Batch Mode Active Learning for Multimedia Pattern Recognition

by

Shayok Chakraborty

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2013 by the
Graduate Supervisory Committee:

Sethuraman Panchanathan, Chair
Vineeth Nallure Balasubramanian
Baoxin Li
Hans Mittelmann
Jieping Ye

ARIZONA STATE UNIVERSITY

May 2013

ABSTRACT

The rapid escalation of technology and the widespread emergence of modern technological equipments have resulted in the generation of humongous amounts of digital data (in the form of images, videos and text). This has expanded the possibility of solving real world problems using computational learning frameworks. However, while gathering a large amount of data is cheap and easy, annotating them with class labels is an expensive process in terms of time, labor and human expertise. This has paved the way for research in the field of active learning. Such algorithms automatically select the salient and exemplar instances from large quantities of unlabeled data and are effective in reducing human labeling effort in inducing classification models. To utilize the possible presence of multiple labeling agents, there have been attempts towards a batch mode form of active learning, where a batch of data instances is selected simultaneously for manual annotation. This dissertation is aimed at the development of novel batch mode active learning algorithms to reduce manual effort in training classification models in real world multimedia pattern recognition applications. Four major contributions are proposed in this work: (i) a framework for dynamic batch mode active learning, where the batch size and the specific data instances to be queried are selected adaptively through a single formulation, based on the complexity of the data stream in question, (ii) a batch mode active learning strategy for fuzzy label classification problems, where there is an inherent imprecision and vagueness in the class label definitions, (iii) batch mode active learning algorithms based on convex relaxations of an NP-hard integer quadratic programming (IQP) problem, with guaranteed bounds on the solution quality and (iv) an active matrix completion algorithm and its application to solve several variants of the active learning problem (transductive active learning, multi-label active learning, active feature acquisition and active learning for regression). These contributions are validated on the face recognition and facial expression recognition problems (which are commonly encountered in real world applications like

robotics, security and assistive technology for the blind and the visually impaired) and also on collaborative filtering applications like movie recommendation.

DEDICATION

To my parents for inspiring me to be a scientist and for all their sacrifices.

ACKNOWLEDGEMENTS

My tenure at Arizona State University has been influenced and guided by a number of people to whom I am deeply indebted. Without their help, friendship and support, this thesis would have never seen the light of day. I am happy to take this opportunity to thank those who have influenced me during my graduate career.

I would like to express my heartfelt gratitude to my mentor and advisor Dr. Sethuraman Panchanathan. I could not have asked for a better role model, who is inspirational, supportive and patient. He gave me the freedom to pursue my own research interests and instilled in me an urge to perform cutting-edge research throughout the course of my PhD. I feel exceedingly privileged to have had his guidance and support through every high and low of my career.

I would like to thank Dr. Jieping Ye for sparing his time to interact with me and for providing me with valuable insights about my work that have shaped my research and thinking. His courses on data mining, machine learning and linear algebra have equipped me with the fundamental knowledge and have motivated me to pursue a career in the machine learning / data mining domain. I would like to convey my sincere gratitude to Dr. Vineeth Balasubramanian for his precious advice throughout my doctoral studies. He has taught me a great deal about computer vision, machine learning and how to look at real problems and results from an insatiably curious and scientific point of view. I shared my first graduate level publication with him (IEEE CVPR 2008), which was a landmark experience in my graduate career. Most importantly, Vineeth has been a good friend and ally throughout the journey. I would also like to thank Dr. Baoxin Li and Dr. Hans Mittelmann for serving on my dissertation committee and for providing valuable feedback on my research.

It has been an enriching experience working with all my fellow members at the Center for Cognitive Ubiquitous Computing (CUbiC) at Arizona State University. I am really proud to be a member of this lab, which has produced several outstanding PhD graduates over the years. I have immensely benefitted from our Machine Learning discussions with Vineeth, CK, Sunaad, Ramkiran, Ashok, Prasanth and Hemanth. My sincere gratitude goes to Terri, Troy, Sreekar, Gaurav, John, Mohammad, Dirk, Lakshmie, David, Daniel, Hiranmayi, Lakshmi, Jessie, Mike, Arash, Ramin, Eric, Shantanu, Bijan, Kian, Derrick and Chelsea for all their help, support and kindness that made CUbiC a home away from home. I would also like to thank Kathy for her prompt help whenever I needed. I would like to extend my gratitude to all the faculty and staff at Arizona State University for providing me with all the necessary support during the course of my PhD tenure. In particular, I would like to thank Dr. Gerald Farin for his guidance in the initial stages of my doctoral career.

During the course of my PhD, I have had the opportunity to collaborate with Dr. Juan Nolazco, Paola Garcia and Roberto Aceves (Technologico de Monterrey, Mexico). I would like to thank them for their help and for the thought-provoking conversations. I would also like to thank Dr. Jay Stokes (Microsoft Research, Redmond) for mentoring me during my internship at MSR. My doctoral work has been funded generously by grants from the National Science Foundation (NSF IIS-0326544, NSF IIS-HCC 1116360) and the ASU Office of Knowledge and Enterprise Development. I sincerely thank them for their kind support.

I would also like to take this opportunity to thank all my friends here in Arizona (too many to list here, but you know who you are!) who have enriched my life with their warmth and concern. They have kept me good company throughout my PhD career and have left me with wonderful memories of good times.

Last, but most importantly, I would like to dedicate this work to my parents, who have been an unwavering source of support and inspiration in every step of my life and career. Both of them being scientists, they have instilled in me the constant quest for scientific knowledge, which motivated me to pursue a doctoral degree in science and engineering. They have always stood beside me through the peaks and valleys of my PhD. I would not be what I am today without their love, care and support. To my dear parents, words fail to express my gratitude.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS vii

LIST OF TABLES xiii

LIST OF FIGURES xv

CHAPTER

1 INTRODUCTION AND MOTIVATION 1

1.1 Inspiration from Human Centered Multimedia Computing 1

1.2 Active Learning 4

1.3 General Approach to Active Learning 6

1.4 Active Learning in Education 7

1.5 Generalizations of Active Learning 9

Active Learning for Structured Outputs 9

Active Feature Acquisition 10

Active Class Selection 11

Active Clustering 12

Active Learning for Regression 12

Multi-Label Active Learning 13

Multiple Instance Active Learning 14

1.6 Analysis of Active Learning 15

Empirical Analysis 15

Theoretical Analysis 16

1.7 Related Research Areas 18

Semi-Supervised Learning 18

Reinforcement Learning 19

Equivalence Query Learning 20

CHAPTER Page

1.8 Batch Mode Active Learning (BMAL) 20
1.9 Rationale and Contributions 21
1.10 Thesis Overview 22
2 RELATED WORK 24
2.1 Online Active Learning 25
2.2 Pool based Active Learning with Serial Query 26
SVM based methods 27
Statistical methods 28
Ensemble based methods 31
Other miscellaneous approaches 32
2.3 Batch Mode Active Learning (BMAL) 36
3 DYNAMIC BATCH MODE ACTIVE LEARNING 42
3.1 Clustering based Batch Size Selection : An Intuitive Approach 43
3.2 Dynamic Batch Mode Active Learning via Stochastic Gradient Descent (SGD) 45
Solving the Optimization Problem 48
3.3 Dynamic Batch Mode Active Learning via Submodular Optimization . . . 50
Submodularity of the Objective Function 53
Greedy Solution to the Optimization Problem 54
3.4 Using the Proposed Frameworks for Static BMAL 55
3.5 Experiments and Results 56
Datasets and Feature Extraction 56
Experiment 1: Dynamic vs. Static BMAL 57
Experiment 2: Batch Selection Criteria of BMAL Algorithms For a Given Batch Size 61

Experiment 3: Performance of the Proposed Dynamic BMAL with Varying Complexities of a Video Stream 63
Experiment 4: Effect of the Cost Parameter 66
Experiment 5: Study of Solution Quality 67
Experiment 6: GTAM Algorithm for Label Prediction 68
3.6 Discussions 69
4 BATCH MODE ACTIVE LEARNING FOR FUZZY LABEL PROBLEMS : AN ANALYSIS WITH FACIAL EXPRESSION RECOGNITION 72
4.1 Fuzzy Sets and Membership Functions 73
4.2 Batch Mode Active Learning for Fuzzy Label Problems 74
4.3 Submodularity of the Objective Function 76
4.4 Greedy Solution with Performance Guarantees 77
4.5 Experiments and Results 79
Datasets and Feature Extraction 79
Classification Model 79
Experimental Setup 80
Ground Truth 80
Experiment 1: Fuzzy BMAL vs. Random Sampling 81
Experiment 2: Fuzzy BMAL vs. Crisp BMAL 82
4.6 Discussions 83
5 ACTIVE BATCH SELECTION VIA CONVEX RELAXATIONS WITH GUARANTEED SOLUTION BOUNDS 85
5.1 The Proposed Batch Mode Active Learning Formulation 86
5.2 BatchRank : Convex Relaxation I 88
Solution Bound of BatchRank 90
5.3 BatchRand: Convex Relaxation II 92
Multi-dimensional Relaxation 92

CHAPTER Page

Semi-Definite Programming (SDP) Relaxation 93
Probabilistic Solution Guarantee 94
5.4 Computational Considerations 95
5.5 Experiments and Results 96
Datasets and Feature Extraction 96
Competing Algorithms and Experiment Set-up 98
Experiment 1: Batch Mode Active Learning on Binary and Multi-class Datasets 99
Experiment 2: Batch Mode Active Learning on Multi-Label Datasets . . . 101
Experiment 3: Run Time Analysis 102
Experiment 4: Validation of the Solution Bounds 104
Experiment 5: Noise Sensitivity 105
Experiment 6: Population Imbalance 107
Experiment 7: Visual Demonstration 108
5.6 Discussions 110
6 ACTIVE MATRIX COMPLETION 111
6.1 Background 111
6.2 Matrix Completion : A Brief Survey 114
6.3 Active Matrix Completion using the Conditional Gaussian Distribution . . 115
GLasso 116
MissGlasso 116
6.4 Active Matrix Completion using Query by Committee (QBC) 120
6.5 Active Matrix Completion using Committee Stability 121
6.6 Computational Considerations 122
Efficient Inverse Covariance Estimation 122
Efficient Low Rank Matrix Completion 124
6.7 Experiments and Results 125

Experiment 1: Image Datasets 126
Experiment 2: Recommendation Systems 128
Experiment 3: Computation Time Analysis 130
6.8 Extension to Active Learning Problem Settings 131
Transductive Active Learning 132
Multi-label Active Learning 134
Active Learning in Regression 135
Active Feature Acquisition 137
6.9 Discussions 138
7 GENERALIZATIONS AND EXTENSIONS 140
7.1 Varying the Criteria for Batch Selection 140
Sparsity based Objective Function 141
Perplexity based Objective Function 142
Comparison against the heuristic approaches 142
7.2 Learning from Multiple Sources of Information 143
Batch Mode Active Learning from Multiple Sources of Information . . . 145
7.3 Context Aware Learning 148
Batch Mode Active Learning Framework to Incorporate Contextual Information 149
7.4 Discussions 151
8 RELATED CONTRIBUTIONS 152
8.1 Theory of Conformal Predictions 153
8.2 Generalized Query by Transduction (GQBT) 157
Why Generalized QBT? 161
Combining multiple criteria for active learning 163
8.3 Experiments and Results 163

CHAPTER Page

Application to Face Recognition 166

8.4 Discussions 168

9 CONCLUSIONS AND FUTURE WORK 170

9.1 Summary of Contributions 171

9.2 Future Work 175

Dynamic Batch Mode Active Learning 176

Batch Mode Active Learning for Fuzzy Classification 176

Batch Mode Active Learning with Guaranteed Solution Bounds 177

Active Matrix Completion 177

Other Possible Directions of Future Work 178

REFERENCES 181

APPENDIX

APPENDIX A 209

LIST OF TABLES

Table Page

3.1 Mean predicted batch size (PBS) and percent labeling cost reduction (LCR) using SGD based dynamic selection against static selection with batch size 80. 59

3.2 Mean predicted batch size (PBS) and percent labeling cost reduction (LCR) using submodularity based dynamic selection against static selection with batch size 80. 59

3.3 Average time taken (in seconds) to query a batch of 10 images from an unlabeled video with 250 images. 62

3.4 Test set accuracies using Proposed and Clustering based Dynamic BMAL on the VidTIMIT dataset with increasing proportions of new identities. 65

3.5 Test set accuracies using Proposed and Clustering based Dynamic BMAL on the MOBIO dataset with increasing proportions of new identities. 65

4.1 Average time (in seconds) taken by the Fuzzy and Discriminative Batch Mode Active Learning techniques to query a batch of 10 points from the unlabeled pool. 83

5.1 Dataset Details 99

5.2 Time taken (in seconds) to query a batch of samples from an unlabeled set. (Binary and Multi-class Datasets). 103

5.3 Time taken (in seconds) to query a batch of samples from an unlabeled set (Multi-label Datasets). 104

6.1 Recommendation Datasets Details 129

6.2 Average time taken (seconds) to query a batch of indices from the matrix. . . 131

Table Page

8.1 Datasets from the UCI Machine Learning repository used in our experiments. An equal number of examples from each class was used in the initial training set. For example, for the Breast Cancer dataset, 5 examples from each class were used to form the initial training set of 10 examples. 165

8.2 Label complexities of each of the methods for all the datasets. Label complexity is defined as the percentage of the unlabeled pool that is queried to reach the peak accuracy in the active learning process. Note the low label complexities of the proposed approach in all the cases. Also, note that the label complexities for the other methods on datasets like Waveform and Image Segmentation are very high although the accuracy did increase at a reasonable rate in the active learning process in Figure 8.5. This only implies that these methods reached their peak accuracy when the unlabeled pool was almost exhausted. 167

LIST OF FIGURES

Figure Page

1.1 A first prototype of the Social Interaction Assistant 2

1.2 General Schema of a Passive Learner 5

1.3 General Schema of an Active Learner 5

1.4 The Cone of Learning 7

2.1 Categories of Active Learning 24

3.1 Sample images from the VidTIMIT and MOBIO datasets 57

3.2 Dynamic vs Static BMAL on the VidTIMIT and MOBIO datasets (static batch size = 10). Best viewed in color. 59

3.3 Dynamic vs Static BMAL on the VidTIMIT and MOBIO datasets (static batch size = 80). Best viewed in color. 60

3.4 Batch Mode Active Learning on the VidTIMIT and MOBIO datasets (Best viewed in color). 62

3.5 Study of the Proposed Dynamic Batch Selection Frameworks with Varying Complexities of a Video Stream 64

3.6 Effect of the Cost Parameter in the SGD based Dynamic BMAL 67

3.7 Validation of Solution Quality for the SGD and Submodularity based Dynamic BMAL 68

3.8 Validation of the Efficacy of GTAM 69

4.1 The Social Interaction Assistant for individuals with visual impairments [1] . 72

4.2 Comparison of Fuzzy BMAL against Random Sampling (Best viewed in color) 81

4.3 Comparison of Fuzzy BMAL against Crisp BMAL and the Discriminative BMAL algorithm (Best viewed in color) 82

5.1 Batch Mode Active Learning on the UCI datasets. (Best viewed in color.) . . 100

5.2 Batch Mode Active Learning Face Recognition and Facial Expression Recognition datasets. (Best viewed in color.) 101
5.3 Multi-label Batch Mode Active Learning on the Scene and Yeast datasets. (Best viewed in color.) 103
5.4 Validation of Solution Bounds of BatchRank and BatchRand. (Best viewed in color.) 105
5.5 Noise Sensitivity of BatchRank and BatchRand on the VidTIMIT dataset. (Best viewed in color.) 106
5.6 Population Imbalance : VidTIMIT dataset. (Best viewed in color.) 107
5.7 Population Imbalance : Confusion matrices for Random Selection, BatchRank and BatchRand (Max trace = 4500) : VidTIMIT dataset. 108
5.8 Batch of images selected using BatchRand 109
5.9 Batch of images selected using BatchRank 109
5.10 Batch of images selected using Random Sampling 109
5.11 Total pairwise distance among the selected images 110
6.1 GrayScale images used in our experiments 126
6.2 Active Matrix Completion on Image Datasets (Best viewed in color). Degree of Sparsity = 60% 127
6.3 Active Matrix Completion on Recommendation Datasets (Best viewed in color) 130
6.4 Transductive Active Learning using Active Matrix Completion (Best viewed in color). 133
6.5 Multi-Label Active Learning using Active Matrix Completion (Best viewed in color). 135
6.6 Pose images from the FacePix dataset 136
6.7 Active Learning in Regression using Active Matrix Completion (Best viewed in color). 137

6.8 Active Feature Acquisition using Active Matrix Completion (Best viewed in color). 139
7.1 Performances of different Batch Mode Active Learning schemes on the VidTIMIT and MBGC datasets (Best viewed in color). 143
7.2 Categorization of approaches towards multimodal biometrics 144
7.3 An overview of the approaches to information fusion 145
7.4 Batch Mode Active Learning from multiple sources on the VidTIMIT and MBGC datasets. 147
7.5 Context Aware Learning on the VidTIMIT and MBGC datasets. 150
8.1 An illustration of the non-conformity measure defined for k-NN 155
8.2 Performance of the CP Framework on the Cardiac Patient Dataset. Note that the errors are calibrated at each of the confidence levels. For instance, at 80% confidence level, the number of errors will always be less than 20% of the total number of test examples. 157
8.3 Comparison of the proposed GQBT approach with Ho and Wechsler's QBT approach on the Musk dataset from the UCI Machine Learning repository. Note that our approach reaches the peak accuracy by querying ≈ 80 examples, while the latter needs ≈ 160 examples. 162
8.4 Performance comparison on the Musk dataset (as in Figure 8.3). Note the reduction in label complexity obtained by combining the p-values from the two non-conformity measures discussed in Section 8.2. The proposed approach needs only ≈ 50 examples to reach the peak accuracy. 164
8.5 Results with datasets from the UCI Machine Learning repository. In the Musk dataset, the results started with an accuracy of $\approx 70\%$, but since all methods had similar initial accuracies, the graph is shown from 85% accuracy onwards, where the differences in performance are clearly seen. 166

8.6 Results obtained on the VidTIMIT dataset. Note that the GQBT approach led to a significantly higher peak accuracy, and had a lower label complexity of 58.8% to reach the peak accuracy. Label complexities of the other methods: Ho and Wechsler's QBT - 98.2%; Query by Committee - 100%; Margin-based SVM - 89%; Random sampling - 99.6% 168

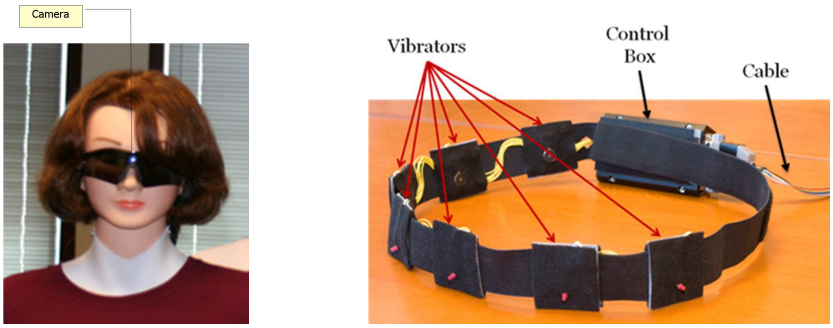
Chapter 1

INTRODUCTION AND MOTIVATION

1.1 Inspiration from Human Centered Multimedia Computing

Over the last decade, there has been an increasing focus on the development of assistive technology to aid physically challenged individuals in their daily life activities. A prominent number of these devices are based on the effective analysis and interpretation of video data. For instance, people with severe paralysis often have communication abilities that are limited to "yes" and "no" responses made with small head, hand or eye movements. Still, they desire to express themselves in conversations with their families and caregivers and initiate topics of discussion. To enable this, the "Camera Mouse" interface was developed by researchers at Boston University [2]. The system tracks the user's movements using a video camera and translates them into movements of the mouse pointer on the screen. Further, the increasing focus on accessibility has resulted in the design and development of several assistive technologies to aid people with visual impairments in their daily activities. Most of these devices have been centered on enhancing the interaction of a user who is blind or visually impaired with objects and environments, such as a computer monitor, personal digital assistant, cellphone, road traffic, or a grocery store. Although these efforts are very essential for the quality of life of these individuals, there is also a need (which has so far not been seriously considered) to enrich the interactions of individuals who are blind, with other individuals.

Non-verbal cues (including prosody, elements of the physical environment, the appearance of communicators and physical movements) account for as much as 65% of the information communicated during social interactions [1]. However, more than 1.1 million individuals in the US who are legally blind (and 37 million worldwide) have a



(a) A pair of glasses with a small camera mounted on the nose-bridge

(b) A first prototype of the haptic belt. A wireless version of this belt has already been built in the second phase

Figure 1.1: A first prototype of the Social Interaction Assistant

limited experience of this privilege of social interactions. These individuals continue to be faced with fundamental challenges in coping with everyday interactions in their social lives. To address this basic need, the Center for Cognitive Ubiquitous Computing (CUbiC) (http://cubic.asu.edu) at Arizona State University has focused its efforts on the design and development of a Social Interaction Assistant (SIA) that is intended to enrich the experience of visually challenged individuals by providing real-time access to information about individuals and their surrounds.

In our current prototype, the Interaction Assistant device consists of a pair of glasses with a small camera mounted on the nosebridge (as shown in Figure 1.1(a)). The incoming video stream is analyzed using computer vision algorithms to extract information about the surroundings. When a person comes in the field of view of the camera, his/her face is captured using face detection algorithms. The identity of the person is then determined by a recognition engine through a similarity match with a database of stored images [3]. Most assistive devices provide only audio outputs, which is not a practical solution for visually impaired individuals in the context of social interactions, as they use their ears as their "eyes" to perceive the environment. To overcome this fundamental limitation, we have designed a vibrotactile haptic belt, which consists of a set of 7 vibrators,

to be worn by the user around his waist (Figure 1.1(b)). Information about the direction of an approaching individual is conveyed through the location of vibration, and the distance to the user is encoded in the duration of vibration. For more details about the functioning of the system, please refer [4].

The video camera integrated into the SIA system has a high frame rate (typically 25 frames per second). Thus, within a short duration of time, a staggering number of images will be acquired by the system. Moreover, these images will have considerable redundancy among them because of the high frame rate of the camera. This massive amount of superfluous data needs to be efficiently processed to ensure reliable functioning of the end-to-end system. Further, to train the underlying classification models with the acquired data, the captured images need to be provided with class labels. Manual labeling of such a large scale data is an expensive process in terms of time, labor and human expertise. Thus, a machine learning algorithm which can automatically select the salient and exemplar instances for manual labeling from vast quantities of unlabeled data will be paramount importance in facilitating the learning process in such an application.

The problem of hand labeling large amounts of data is commonly encountered in a variety of applications in this era. Over the last couple of decades, technology has advanced in leaps and bounds. This has resulted in the frequent generation of humongous quantities of digital data (in the form of images, videos and text among others). These data are typically unlabeled and need substantial human effort for annotation. A few examples are presented here:

- **Speech Recognition:** Accurate labeling of speech utterances is extremely time consuming and requires trained linguists. Zhu [5] reported that annotation at the word level can take ten times longer than the actual audio (e.g. one minute of speech takes ten minutes to label) and annotating phonemes can take 400 times as long (nearly

seven hours). The problem becomes even more severe and compounded for rare languages or dialects.

- **Information Extraction:** Good information extraction systems must be trained using labeled documents with detailed annotations. Users hand label entities or relations of interest in text, such as person and organization names, or whether a person works for a particular organization. Locating entities and relations can take a half-hour or more for even simple newswire stories (Settles *et al.* [6]). Annotations for other knowledge domains may require additional expertise, e.g. annotating gene and disease mentions for biomedical information extraction usually requires PhD-level biologists.

- **Text Classification:** Learning to classify documents (e.g. articles or web pages) requires that users label each document with particular labels, like "relevant" or "not relevant". Annotation of thousands of these instances can be tedious and even redundant.

The aforementioned applications motivate the development of a framework that can automatically identify the representative samples from vast amounts of redundant and unlabeled data. This can tremendously reduce human annotation effort in training classification / regression models to identify patterns in the data. In the machine learning literature, such a framework is referred to as *active learning*.

1.2 Active Learning

The primary goal in any classification problem is to learn a function $f : X \rightarrow C$, which maps input feature vectors X into the corresponding output classes C. To develop a robust recognition engine, it is indispensable to have a large amount of labeled data in the form of a training set. Usually, this data is sampled at random from the underlying distribution

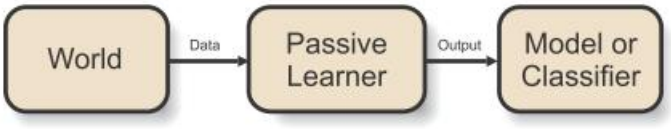


Figure 1.2: General Schema of a Passive Learner

Figure 1.3: General Schema of an Active Learner

and is then used to induce the classifier. This methodology is called *passive learning*. A passive learner receives a randomly selected dataset from the world and outputs a classifier [7] (as depicted in Figure 1.2).

However, as noted in Section 1.1, while gathering a large amount of unlabeled data is cheap and easy, annotating them with class labels involves significant human effort. To alleviate this problem, *active learning* strategies have been proposed in the literature. Instead of randomly selecting data points for manual labeling, active learners gather information about the world by querying the class labels of certain specific unlabeled points and receiving responses. The queries are not pre-defined, but are selected dynamically based on the responses to the previous queries. This tremendously reduces the human annotation effort as only a few points, that are identified by the algorithm, need to be labeled manually. Moreover, the ability of the active learner to adaptively query the world based on past experience endows it with greater generalization capability, which makes it better than the standard passive learner. Figure 1.3 depicts the general schema of an active learner.

1.3 General Approach to Active Learning

The fundamental step in active learning is to define the notion of a **model** M and its **model quality** (or **model loss** $Loss(M)$). The definition of model and the associated model loss can be tailored according to the specific application at hand. Given this notion of model loss, an active learner selects the nest query as the one that will result in minimal loss of the future model. This approach is *myopic* in the sense that the learner is attempting to greedily ask the single best next query. Myopia is a standard approximation used in sequential decision making problems [8, 9, 10]. While considering to ask a potential query **q**, the learner needs to assess the loss of the subsequent model M'. The posterior model M' is the original model M updated with query q and response x. Since the learner has no knowledge of the true response x of the potential query, it needs to perform some kind of averaging or aggregation. A natural approach is to maintain a distribution over the possible responses to each query. The *expected* model loss can then be computed for a given query, where the expectation is taken over all the possible responses to that query:

$$Loss(q) = E_x Loss(M') \tag{1.1}$$

This active learning framework results in selecting the query producing the minimal expected model loss. In statistics, a standard alternative to minimizing the expected model loss is to minimize the maximum loss [11]. This implies that the response x will always be the response that gives the highest model loss:

$$Loss(q) = \max_x Loss(M') \tag{1.2}$$

This active learning strategy results in selecting the query that produces the minimax model loss. Both the averaging and the aggregation methods are useful - one may be more advantageous over the other in specific situations. The general schema of an active learner is depicted in Algorithm 1.

Algorithm 1 General Schema of an Active Learner [7]

```
for i = 1 to totalQueries do
    for each q in potentialQueries do
        Evaluate Loss(q)
    end for
    Ask query q for which Loss(q) is lowest
    Update model M with query q and response x
end for
return Model M
```

1.4 Active Learning in Education

The concept of active learning was initially promoted in the domain of education. Most of the time, in a typical classroom setting, students are involved only passively in learning; they merely listen to the instructor, glance occasionally at the blackboard or the slide and read (when required) text books. Research shows that such passive involvement generally leads to a limited retention of knowledge by students, as indicated in the "cone of learning" shown in Figure 1.4 [12]:

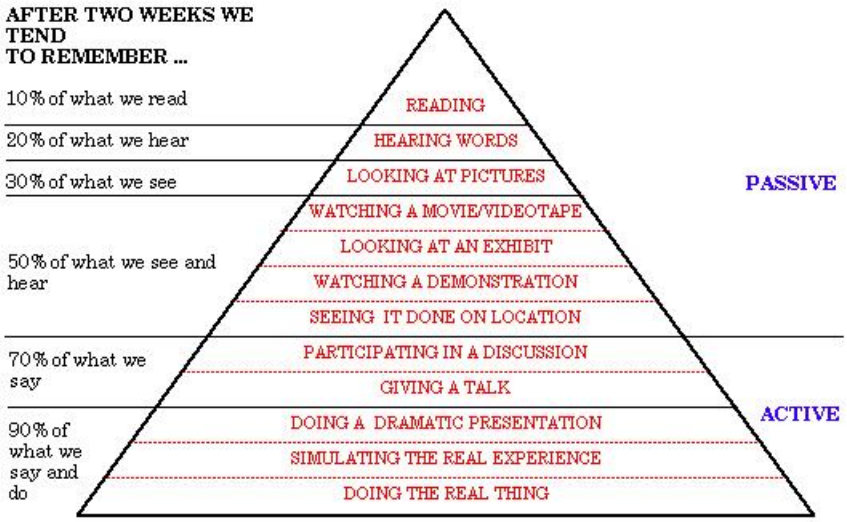


Figure 1.4: The Cone of Learning

However, research clearly supports the widely accepted proposition that students need to do more than just listen to learn [13]. By re-organizing or adapting the ways they

present material to students, instructors can create an environment in which knowledge retention is significantly increased, subject to the co-operation of the students. A method to implement such an effective learning environment is through *active learning*. Active learning involves students directly and actively in the learning process itself. Bonwell and Eison [14] described active learning in the following fashion: "When using active learning, students are engaged in more activities than just listening. They are involved in dialog, debate, writing, and problem solving, as well as higher-order thinking, e.g., analysis, synthesis, evaluation".

Similarly, Johnson *et al.* [15] defined cooperative learning as "the instructional use of small groups so that students work together to maximize their own and each other's learning. Five essential components must be present for small-group learning to be truly cooperative:

1. clear positive interdependence between students
2. face to face interaction
3. individual accountability
4. emphasize interpersonal and small-group skill
5. processes must be in place for group review to improve effectiveness."

Prince [16] surveyed many different studies on active, collaborative, cooperative and problem-based learning. Many faculty members across the Foundation Coalition (FC) have been using active/cooperative learning student teams as an integral part of their courses since the inception of the Coalition in 1993. These faculty members have coupled their experience and expertise with research to create a number of resources that will help them use active/cooperative learning together with their lectures. They unanimously agree that proper active learning techniques have a powerful impact upon students' learning, in stimulating students' thinking and in the effective transfer of knowledge.

In the machine learning literature, a student corresponds to a classification model while a teacher corresponds to the universe of data samples the model is supposed to learn from. In the passive learning setting, data points are selected at random from the universe to train the model. This is equivalent to the scenario when the students merely listen to the instructor without actively participating in the learning process. In contrary, in active learning, similar to the classroom environment, the learner actively learns from the universe by asking intelligent queries and receiving responses. This has the potential to produce a better model as it gets trained on the salient and exemplar instances from the population. Thus, active learning in education is analogous to active machine learning. This concept was therefore introduced in the machine learning literature about a couple of decades back and has been a dynamic research topic since then.

1.5 Generalizations of Active Learning

In this section, we discuss some generalizations and extensions of the standard active learning framework to different problem settings.

Active Learning for Structured Outputs

Many important learning problems involve predicting structured outputs on instances, such as sequences and trees. For example, an information extraction problem can be viewed as a sequence labeling task. Let $x =< x_1, \ldots, x_T >$ be an observation sequence of length T with a corresponding label sequence $y =< y_1, \ldots, y_T >$. Words in a sentence correspond to tokens in the input sequence x, which are mapped to labels in y. The labels indicate whether a given word belongs to a particular entity class of interest. Unlike simpler classification tasks, each instance x in this setting is not represented by a single feature vector, but rather a structured sequence of feature vectors, one for each token (or word). For such problems, labels are typically predicted by a sequence model based on

a probabilistic finite state machines, such as Conditional Random Fields (CRFs) or Hidden Markov Models (HMMs). Settles and Craven [17] evaluated a large number of active learning algorithms for sequence labeling tasks using probabilistic sequence models like CRFs. Most of these algorithms can be generalized to other probabilistic sequence models such as HMMs [18, 19] and probabilistic context-free grammars [20, 21]. Thompson *et al.* [22] proposed query strategies for structured output tasks like semantic parsing and information extraction using inductive logic programming methods.

Active Feature Acquisition

In some learning problems, instances may have incomplete feature descriptions due to errors in the acquisition process, reluctance of subjects to divulge information etc. Consider a learning model used in medical diagnosis which has access to some patient symptom information, but not other data that require complex, expensive, or risky medical procedures. Here, the task of the model is to suggest a diagnosis using incomplete patient information as the feature set. Active feature selection can alleviate these problems by allowing the learner to request more complete feature information. The assumption is that additional features can be obtained at a cost, such as running additional diagnostic procedures. The goal in active feature acquisition is to select the most informative features to obtain during training, rather than randomly or exhaustively acquiring all new features for all training instances. Several approaches have been proposed for active feature selection and data acquisition [23, 24, 25].

Similarly, active classification considers the case in which missing feature values may be obtained during classification (test time) rather than during training. Greiner *et al.* [26] introduced this setting and provided a PAC-style theoretical analysis of learning such classifiers given a fixed budget. Variants of naive Bayes [27] and decision tree [28] classifiers have also been proposed to minimize costs at classification time. Typically, these are

evaluated in terms of their total cost (feature acquisition plus misclassification) as a function of the number of missing values. The approaches are flexible enough to incorporate other types of costs, such as delays between query time and value acquisition [29]. A different approach is to model the feature acquisition task as a sequence of decisions to either acquire more information or to terminate and make a prediction [30]. Kapoor and Horvitz [31] developed an algorithm that bridged active learning and real-time diagnostic feature acquisition into a holistic approach to information acquisition that simultaneously considered the extension of the predicitve model and the probing of a case at hand. Sindhwani *et al.* [32] addressed the problem of active dual supervision to optimally query an example and feature labeling oracle to simultaneously collect two different forms of supervision, with the objective of building the best classifier in the most cost-effective manner. Kong *et al.* [33] studied the dual active feature and sample selection problem for graph classification. The authors demonstrated how to find a useful query graph and a set of optimal features simultaneously to minimize the labeling efforts in graph classification.

Active Class Selection

The inherent assumption in an active learning problem is that a large amount of unlabeled data is readily available, but labeling the data is time consuming and expensive. Active class selection considers the opposite problem, where a learner is allowed to query a known class label, and obtaining each instance incurs a cost. Lomasky *et al.* [34] proposed several active class selection query algorithms for an "artificial nose" task, in which a machine learns to discriminate between different vapor types (the class labels) which must be chemically synthesized (to generate the instances). Some of their approaches show significant gains over uniform class sampling, which is the passive learning equivalent.

Active Clustering

Active learning has also been judiciously used in unsupervised learning, where the task is to organize a large amount of unlabeled data in a meaningful way. Typical examples of such algorithms include clustering, which exploit the latent structure in the data to derive underlying patterns. Hofmann and Buhmann [35] proposed an active clustering algorithm for proximity data based on an expected value of the information criterion. Some clutsreing algorithm operate under certain constraints, where a user can specify apriori that two instances must belong to the same cluster, or that two others cannot. Grira *et al.* [36] explored an active variant of this approach for image databases where queries take the form of "must-link" and "cannot-link" constraints on similar or dissimilar images. Huang and Mitchell [37] presented an active learning framework that integrated four different types of user feedback into the clustering algorithm and provided empirical evidence of substantial improvement in text clustering when user input was incorporated. Wauthier *et al.* [38] developed an active learning algorithm for spectral clustering that incrementally measured only those similarities which are most likely to remove uncertainty in an intermediate clustering solution. Biswas and Jacobs [39] proposed an active clustering algorithm which selected the most useful pairs to be manually annotated; the informativeness of a pair of points was computed based on the expected change in clustering that could be induced using the points in question.

Active Learning for Regression

Although primarily used for classification, active learning has also been applied in regression problems to select the samples that are most informative in learning the regression function. Castro *et al.* [40] analyzed the theoretical capabilities of active learning for estimating regression functions in the presence of noise. The proposed theory showed

promise in a number of applications, including field estimation using wireless sensor networks and fault line detection. Sugiyama and Rubens [41] proposed an ensemble active learning algorithm for performing active learning and model selection simultaneously in a linear regression problem. Sugiyama [42] also proposed an active learning approach for linear regression using importance weighted least squares learning method based on conditional expectation of the generalization error. Yu *et al.* [43] proposed a transductive active learning approach which exploited the presence of unlabeled data in a linear regression learning problem. Burbidge *et al.* [44] presented a variance based Query by Committee algorithm for active point selection in a regression setting.

Multi-Label Active Learning

Multi-label classification is a generalization of conventional classification problems, where each data sample can have multiple labels [45, 46]. For instance, classifying the contents of a natural scenery image is a multi-label problem, as a single image can have multiple contents (like sunset, ocean, mountains etc.) associated with it. Annotating a data point in a multi-label scenario requires a human oracle to check the presence/absence of every possible class in the data point. Thus, the need for active learning in a multi-label setting is even more pronounced. There has been some previous effort in the domain of multi-label active learning. Singh *et al.* [47] and Brinker [48] proposed multi-label active learning strategies based on uncertainty sampling using SVMs, where the uncertainty was quantified as the distance from the hyperplane and using the entropy of the learner. Zhang *et al.* [49] proposed a multi-label active learning strategy where the uncertainties across multiple views were fused for active sample selection. Li *et al.* [50] proposed two loss strategies Max Loss and Mean Max Loss for SVM-based multi-label active learning. Along similar lines, Yang *et al.* [51] proposed a multi-label active learning strategy for text classification, where the sample selection was based on expected reduction in model loss. Hung

and Lin [52] proposed a multi-label active learning framework which was characterized by a major learner for making predictions and an auxiliary learner for helping with query decisions and a query criterion based on the disagreement between the two learners. However, all these methods are based on querying *all the labels* of the selected samples and do not exploit the inherent correlations among the labels of a given sample. Qi *et al.* [53] proposed an efficient online adaptation model, based on the minimization of a multi-label Bayesian classification error bound, which queried informative sample-label pairs instead of all the labels of an unlabeled sample.

Multiple Instance Active Learning

In multiple instance (MI) learning problems, the instances are naturally organized into bags and it is the bags instead of the individual instances, that are labeled for training [54]. A bag $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ is labeled negative if every instance it contains is negative. A bag is labeled positive if at least one of its instances is positive (positive bags may also contain negative instances). This framework has been applied to a wide variety of tasks including drug activity prediction, content based image retrieval, text classification, stock prediction and protein family modeling. Multiple instance active learning can reduce the labeling burden in problem domains where labels can be acquired at both bag-level and instance-level granularities. This approach is well-motivated in learning settings where it is cheap to acquire bag labels and possible (but expensive) to acquire more fine-grained instance labels. Settles *et al.* [55, 56] proposed an active query selection strategy to learn from labels at mixed levels of granularity and demonstrated that learning from instance labels can significantly improve performance of a basic MI learning algorithm in the content based image retrieval and text classification domains. Liu *et al.* [57] proposed three strategies for multiple instance active learning - (i) selecting bags only, (ii) selecting instances only and (iii) selecting both bags and instances. The authors empirically established that

selecting both bags and instances outperformed the other two strategies of selecting instances or bags individually.

1.6 Analysis of Active Learning

In this section, we discuss some of the empirical and theoretical evidence for how and when active learning approaches can be successful.

Empirical Analysis

Active learning is a well-studied topic in machine learning research. Numerous published results and increased industry adoption seem to indicate that active learning methods have matured to the point of practical use in many applications. However, there are some negative results on active learning that have been published in the literature. Schein and Ungar [58] analyzed active learning using the logistic regression model and showed that it can sometimes require more labeled instances than passive learning. Guo and Schuurmans [59] reported that active learning query strategies are sometimes worse than random sampling. Gasperin [60] presented negative results for active learning in an anaphora resolution task. Baldridge and Palmer [61] found an inconsistency in how well active learning helps that seems to be correlated with the proficiency of the annotator (a domain expert was better utilized by an active learner than a domain novice, who was better suited to a passive learner). Nevertheless, as per majority of the published research, active learning is effective in reducing the number of labeled instances to achieve a given level of accuracy. Tomanek and Olsson [62] reported in a survey that 91% of researchers who used active learning in large-scale annotation projects had their expectations fully or partially met.

Theoretical Analysis

There have been some theoretical analyses on the performance of active learning algorithms. They attempt to derive a bound on the number of queries required to learn a sufficiently accurate model for a given task and prove theoretically that this number is less than in the passive supervised setting. A strong theoretical result of pool based active learning using the Query by Committee (QBC) algorithm was proposed by Freund *et al.* [63]. The authors showed that, under a Bayesian assumption, it is possible to achieve a generalization error ε after seeing $O(d/\varepsilon)$ unlabeled instances, where d is the Vapnik-Chervonenkis (VC) dimension [64] of the model space and requesting only $O(d \log 1/\varepsilon)$ labels. Bachrach *et al.* [65] presented improvements of the proposed approach by limiting the version space via kernel functions. Wang and Zhou [66, 67] theoretically charaterized the sample complexity of active learning in the non-realizable case under multi-view settings in the presence of Tsybakov noise. Raginsky and Rakhlin [68] developed unified information theoretic algorithms for deriving lower bounds for passive and active learning schemes using the Alexander's capacity function. Golovin *et al.* [69] proposed a novel Bayesian active learning algorithm in the presence of noise and theoretically proved that it was competitive with the optimal policy.

Dasgupta *et al.* [70] proposed a variant of the perceptron update rule, which could achieve the same label complexity bounds as reported for QBC. The authors showed that a standard perceptron makes a poor active learner in general requiring $O(1/\varepsilon^2)$ labels as a lower bound. Dasgupta [71] also provided theoretical upper and lower bounds for active learning in the more general pool-based setting. The same author [72] further showed that for homogeneous linear separators in $\Re^d$, the number of labels needed by an active learner to achieve an error rate less than or equal to ε is $O(d \log^2 1/\varepsilon)$, which is exponentially smaller than the usual $\Omega(d/\varepsilon)$ sample complexity of supervised learning. Dasgupta *et*

al. also proposed an importance-weighting based practical approach which has guaranteed label complexity bounds [73]. Balcan *et al.* [74] showed that asymptotically, active learning strategies are better than supervised learning in the limit. Very recently, Dasgupta [75] discussed mathematical properties and empirical performance about the two common intuitions of active learning - selecting query points to shrink the space of the candidate classifiers rapidly and exploiting natural clusters in the unlabeled dataset.

Most of these results have used theoretical frameworks similar to the standard PAC model (probably approximately correct) [76] and assume that some model in the hypothesis class can perfectly classify the instances and the data are noise-free. To overcome these limitations, there have been recent work on *agnostic active learning* (Balcan *et al.* [77]) which only assumes that the unlabeled data are drawn i.i.d from a fixed distribution. Hanneke [78] provided upper bounds on the query complexity for the agnostic setting. Dasgupta *et al.* [79] proposed an efficinet query selection algorithm by presenting a polynomial time reduction from active learning to supervised learning for arbitrary input distributions and model classes. Beygelzimer *et al.* [80] presented a theoretical analysis of an agnostic active learning algorithm that works without the notion of a version space (unlike all previous active learning approaches) and is therefore computationally efficient. Wang [81] proposed sufficient conditions under which agnostic active learning is strictly superior to passive supervised learning. The author established that under some noise condition, if the Bayesian classification boundary and the underlying distribution are smooth to a finite order, active learning achieves polynomial improvement in the label complexity. These agnostic active learning approaches explicitly use complexity bounds to determine which hypotheses are viable and queries can be assessed by how valuable they are in distinguishing among these viable hypotheses. These methods have attractive PAC-style convergence guarantees and complexity bounds that are, in many cases, significantly better than passive learning.

However, most positive theoretical results have been based on intractable algorithms, or methods otherwise too prohibitively complex to be used in practice. Moreover, these studies have largely only been for simple classification problems. In fact, most are limited to binary classification with the goal of minimizing $0/1$-loss and are not easily adapted to other objective functions that may be more appropriate for many applications. Furthermore, some of these methods require an explicit enumeration over the version space, which is not only often intractable but difficult to even consider for complex learning models. However, some recent theoretical work has begun to address these issues, coupled with promising empirical results [82, 73].

1.7 Related Research Areas

Research in the field of active learning is driven by two main ideas - the learner should not be strictly passive and a large amount of unlabeled data is readily available. There are a few areas in the machine learning literature which are based on similar settings - we outline them briefly in this section.

Semi-Supervised Learning

Semi-supervised learning techniques [5] exploit the unlabeled data samples to learn a good classification model. A basic semi-supervised learning technique is self-training, where the learner is first trained with a small amount of labeled data, and then used to classify the unlabeled data. Typically the most confident unlabeled instances, together with their predicted labels, are added to the training set and the process repeats. A complementary technique in active learning is uncertainty sampling (for details, please refer Chapter 2), where the instances about which the model is least confident are selected for querying. Similarly, co-training [83] uses ensemble methods for semi-supervised learning where

models are trained separately with labeled data which are then used to classify the unlabeled data. The unlabeled samples for which a particular model's confidence of prediction exceeds a certain threshold are used to train the other models. Query by Committee is the active learning counterpart, where the unlabeled sample for which the ensemble of learners disagree the most is selected for manual annotation. We thus note that active learning and semi-supervised learning share a few conceptual overlaps. Some active learning formulations, in fact, are based on semi-supervised learning algorithms [84, 85, 86, 87]. Guillory and Bilmes [88] considered the problem of active semi-supervised learning in an offline transductive setting and proved that the error bound on undirected weighted graphs can be generalized by replacing graph cuts with an arbitrary symmetric submodular function. He *et al.* [89] proposed a novel active learning algorithm called Graph Regularized Experimental Design (GRED) where active and semi-supervised learning are combined into a single framework for pixel selection and colorization for the task of image compression.

Reinforcement Learning

In reinforcement learning [90], the learner interacts with the world via actions and tries to find an optimal policy of behavior with respect to "rewards" it receives from the environment. In order to perform well, the learner must be proactive. It is easy to converge on a policy of actions that have worked well in the past but are sub-optimal. In order to improve, a reinforcement learner must take risks and try out actions for which it is uncertain about the outcome, just as an active learner requests labels for instances it is uncertain how to label. This is often called the "exploration-exploitation" trade-off in the reinforcement learning literature. Mihalkova and Mooney [91] proposed an active reinforcement learning approach which aimed to reduce the number of actions required to find an optimal policy. Epshteyn *et al.* [92] proposed an active reinforcement learning framework to learn the transition probabilities of a Markov Decision Process (MDP) by exploring the

regions of space in which the optimal policy is most sensitive. Hoi and Jin [93] proposed a min-max approach to actively learn the

kernel matrix that selected the example pairs leading to the largest classification margin even when the class assignments to the selected pairs are incorrect.

Equivalence Query Learning

An area closely related to active learning is learning with equivalence queries [94]. In such a setting, instead of generating an instance to be labeled by the oracle, the learner instead generates a hypothesis of the target concept class, and the oracle either confirms or denies that the hypothesis is correct. If it is incorrect, the oracle should provide a counter-example, i.e. an instance that would be labeled differently by the true concept and the query hypothesis. However, there are only a few practical applications of equivalence query learning, because an oracle often does not know (or cannot provide) an exact description of the concept class for most real-world problems.

1.8 Batch Mode Active Learning (BMAL)

In a typical active learning setting, the learner is exposed to a pool of unlabeled instances. It is assumed that the data is independent and identically distributed (i.i.d) according to some underlying distribution $F(x)$ and the class labels y are distributed according to some conditional distribution $P(y|x)$. Given an unlabeled pool U, an active learner has three components - (f,q,X). The first component f is the classifier that is trained on the current training set X. The second component q is the query function which decides which instance in the unlabeled pool is to be queried next for its class label. After a particular point is selected, it is supplied to a human oracle for labeling and is then appended to the training set. The model is updated and the process is continued iteratively until some stopping

criterion is satisfied. The active learner finally returns the classifier after a predetermined number of queries.

However, selecting a single instance at a time for manual annotation requires frequent model retraining as the classifiers need to be updated after every single query. With the advent of technologies like the Amazon Mechanical Turk, it is now possible to leverage the intelligence of multiple human users simultaneously in labeling data instances to train a classification model. To address this need, *batch mode active learning* (BMAL) algorithms have been proposed in recent years. Such techniques select a batch of points simultaneously from an unlabeled set for manual labeling and are effective in utilizing the presence of parallel labeling agents and avoiding frequent classifier updates. Sample applications of BMAL include content based image retrieval [95, 96], medical image classification [97] and text classification [98].

1.9 Rationale and Contributions

This work intends to develop novel batch mode active learning algorithms to reduce human annotation effort in real world machine learning problems. This is of tremendous practical importance given the humongous amount of data that are being generated everyday in today's digital world. The proposed frameworks can be judiciously used to develop machine learning models with minimal human effort for a variety of real world applications. Specifically, this work aims to provide four major contributions:

- A dynamic batch mode active learning framework which simultaneously solves for both the batch size and the specific points that need to be queried for manual annotation from an unlabeled set of instances, through a single formulation.
- A batch mode active learning strategy for fuzzy label classification problems where there is an inherent imprecision and vagueness in the class label definitions.

- Batch mode active learning algorithms based on convex relaxations of an NP-hard integer quadratic programming (IQP) problem, with guaranteed bounds on the solution quality
- An active matrix completion algorithm and its application to solve several variants of the active learning problem (transductive active learning, multi-label active learning, active feature acquisition and active learning for regression).

These contributions are validated on the face and expression recognition problems on several challenging biometric datasets (more details on the datasets are presented in the subsequent chapters). Automated recognition of human identity and human facial expression are fundamental problems that need to be solved as part and parcel of the Social Interaction Assistant system for the visually impaired (as described in Section 1.1). Integration of the proposed approaches in the design of such an assistive technology will hopefully reduce the human annotation effort and aid in the development of reliable machine learning models for the challenging recognition tasks. Although validated on these applications in this work, the proposed frameworks are generic and can be used in any application where a large amount of unlabeled data is readily available, but labeling the data is time consuming and expensive.

1.10 Thesis Overview

The subsequent chapters in this thesis are organized as follows: Chapter 2 discusses related work on active learning that has been proposed in the literature, Chapters 3, 4, 5 and 6 describe the proposed contributions in details and present the results obtained. Chapter 7 depicts the generalizibility of the proposed batch mode active learning framework by extending it to related problems (like learning from multiple sources of information and context aware learning); Chapter 8 details an online active learning algorithm based on

the Conformal Predictions (CP) Theory. Finally, Chapter 9 concludes with discussions and pointers to future work.

Chapter 2

RELATED WORK

In this chapter, we present a detailed survey of the different active learning algorithms that have been proposed in the literature. Active learning can be categorized broadly as shown in Figure 2.1. At the highest level, we can divide such methods into two types - *pool-based* and *online*. Pool based active learning is further divided into *Serial Query based Active Learning* and *Batch Mode Active Learning*. A comprehensive review of these categories can be found in [99].

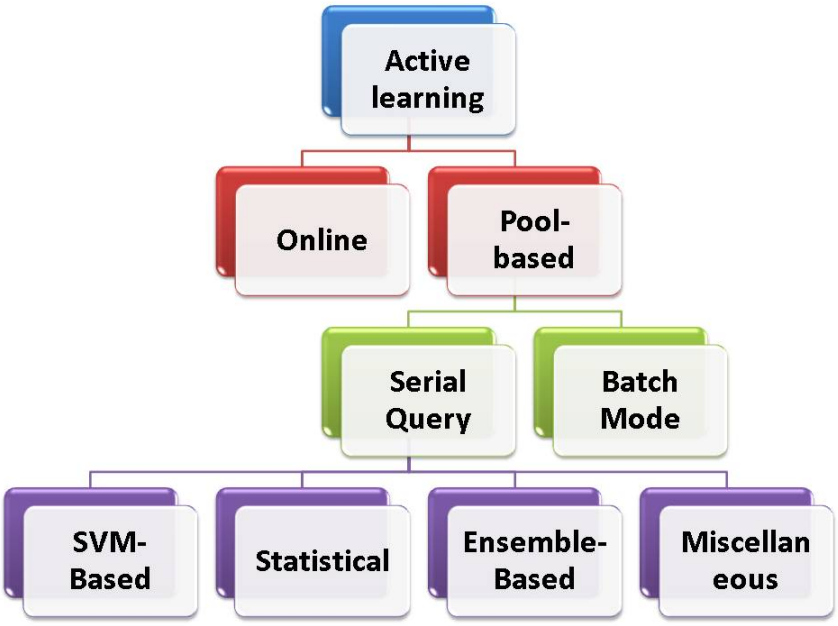


Figure 2.1: Categories of Active Learning

In a pool-based setting, the learner is exposed to a pool of unlabeled instances and it iteratively selects one point at a time from the pool to be labeled manually. This is continued until the pool gets exhausted or some stopping criterion is satisfied. In contrast, in an online setting, the learner does not have access to the entire unlabeled pool at once, but

encounters the points sequentially over a period of time. At each instant, the model has to decide whether to query the given point and update the hypothesis. We now present a review of the online active learning approaches, followed by the more popular serial query pool based active learning algorithms. We subsequently present existing work on batch mode active learning, which is the main focus of this dissertation.

2.1 Online Active Learning

Sculley [100] proposed an online active learning framework to develop an automated spam email filter system. The author used three online active learning techniques - label efficient *b*-sampling, logistic margin sampling and fixed margin sampling and concluded that they can dramatically reduce the labeling cost to design a spam filter. Bianchi *et al.* [101] provided regret bounds on an active learning algorithm for learning linear thresholds from an i.i.d. stream of samples. Monteleoni and Kaariainen [102] analyzed the performances of two online active learning algorithms in the optical character recognition problem. The algorithms - DKM and CBGZ and their combined variants were seen to consistently outperform random sampling. Dredze and Crammer [103] proposed an online active learner for natural language processing, where the distance of a point from the margin in a large-margin classifier was combined with parameter confidence. More recently, Ho and Wechsler [104] used the transductive confidence machine framework and the theory of conformal predictions to select query points in an online setting. The points were queried based on the difference between the top two $p-values$ computed using the conformal predictions framework. Balasubramanian *et al.* [105] proposed a generalized version of the above approach based on eigen-decomposition of matrices, where all the $p-values$ of a given point were incorporated to decide whether or not to query the particular point. The authors applied the transductive active learning approach in the online setting to face recognition. The query by committee (QBC) algorithm is popularly used

to query points in an online setting based on the level of disagreement among a group of classifiers. If the disagreement is above a certain threshold, the point is queried for its label. Melville *et al.* [106] used the Jensen Shannon divergence as a disagreement measure to select query points in such a setting. Attenberg and Provost [107] combined ideas from decision theory, cost-sensitive learning, online density estimation to develop a framework for active inference and learning in the online setting. Mesterharm and Pazzani [108] used online learning algorithms to solve active learning problems and provided performance bounds of the algorithm in both the online and batch settings. Chu *et al.* [109] proposed an online active learning framework based on variance minimization. Guillory and Bilmes [88, 110] proposed an online prediction version of the sub-modular set cover with connections to ranking and repeated active learning. In each round, the learning algorithm receives a sequence of items together with a monotone sub-modular function and suffers loss equal to the number of items needed, when items are selected in order of the chosen sequence, to achieve a coverage constraint. The fundamental challenge in online active learning is to design an appropriate query function without having access to the entire unlabeled set; that is, the query function needs to be implemented on each data instance as it arrives, without any knowledge of the samples that are to be encountered in future.

2.2 Pool based Active Learning with Serial Query

Majority of the existing active learning approaches have been applied in the pool based setting. These methods can be broadly categorized as follows: (i) SVM based methods, (ii) Statistical methods, (iii) Ensemble based methods and (iv) Other miscellaneous approaches.

SVM based methods

A sizable number of the pool based approaches are based on the Support Vector Machines (SVM) algorithm. Tong and Koller [111] [112] designed the query function to select the unlabeled point which is closest to the SVM decision boundary in the feature space. Tong and Chang [113] applied the same concept in the image retrieval problem where in every iteration, the point that was closest to the decision boundary was returned for labeling. Schohn and Cohn [114] applied active learning with SVMs in the document classification task and concluded that the classifier trained through active learning often outperforms those that were trained on all the available data. Mitra *et al.* [115] assigned a confidence c to examples within the current decision boundary indicating whether or not they were true support vectors. Points were then queried probabilistically according to this confidence factor. Another active learning scheme using SVMs was proposed by Campbell *et al.* [116] where the next point to be queried was the one which minimized a predetermined risk function. Cheng and Wang [117] used Co-SVMs in the image retrieval problem where two SVMs trained separately on color and texture features were used to classify unlabeled data - the points which were differently classified by the two SVMs were chosen to be labeled. Osugi *et al.* [118] proposed a probabilistic method of active learning which decided between labeling examples near the decision boundary and exploring the input data space for unknown pockets of points. Ebert *et al.* [119] analyzed different sampling criteria including a novel density based approach and corroborated the importance of combining exploration and exploitation for active learning. The authors also demonstrated that a time varying combination of sampling criteria often leads to improved performance. Loy *et al.* [120] presented a new unified framework for joint exploration and exploitation active learning without any heuristic weighting to learn from video streams.

Statistical methods

Statistical approaches quantify the informativeness of a data instance based on some statistical properties of the learner. Such methods can be further categorized as follows:

Uncertainty Sampling: The most commonly used query framework is uncertainty sampling, where a learner queries instances about which it is maximally uncertain. Uncertainty can be quantified in various ways - the expected $0/1$ loss, which is computed as 1 minus the maximum posterior probability under the current model, margin sampling [19] and the most popular Shannon's entropy [121]. Holub *et al.* [122] proposed an active learning framework that attempted to minimize the expected entropy of the labels of the data points in the unlabeled pool. MacKay [123] introduced information-theoretic approaches to active learning by measuring the informativeness of each data point within a Bayesian learning framework. Cohn *et al.* [124] described a rudimentary form of active learning which they called selective sampling. Here, the learner proceeded by examining the information already provided and then deriving a "region of uncertainty" where it believed misclassification was still possible. Ho and Wechsler [125] investigated a transductive framework to active learning where they used k nearest neighbors as the classifier. Li and Sethi [126] proposed an algorithm that identified samples that had more uncertainty associated with them, as measured by the conditional error. Tang *et al.* [127] used entropy based uncertainty scores to quantify the representativeness of a data point in a natural language parsing application, which was used to design the query function. Lewis and Gale [128] applied a probabilistic framework to active learning where the most uncertain point was chosen for manual annotation. Park *et al.* [129] developed a framework for optimal experimental design under the Gaussian Process Poisson model using uncertainty sampling to estimate the non-linear function of a neuron's stimulus sensitivity. Yan *et al.* [130] proposed a strategy to simultaneously learn the most uncertain samples and

the annotators to query the labels from for active learning, even when the expertise of the annotators may not be consistent across the task domain. Kowdle *et al.* [131] presented an uncertainty sampling based active learning approach for piecewise planar 3D reconstruction of a scene. Roder *et al.* [132] proposed an uncertainty based active learning approach where the class conditional probability $p(y|x)$ was viewed as a random variable and modeled using a second order distribution.

Expected Model Change: An active learning framework based on expected model change uses a decision-theoretic approach and selects the instance that would impart the greatest change to the current model if its label was known. An example query strategy in this framework is the expected gradient length (EGL) approach where the change imparted to the model is measured by the length of the training gradient. The learner queries the instance which, if queried and added to the training set, would result in the new training gradient of largest magnitude. This strategy was introduced by Settles *et al.* [133] and has also been applied to probabilistic sequence models like CRFs [17].

Expected Error Reduction: This type of active learning algorithms aim to quantify the amount of reduction of the generalization error. The idea is to estimate the expected future error of a model trained using $L \cup \{x, y\}$ on the remaining unlabeled instances in the unlabeled pool and query the instance with minimum expected future error (sometimes called the risk). Roy and McCallum [134] first proposed the expected error reduction framework for text classification using naive Bayes. The authors adopted a sampling approach to estimate the expected reduction in error due to the labeling of a query. The future error rate was estimated by log-loss using the entropy of the posterior class distribution on a sample of the unlabeled examples. Zhu *et al.* [86] combined this framework with a semi-supervised learning approach resulting in a dramatic improvement over random or uncertainty sampling. Guo and Greiner [135] employed an optimistic variant that biased the expectation towards the most likely label for computational convenience.

The framework had the dual advantage of being near-optimal and being independent of the model class. He and Cai [136] proposed a novel active subspace learning algorithm based on expected error reduction, which used the most informative data samples to learn an optimal subspace.

Variance Reduction: Minimizing the expectation of a loss function directly is expensive, and in general this cannot be done in closed form. However, we can still reduce generalization error indirectly by minimizing output variance, which sometimes has a closed-form solution. Consider a regression problem, where the learning objective is to minimize the standard error (i.e. squared-loss). The learner's expected future error can be decomposed as:

$$E_T[(\hat{y}-y)^2|x] = E[(y-E[y|x])^2] + (E_L[\hat{y}]-E[y|x]])^2 + E_L[(\hat{y}-E_L[\hat{y}])^2]$$

where E_L is an expectation over the labeled set L, $E[.]$ is an expectation over the conditional density $P(y|x)$ and E_T is an expectation over both; $\hat{y}$ is the model's predicted output for a given instance x and y indicates the true label for that instance. The first term represents the noise, the second term, the bias and the third term the variance of the model. Therefore, minimizing the variance is guaranteed to minimize the future generalization error of the model (since the learner can do nothing about the bias and or noise components). Cohn [137] and Cohn *et al.* [138] presented the first statistical analyses of active learning for regression using the estimated distribution of the models output. They showed that this can be done in closed-form for neural networks, Gaussian mixture models, and locally-weighted linear regression. Bellare *et al.* [139] proposed an active learning algorithm that approximately maximized the recall of a classifier under precision constraints with provably sub-linear label complexity for the task of entity matching.

Ensemble based methods

In ensemble based approaches, the Query by Committee (QBC) algorithm has been extensively applied. Seung *et al.* [140] proposed the first QBC approach by sampling a committee of two random hypotheses that are consistent with the current labeled set. Freund *et al.* [63], as well as Liere and Tadepalli [141] used the disagreement measure among a committee of classifiers to select points from an unlabeled pool. McCallum and Nigam [84] modified the Query by Committee method for estimating the document density while applying active learning to the text classification problem. They also combined active learning with Expectation Maximization to take advantage of the word co-occurrence information among the documents in the unlabeled pool. Melville and Mooney [24] proposed an ensemble based active learning method that encouraged diversity among committee members. Abe and Mamitsuka [142] combined QBC with boosting and bagging. The point to be queried next was the one on which the weighted majority voting by the current hypothesis had the least margin. Argamon and Dagan [143] proposed a Query by Committee algorithm in which the committee members were probabilistically selected from a distribution conditioned by the current training set. Muslea *et al.* [85] proposed a naive form of QBC, which they called *co-testing*, where an unlabeled point was randomly selected on which the existing views disagreed. Zhang and Sun [144] proposed a multile view multiple learner (MVML) active learning approach where the selecting sampling strategy was implemented in three ways - by choosing samples just considering the disagreement between the two views, just considering the disagreement within each view and considering both the within-view and between-view disagreements. Another form of QBC was proposed with the nearest neighbor classifier [145] [146] where each neighbor was allowed to vote on the class label of an unlabeled point, with the proportion of these votes representing the posterior label probability which was used as a disagreement

measure for instance selection. Gao and Koller [147] presented an active classification algoroithm where each classifier in a large ensemble is viewed as an observation which can influence the classification process. Observations were selected dynamically based on previous observations using a value-theoretic computation that balanced an estimate of the expected classification gain as well as its computational cost. Yang [148] studied the probelem of active learning in a stream based setting where the distribution of the samples changed over time and proved upper bounds on the number of prediction mistakes and number of label requests for disagreement based active learning algorithms.

Other miscellaneous approaches

In other kinds of pool based approaches, Baram *et al.* [149] proposed a master algorithm which estimated the progress of each active learner in an ensemble during a learning session and then dynamically switched over to the best performing one at each stage. Using three active learning algorithms (Simple, Kernel Farthest First and Self-Conf) to construct an ensemble, the authors empirically established that combining them online resulted in a better performance than using any one of them. Blum and Chawla [150] developed an algorithm based on graph-cuts to learn from both labeled and unlabeled data. Nigam *et al.* [151] combined the Expectation Maximization (EM) algorithm with naive Bayes classifier to learn from labeled and unlabeled text documents. Pelleg and Moore [152] proposed a mixture model approach to solve the problem of anomalous rare category identification in an unlabeled set with minimal human effort. Schein and Ungar [153] extended the A-optimality criterion to pool based active learning using Logistic Regression classifiers. Thompson *et al.* [22] applied the active learning framework to two non-classification tasks: semantic parsing and information extraction. They concluded that about 44% reduction in annotation cost was achieved using active learning in these complex tasks. Clustering techniques have also been used to boost the performance of pool-based active

learning [82] [154]. There have also been efforts in incorporating contextual information in active learning. Very recently, Kapoor *et al.* [155] incorporated *match* and *non-match* constraints in active learning for face recognition. Qi *et al.* [156] presented a 2D active learning scheme where sampling was done along both sample and label dimensions. The authors proposed to select sample-label pairs to minimize a multi-label Bayesian classification error bound. Kothari and Jain [157] proposed a genetic algorithm based active learning strategy to iteratively refine the class membership of the unlabeled patterns so that the maximum a posteriori (MAP) based predicted labels of the points in the labeled dataset were in agreement with the known labels. Joshi *et al.* [158] proposed an active learning algorithm, where two data points were selected in each iteration, one from the training set and one from the unlabeled set and the user had to give feedback regarding whether or not the two samples belonged to the same class. This was the first effort in designing an active learning framework where the user feedback was binary (yes/no) type. Guo and Greiner [135] proposed an active learning approach that exploited the discriminative partition information contained in the unlabeled instances and selected the query instance that provided the maximum conditional mutual information about the labels of the unlabeled samples given the labeled data. Tong and Koller [159] proposed an active learning framework for parameter estimation in Bayesian networks.

Park and Pillow [160] presented a novel active learning scheme under hierarchical, conditionally Gaussian priors that uses sequential Markov Chain Monte Carlo sampling to model a mixture-of-Gaussians representation of a neuron's receptive field (RF) field and selects optimal stimuli using an approximate infomax criterion. Osborne *et al.* [161] proposed several techniques for evidence estimation using the Bayesian Quadrature method, including approximately imposing a positivity constraint, approximately marginalizing hyper-parameters and using active sampling to select the locations of function evaluations and concluded that the active learning approach yielded the most significant gains for in-

tegral estimation. Tyagi and Cevher [162] proposed a randomized active sampling scheme for estimating multi-index functions of the form $f(x) = g(Ax)$ from point evaluations of f, where the function f is defined on the ℓ_2-ball in $\Re^d$, g is twice continuously differentiable almost everywhere and $A \in \Re^{k \times d}$ is a rank k matrix where $k << d$. Sawade *et al.* [163] devised an active comparison algorithm (for comparing the risks of two given prediction models) that selects instances according to a sampling distribution which maximizes the power of a statistical test applied to the observed empirical risks and thereby minimzes the likelihood of selecting the inferior model. The same authors [164, 165] further proposed an active estimation procedure based on variance minimization to estimate the F_α-measure of a given model on a fixed labeling budget. Ailon [166] proposed an active learning framework for pairwise ranking using a query (and time) efficient decomposition procedure reducing the problem to smaller sub-problems in which the optimal loss was high and uniform sampling sufficed. Jamieson and Nowak [167] examined the same problem and proposed an active learning algorithm that exploited the natural relationships among the objects to reduce the number of pairwise comparisons as compared to standard sorting based methods ($n \log_2 n$). Along similar lines, Charlin *et al.* [168] addressed the problem of active learning of user preferences for matching problems by introducing a novel method for determining probabilistic matchings and developing strategies that are sensitive to the specific matching objective. Jain *et al.* [169] proposed two hashing based solutions (to retrieve near points in sub-linear time) for pool-based active learning and empirically demonstrated the practicality of the algorithm to perform active selection with millions of unlabeled samples. Garnett *et al.* [170] used Bayesian decision theory to solve the problems of active search (actively uncover as many members of a given class as possible) and active surveying (to actively query points to ultimately predict the class proportion of a given class). Rashidi and Cook [171] proposed a novel active learning method called RIQY (Rule Induced active learning QuerY), which can construct generic active learning queries based on rule induction from multiple unlabeled instances. Ju-

dah *et al.* [172] introduced a new approach based on reducing active imitation learning to i.i.d active learning. Deng *et al.* [173] proposed a novel minimax bandit model for active learning in the context of personalized treatment via biomarkers. Active learning has been extensively used to address problems in multimedia and computer vision such as video / image annotation [174, 175, 176, 177, 178], video search [179], image segmentation [180] and scene understanding [181] among others. Recent efforts in this area have included a novel min-max approach to systematically combine multiple criteria (such as informativeness and representativeness) for active sample selection [182], to appropriately consider the information overlap across different domains for batch selection [183], applying active learning for link classification in signed networks [184, 185, 186, 187] as well as application of active learning approaches to rapidly improve a multi-task adaptive filtering system with minimal user/task-level feedback [188]. At the AISTATS 2010 challenge on active learning, several novel methodologies were proposed to address practical challenges like large, noisy data, irrelevant attributes, missing values and mixed variable types [189, 190, 191, 192, 193].

Salganicoff *et al.* [194] applied active learning to the vision based grasping problem. The authors combined the Interval Estimation (IE) active learning approach with the classification tree algorithm ID-3 to develop a system which actively learns to select the grasp approach directions. Morales *et al.* [195] applied active learning to measure the grasping reliability. The algorithm accumulated the information gathered through successive grasping attempts and chose the best configuration to grasp a given object. Dima [196] proposed the Unlabeled Data Filtering (UDF) algorithm to solve the initialization problem in active learning in robotics. Zhang and Kim [197] proposed an active learning based path planner to plan the optimal path between a source and a destination in a path planning application. The system learned incrementally and developed its knowledge to plan suitable paths in real time. Cantin *et al.* [198] developed an active policy learning

approach to be used in an exploration application. The method balanced exploration and exploitation and used probabilistic active learning algorithm to predict the policies that would produce higher expected gains. Dima *et al.* [199] described an active learning algorithm based on kernel density estimation to identify exemplar images in a dataset. They applied the concept to the problem of terrain classification and obstacle detection by autonomous outdoor robots and concluded that the algorithm achieved comparable performance in accuracy by labeling only a few sample and informative images. Tapus and Mataric [200] developed active learning strategies to help stroke patients. Wiens and Guttag [201] applied active learning algorithms to perform patient-adaptive and task-adaptive heartbeat classification. Burl and Wang [202] applied active learning to study the behavior of complex systems using physics based simulation codes. However, all these approaches have been based on serial query strategies; we now review existing work on batch mode active learning, which is the primary focus of this thesis.

2.3 Batch Mode Active Learning (BMAL)

As mentioned in Section 1.8, batch mode active learning algorithms are effective in utilizing the presence of multiple labeling oracles and avoiding frequent classifier training, as they select a batch of points simultaneously for manual annotation. Existing approaches for BMAL have largely been based on extending pool-based active learning methods to select multiple instances simultaneously. They use greedy heuristics and select the top k instances (k being the required batch size) from the unlabeled set for manual annotation. Brinker [203] extended the version space concept proposed in [111] to query a diverse batch of points using SVMs, where diversity was measured as the angle induced by the hyperplane of the currently selected point to the hyperplanes of the already selected points. Ding *et al.* [204] used cluster diversity and most possible error approximation bound as the batch selection criteria. Zhang *et al.* [205] proposed a BMAL scheme that selected a

diverse batch of points using the farthest-first traversal strategy. Schohn and Cohn [114] proposed to query a batch of points based on their distance from the separating hyperplane for a linear SVM. Xu *et al.* [206] proposed an SVM based BMAL strategy which combined representativeness and diversity measures for batch selection. Demir *et al.* [207] proposed an active learning algorithm for classification of remote sensing images using a kernel clustering based strategy to assess the diversity and informativeness of samples from each cluster. Ananthakrishnan *et al.* [208] presented a BMAL scheme for machine translation systems that attempted to maximize the in-domain coverage by selecting sentences which represent a balance between domain match, translation difficulty and batch diversity. Shi *et al.* [209] proposed three criteria (minimum redundancy, maximum uncertainty and maximum impact) to exploit the link based dependencies in a network and actively select a batch of instances for user query.

However, extending the pool-based setting to the batch setting by considering the top k instances does not account for other factors such as information overlap between the selected points in a batch. More recently, this has led to newer efforts that are specifically intended to select batches of points using appropriate optimization strategies. Hoi *et al.* [95, 97] used the Fischer information matrix as a measure of model uncertainty and proposed to query the set of points that maximally reduced the Fischer information. The batch selection criterion was formulated as a trace norm minimization problem:

$$\min_{q,M} trace(M)$$

s.t.:

$$M \geq I_p^{1/2} I_q^{-1} I_p^{1/2}$$

$$\sum_{i=1}^{n} q_i, q_i \geq 0, i = 1, \ldots n$$

Here, $p(x)$ was the distribution of all unlabeled examples and $q(x)$ was the distribution of the unlabeled examples that were chosen for labeling, I_p and I_q were the Fischer informa-

tion matrices of the classification model for distributions $p(x)$ and $q(x)$ respectively and M was a slack matrix to upper bound the objective function. The optimization problem was solved using semidefinite programming (SDP). The same authors [210] proposed a BMAL scheme based on SVMs where a kernel function was first learned from a mixture of labeled and unlabeled samples, which was then used to identify the informative and diverse examples through a min-max framework. Joshi *et al.* [211] introduced a batch mode active learning framework using submodular functions for multi-class image classification. The authors combined the uncertainty and diversity criteria into a submodular objective function, which was solved using an iterative greedy algorithm. Shi and Zhao [212] proposed a unified framework integrating sparse representation and batch mode active learning. Based on the existing sparse family of classifiers, the authors defined the corresponding batch mode sparse active learning family and explored their shared properties. Azimi *et al.* [213] proposed a batch mode active learning algorithm that first used the Monte Carlo simulation to estimate the distribution of the unlabeled samples and then attempted to select a batch of k instances that best matched this distribution. Zhao *et al.* [214, 215] proposed a graph-based transductive BMAL framework based on label propagation. Vijayanarasimhan *et al.* [216] formuated the BMAL problem as a continuous optimization where the subset of possible queries was determined that maximized the improvement to the classifier's objective without exceeding a specified budget. Guo and Schuurmans [59] proposed a discriminative strategy that selected a batch of points which maximized the log-likelihoods of the selected points with respect to their assigned class labels and minimized the entropy of the unselected points in the unlabeled pool. Specifically, the algorithm solved for a binary matrix μ, which optimistically assigned class labels to the unlabeled points and simultaneously decided the points to be selected in the batch, through the following objective function:

$$\max_{\mu} \sum_{i \in L_t} \log P(y_i|x_i, w^{t+1}) + \beta \sum_{j \in U_t} v_j^{t+1} \mu_j^T - \alpha \sum_{j \in U_t} (1 - \mu_j e) H(y|x_j, w^{t+1})$$

s.t.:

$$\mu \in \{0,1\}^{|U_t| \times 2}$$

$$\mu \circ E = m$$

$$\mu_j e \leq 1, \forall j$$

$$\mathbf{1}^T \mu \leq (\frac{1}{2} + \varepsilon) m e^T$$

Here, L_t and U_t were the current training and unlabeled sets at time t, v_j^{t+1} was a row vector $[\log P(y = 1|x_j, w^{t+1}), \log P(y = -1|x_j, w^{t+1})]$, e was a two entry column vector of 1s, $\mathbf{1}$ is a $|U_t|$ entry column vector with all 1s, E was a $|U_t| \times 2$ matrix with all entries 1, ε was a user-provided parameter that controlled class balance during instance selection and β was a parameter that was used to adjust the belief in the guessed labels. The selection variable μ chose instances from U_t and also selected labels for the selected instances. Solving this optimization yielded the optimal μ for instance selection for iteration $t+1$.

Very recently, Guo [217] proposed a batch mode active learning scheme which maximized the mutual information between the labeled and unlabeled sets and was independent of the classification model used. Let L and U be the current labeled and unlabeled sets and Q be a set of cardinality b denoting the set of points that are selected. The selection strategy was formulated as the following optimization problem:

$$Q^* = \arg \max_{|Q|=b, Q \subseteq U} I(X_{L \cup Q}, X_{U \setminus Q}) = \arg \max_{|Q|=b, Q \subseteq U} \log |\Sigma_{L'L'}| + \log |\Sigma_{U'U'}|$$

where $L' = L \cup Q$ and $U' = U \setminus Q$. The mutual information criterion depended only on the covariance matrices computed using the kernel functions over the instances. The maximum mutual information strategy attempted to select the batch of b instances from the unlabeled set U to label, to maximize the log determinants of the covariance matrices over the produced sets L' and U'. The methods described in [59] and [217] have well defined

mathematical basis and have been shown to be the best performing BMAL schemes till date [217].

All the aforementioned techniques of batch mode active learning assume that the batch size (number of data points to be queried from an unlabeled set to be specified in advance. This may not be a practical assumption as it is difficult to decide on a number at random and without any knowledge of the data stream in question. Moreover, in many real world applications, the label of each data point is fuzzy, that is, it is possible for one point to belong to multiple classes with varying degrees. To the best of our knowledge, no BMAL technique has been proposed to explicitly handle fuzzy label problems. Further, the state-of-the-art BMAL schemes [59, 217] solve the batch selection problem by convex relaxations of NP-hard integer programming problems. Even though they have been empirically shown to demonstrate good performance, no formal guarantee has been established on the qualities of the convex relaxations. Recently, the problem of low rank matrix completion has gathered significant attention and is being used extensively in applications like machine learning [218], computer vision and graphics [219] and recommendation systems [220] among others. However, the problem of intelligently integrating human expetise in completing a matrix has not been explored till date. This can potentially lead to a better reconstruction of the incomplete matrix and can be of immense practical importance. In order to address these practical issues, we propose four major contributions in this Ph.D dissertation:

1. A framework for dynamic batch mode active learning, where the batch size and the specific points to be selected for manual annotation are simultaneously derived through a single formulation.

2. A BMAL algorithm for fuzzy label classification problems, where there is an inherent imprecision and vagueness in the class label definitions.

3. Batch Mode Active Learning algorithms based on the convex relaxations of an NP-hard integer quadratic programming (IQP) problem, with guaranteed bounds on the solution quality.

4. An active matrix completion framework to leverage human intelligence in completing a data matrix and its application in different variants of the active learning problem (transductive active learning, multi-label active learning, active feature acquisition and active learning in regression).

The contributions are validated on the face recognition and facial expression recognition applications. Reliable face and expression recognition are of paramount importance in a Social Interaction Assistant technology, as detailed in Section 1.1. Automated recognition of identity and facial expression of a subject can enable a visually challenged individual recognize his interaction partner and better understand his/her emotional state, which in turn can facilitate effective social interaction. We detail each of the aforementioned contributions in the subsequent chapters. We also study the generalizibility of our approach by extending it to related problems in machine learning like learning from multiple sources of information and context aware learning.

Chapter 3

DYNAMIC BATCH MODE ACTIVE LEARNING

An ideal batch mode active learning (BMAL) system can be conceptualized as consisting of two main steps: (i) deciding the batch size (number of image frames to be queried from a given unlabeled video stream) and (ii) selecting the most appropriate images from the unlabeled video once the batch size has been determined. Both these steps are critical in ensuring maximum generalization capability of the learner with minimum human labeling effort, which is the primary objective in any active learning application. However, the existing few efforts on batch mode active learning focus only on the second step of identifying a criteria for selecting informative batches of data samples and require the batch size to be specified in advance by the user [59, 217]. In a real world application, deciding on the batch size (number of relevant instances in a data stream) in advance and without any knowledge of the data stream being analyzed, may not lead to a good generalization accuracy. The batch size should depend on the quality and complexity of the samples in the unlabeled stream and also on the level of confidence of the current classifier on the unlabeled data instances. In other words, there is a strong need for dynamic batch selection in BMAL algorithms.

In this chapter, we present two novel batch mode active learning algorithms which adaptively select samples for manual annotation based on the complexity of the data stream being analyzed and the cost of labeling each unlabeled data sample. We develop a formulation for dynamic batch selection which directly optimizes the performance of the updated learner (the learner trained on the current training set together with the newly selected batch). The batch selection problem is solved using the stochastic gradient descent algorithm to simultaneously decide the batch size and identify the specific points that need

to be queried for manual annotation, through a single framework. We also derive a second formulation for dynamic batch selection based on the uncertainty of the current learner. We exploit the properties of sub-modular functions and propose an efficient solution strategy for adaptive batch selection through a single optimization framework. Due to its wide usage, we focus on face based biometric recognition systems as the exemplar application in this paper. Although validated on biometric data, the proposed frameworks are generic and can be used in any application where it is required to select a number of representative entities simultaneously from repetitious samples.

3.1 Clustering based Batch Size Selection : An Intuitive Approach

An intuitive strategy to decide the batch size dynamically is to use a clustering algorithm to segregate the images in the unlabeled pool into relatively pure clusters (in terms of class labels), followed by a method to compute the batch size. Since the number of subjects (and hence the number of clusters) in an unlabeled set is an unknown, we need to exploit the spatial distribution of the points for clustering. This motivates the usage of the DBSCAN algorithm to automatically isolate the high density regions of the unlabeled pool into separate clusters. For details about this method, please refer [221]. Our initial experiments confirmed the efficacy of DBSCAN in isolating the images of different subjects into separate clusters.

The Silhouette Coefficient (based on the cohesion and separation measures of a cluster) is a natural choice to decide the number of points to be queried from each cluster. For the i^{th} point in a cluster, the Silhouette Coefficient is defined as

$$s_i = \frac{(b_i - a_i)}{max(a_i, b_i)} \tag{3.1}$$

where a_i is the average distance of the i^{th} point to all objects in its cluster and b_i is the minimum of the average distances of the i^{th} point to each of the other clusters. The Silhouette

coefficient for an entire cluster can then be computed as the average of the coefficients of each point forming the cluster. It can attain a maximum value of 1, where a high value denotes a compact and well separated cluster. Intuitively, we would like to select few points for a compact and well-separated cluster and more points otherwise. Thus, the number of points to be selected from a cluster should be proportional to (1 - the Silhouette coefficient). Also, we would like to select more points from larger clusters. If m is the total number of points, m_i is the number of points in cluster i, SC_i is the Silhouette coefficient of cluster i and C is a constant, the number of points to be selected from cluster i can thus be defined as:

$$N_i = C * \frac{m_i}{m} * (1 - SC_i) \tag{3.2}$$

This operation is performed for each of the identified clusters to compute the number of points to be selected (the sum of the values obtained across all clusters provides the overall batch size). The dynamically computed batch size for each cluster can now be passed as an input to any standard static BMAL procedure for selecting the required number of points from the corresponding cluster.

The clustering based strategy for dynamic batch size selection suffers from two main drawbacks - (i) it computes the batch size solely based on the spatial distribution of the points in the unlabeled pool; the current training set is not considered in the batch size calculation. Ideally, the knowledge available from the current training data should play a role in deciding the batch size. (ii) It is a two step process, where the batch size is first computed from the cluster structure of the data and then supplied as an input to a standard batch mode active learning algorithm for the point selection process and thereby involves significant computation. To overcome these limitations, we propose two dynamic BMAL frameworks in the following sections.

3.2 Dynamic Batch Mode Active Learning via Stochastic Gradient Descent (SGD)

Consider a BMAL setting which has a current labeled set L_t and a current classifier w^t trained on L_t. The classifier is exposed to an unlabeled video U_t at time t. The objective is to select a batch B from the unlabeled stream in such a way that the classifier w^{t+1}, at time $t+1$, trained on $L_t \cup B$ has maximum generalization capability (we refer to w^{t+1} as the "future model" or "future classifier"). With unlabeled data being available, semi-supervised learning methods have been proposed that train models by minimizing the uncertainty of the labels for the unlabeled instances [222]. That is, to achieve a classifier with good generalization performance, one can minimize the entropy of the missing labels for the unlabeled data. In our active learning framework, we attempt to minimize the entropy of the updated learner on the remaining $|U_t - B|$ images after batch selection. Let C denote the total number of classes. The entropy of the conditional distribution $P(y|x_j, w^{t+1})$ is given by:

$$S(y|x_j, w^{t+1}) = -\sum_{y \in C} P(y|x_j, w^{t+1}) \log P(y|x_j, w^{t+1}) \tag{3.3}$$

Further, to maximize the contribution of the selected unlabeled samples, diversity based selection criteria have been proposed [223] which ensure that the selected samples are less similar with the already available labeled data. In our formulation, we quantify the diversity, ρ_j, of an unlabeled sample x_j as its mean kernelized distance from all the labeled points in the training set:

$$\rho_j = \frac{1}{n_l} \sum_{i=1}^{n_l} \phi(x_i, x_j) \tag{3.4}$$

where n_l is the number of samples in the training set and ϕ denotes the kernel function. Such a distance measure has a concrete theoretical grounding and is popularly used in metrics like the Maximum Mean Discrepancy (MMD) to quantify the difference between two probability distributions [224, 225]. In our experiments, a Gaussian kernel with pa-

rameter 1 was used as the underlying kernel. The two conditions mentioned previously can be satisfied by defining a score function as follows:

$$f(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) \tag{3.5}$$

The first term denotes the sum of the average kernelized distances of each selected unlabeled point from the labeled set, while the second term quantifies the sum of the entropies of the updated learner on each remaining point in the unlabeled stream. λ_1 is a tradeoff parameter governing the relative importance of the two terms.

The problem therefore reduces to selecting a batch B of unlabeled images which produces the maximum score f(B). Let the batch size (number of images to be selected for annotation) be denoted by m, which is an unknown. Since there is no restriction on the batch size m, the obvious solution to this problem is to select *all* the images in the unlabeled video, leaving no image behind. Then, the entropy term becomes 0, and the distance term attains its maximum value. Consequently, f(B) will also attain its maximum score. However, querying all the images for their class labels is not an elegant solution and defeats the basic purpose of active learning. To prevent this, we modify the score function by enforcing a penalty on the batch size as follows:

$$\tilde{f}(B) = \sum_{j \in B} \rho_j - \lambda_1 \sum_{j \in U_t - B} S(y|x_j, w^{t+1}) - \lambda_2 m \tag{3.6}$$

The third term essentially reflects the cost associated with labeling the images, as the value of the objective function decreases with every single image that needs to be labeled. Defining the score function in this way ensures that any and every image is not queried for its class label; only images for which the distance and entropy terms outweigh the labeling cost term, get selected. The coefficient λ_2 is the cost parameter and denotes the cost associated with labeling one unlabeled data sample. This parameter can be set based on the given application. For instance, manually labeling a face image is less tedious as

compared to labeling a voicemail message as urgent/non-urgent (as the human oracle has to listen to the entire message for accurate annotation). Thus, λ_2 will have a smaller value in case of a face recognition application, as compared to a voicemail recognition system. In our experiments, we assume λ_2 to be 1 and also explore the effect of this parameter on the batch size and the accuracy of recognition.

As per Equation (3.6), we need to select a batch B of unlabeled images so as to maximize $\tilde{f}(B)$. Since brute force search methods are prohibitive, we employ numerical optimization techniques to solve this problem. We define a binary vector M of size $|U_t|$ where each entry denotes whether the corresponding point is to be queried for its class label. We rewrite the objective function in Equation (3.6) into an equivalent function in terms of the defined vector M:

$$\max_{M,m} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 m \tag{3.7}$$

s.t.:

$$M_j \in \{0,1\}, \quad \forall j \tag{3.8}$$

In this formulation, note that if an entry of M is 1, the corresponding image will be selected for annotation and if it is 0, the image will not be selected. The number of images to be selected, is therefore equal to the number of non-zero entries in the vector M, or the zero-norm of M. Hence,

$$m = ||M||_0 \approx ||M||_1 = \sum_j M_j \tag{3.9}$$

Here, we have replaced the zero norm of M by its tightest convex approximation, which is the one-norm of M (similar to [226]). Also, from constraint 3.8, the one-norm is simply the sum of the elements of the vector M. Substituting m in terms of M, the formulation becomes:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{3.10}$$

s.t.:

$$M_j \in \{0,1\}, \quad \forall j$$

The above optimization is an integer programming problem and is NP hard. We therefore relax the constraint to make it a continuous optimization problem:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{3.11}$$

s.t.:

$$0 \leq M_j \leq 1, \quad \forall j$$

Solving the Optimization Problem

We first define an objective function $f(M)$ as:

$$f(M) = \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) - \lambda_2 \sum_j M_j \tag{3.12}$$

To solve the optimization problem, we use the Quasi Newton method, which assumes that the function can be approximated as a quadratic in the neighborhood of the optimum point and iteratively updates the variable M to guide the functional value towards this local optima. The first derivative of the function and the Hessian matrix of second derivatives need to be computed as parts of the solution procedure. Assuming w^{t+1} remains constant with small iterative updates of M, the first order derivative vector is obtained by taking the partial of the objective with respect to M:

$$\nabla f(M_j) = \rho_j + \lambda_1 S(y|x_j, w^{t+1}) - \lambda_2 \tag{3.13}$$

The Hessian starts as an identity matrix and is updated according to the BFGS method. In each iteration, a quadratic programming problem is solved which yields an update direction for M. The step size is obtained using a backtrack line search method based on the

Armijo Goldstein equation and guarantees monotonic convergence of the function to the local optimum. The iterations are continued until the change in the value of the objective function is negligible. The final value of M is used to govern the number of points and the specific points to be selected for the given data stream (by greedily setting the top m entries in M as 1 to recover the integer solution, where $m = \sum_j M_j$). Hence, solving a single optimization problem helps in dynamically deciding the batch size as well as selecting the specific points for manual annotation. For further details about the Quasi Newton method, please refer [227].

It is to be noted that the objective function is defined in terms of the future classifier w^{t+1}, which is unknown. To compute the entropy term using w^{t+1} in the Quasi Newton iterations, we therefore need to estimate the class labels of the currently selected batch of unlabeled samples so as to intelligently approximate w^{t+1}. We used the semi-supervised graph-based label propagation method GTAM (Graph Transduction via Alternating Minimization) proposed by Wang *et al.* [228] to derive the labels of the selected unlabeled samples in each Quasi-Newton iteration. This method is efficient in terms of accuracy and computational overhead [228]. We validate the efficiency of this method in our empirical evaluations. The pseudocode of the complete dynamic BMAL algorithm is outlined in Algorithm 2.

We also note that the specific terms in the objective function can be modified based on the particular application in question. For instance, one may want to design an objective function which selects samples by minimizing the uncertainty on the unselected examples and by maximizing the representativeness between the selected and the unselected samples in the unlabeled set. The same strategy based on a penalty on the batch size can be used in the objective function containing the relevant terms.

The proposed dynamic batch selection framework has the computational complexity of $O(n^2)$ (where n is the number of unlabeled data samples), which is the same as the

Algorithm 2 Dynamic Batch Mode Active Learning via Stochastic Gradient Descent (SGD)

Require: Training set L_t, Unlabeled set U_t, parameters λ_1 and λ_2, initial random guess for M, a stopping threshold α

Initialize the Hessian matrix H as the identity matrix I
Evaluate the objective function $f(M)$ (Equation 3.12) and the derivative vector $\nabla f(M)$ (Equation 3.13)
repeat
    Solve the QP problem as required by Quasi-Newton: QP(H,$\nabla f(M)$,M) and let the solution be M^*
    Compute the step size s from the Armijo Goldstein Equations.
    Update M as $M_{new} = M + s(M^* - M)$
    Evaluate the new objective $f(M_{new})$ and the new derivative vector $\nabla f(M_{new})$ using M_{new}
    Calculate the difference in objective value: $diff = abs(f(M)$ - $f(M_{new}))$
    Update the Hessian H using the BFGS Equations
    Update the objective value: $f(M) = f(M_{new})$
    Update the derivative vector: $\nabla f(M) = \nabla f(M_{new})$
    Update the vector M: $M = M_{new}$
until $diff \leq \alpha$
Compute batch size $m = \sum M$ (Equation 3.9)
Greedily set the top m entries in M as 1 to recover the integer solution.
Select m points accordingly

state-of-the-art static BMAL techniques [59, 217], where the batch size needs to be pre-specified. Thus, with the same computational complexity as state-of-the-art static BMAL schemes, we solve for both the batch size and the specific data samples that need to be queried from a given unlabeled data stream.

3.3 Dynamic Batch Mode Active Learning via Submodular Optimization

In this section, we present another novel dynamic batch mode active learning scheme based on sub-modular optimization. Similar to the previous problem, we are given a training set L_t and an unlabeled set U_t for adaptive batch selection. In this method, the uncertainty of an unlabeled sample is computed as the entropy of the current model w^t on this sample (instead of the updated model w^{t+1}, as in the previous formulation). However, since the goal in active learning is to select a batch of unlabeled samples that are maximally

informative for the updated model w^{t+1}, we need to consider a redundancy based criterion (which quantifies the similarity between a pair of samples) if we design the batch selection condition based on the current model w^t. This is because, if two points separately furnish valuable information, but they furnish the same / overlapping information, then both of them together may not be maximally informative for w^{t+1}. The redundancy criterion is important in this formulation, as the objective is to select a batch of useful samples for w^{t+1} using only the current model w^t. This was not necessary in the previous formulation as the performance was directly optimized with respect to the future model w^{t+1}.

In this work, redundancy was quantified as the minimum kernelized distance of an unlabeled point from the already selected batch (other measures of distance or similarity may be used based on the application in question). A greater value of the minimum distance denotes a more promising point from the redundancy perspective. We would like to select a batch of points where each point furnishes useful, but distinctly unique information (we note that in the gradient descent based formulation, the performance was optimized directly with respect to the future classifier w^{t+1} and so the redundancy based term was unnecessary. In the present approach, the performance is quantified in terms of the current model w^t and hence we need the redundancy based term to ensure that we do not select duplicate data samples). For this purpose, we formulate an objective function denoting the score of a set of points B as follows (we have the term quantifying the kernelized distance from the training set, as before):

$$S(B) = \sum_{x_i \in B} \{\rho_i + \lambda_1 E(x_i) + \lambda_2 D(x_i)\} \tag{3.14}$$

where ρ_i is the average kernelized distance of the unlabeled point x_i from the training set, as defined in Section 3.2, $E(x_i)$ is the entropy of x_i based on the current model w^t:

$$E(x_i) = -\sum_{y \in C} P(y|x_i, w^t) \log P(y|x_i, w^t)$$

and

$$D(x_i) = \min_{x_j \in B: j \neq i} < x_i, x_j >$$

which quantifies the similarity of an unlabeled point from the already selected set ($<,>$ denotes the kernelized distance). λ_1 and λ_2 are tradeoff parameters controlling the relative importance of the distance and entropy terms. Since the goal is to select a batch of points with high aggregate uncertainty scores and high distance among them, the objective is to select a set of points which maximizes the score $S(B)$ as defined in Equation (3.14). This score function is monotonically non-decreasing (will be proved later) and since there is no restriction on the batch size, the obvious solution is to select all points in the unlabeled set for manual annotation. Similar to the previous formulation, we therefore impose a penalty on the batch size and modify the score function as follows:

$$S^{new}(B) = \sum_{x_i \in B} \{\rho_i + \lambda_1 E(x_i) + \lambda_2 D(x_i)\} - \lambda_3 |B| \tag{3.15}$$

The last term in Equation (3.15) represents the cardinality of the set B and increases as more points are queried in the batch. λ_3 is the cost parameter which denotes the cost of annotating each unlabeled sample (as discussed earlier). The optimal batch selection criterion can thus be expressed as:

$$\max_{B \subseteq U_t} S^{new}(B) \tag{3.16}$$

Due to the exponential nature of the search space, exhaustive search techniques are not feasible. In the following sections, we derive an efficient strategy to solve the above optimization problem.

Submodularity of the Objective Function

Let Z be a finite set and let $X \subseteq Y \subseteq Z$ be two subsets of Z. Consider an element $x \in Z \backslash Y$. A function $f : 2^Z \rightarrow \Re$ is submodular if

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

That is, a function is submodular if adding an element to a set increases the functional value by at least as much as adding the same element to its superset. This property is called the diminishing returns property [229, 230].

Lemma 1. *The score function $S(B)$, as defined in Equation (3.14), is a submodular set function.*

Proof. Let B_1 and B_2 be two sets formed by selecting unlabeled points from U_t, such that $B_1 \subseteq B_2 \subseteq U_t$ and consider an unselected instance $x \in U_t \backslash B_2$. The increment in the value of the objective function achieved by appending x to the set B_1 is given by

$$S(B_1 \cup \{x\}) - S(B_1) = \rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_1} < x, x_j >$$

Similarly, the increment obtained by appending x to the set B_2 is:

$$S(B_2 \cup \{x\}) - S(B_2) = \rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_2} < x, x_j >$$

Since, $B_1 \subseteq B_2$, the minimum distance of a point x from the other points will always be greater for the set B_1 as there may exist some point x_j in the superset B_2 which is closer to x than any element in its subset B_1. Hence,

$$\min_{x_j \in B_1} EuclidDist(x, x_j) \geq \min_{x_j \in B_2} < x, x_j >$$

Thus, we have,

$$S(B_1 \cup \{x\}) - S(B_1) \geq S(B_2 \cup \{x\}) - S(B_2)$$

This completes the proof of the lemma.

□

Lemma 2. *The score function $S(B)$ is a monotonically non-decreasing function.*

Proof. Let B_1 denote the currently selected set of points and consider an element $x \in U_t \backslash B_1$, U_t being the unlabeled pool. If x is added to the current set, the value of the objective function changes by $\rho_x + \lambda_1 E(x) + \lambda_2 \min_{x_j \in B_1} EuclidDist(x, x_j)$. Both the entropy and distance are non-negative quantities and hence,

$$S(B_1 \cup \{x\}) \geq S(B_1)$$

This completes the proof.

□

Greedy Solution to the Optimization Problem

The problem of maximizing a submodular function is NP-hard. However, Nemhauser *et al.* [229] established that for a function S, which is submodular and non-decreasing, with $S(\phi) = 0$, a greedy algorithm provides an efficient solution with near-optimal results (from the definition of S in Equation (3.14), it is obvious that $S(\phi) = 0$). The greedy algorithm incrementally selects points from the unlabeled set by maximizing the gain in the objective function in each iteration. It presents an incremental ordering of the samples based on their degree of usefulness. A single run of the algorithm over the unlabeled set therefore provides an ordered set of the unlabeled samples based on their information content. The final objective value $S^{new}(B)$ is then computed for every possible batch size by subtracting the weighted set cardinality $\lambda_3 * |B|$ from the corresponding score $S(B)$.

The maximal value of $S^{new}(B)$ represents the desired batch size $|B|$ and the desired set of points in the set B. The pseudo-code is presented in Algorithm 3.

Algorithm 3 Dynamic Batch Mode Active Learning via Submodular Optimization

Require: Training set L_t and Unlabeled set U_t, parameters λ_1, λ_2 and λ_3

Train a classifier w^t on the training set L_t
$B = \{\phi\}$
for $i = 1 \rightarrow |U_t|$ **do**
  for all $x \in U_t \backslash B$ **do**
    $B_{temp} = B \cup \{x\}$
    Compute $S(B_{temp})$ as in Equation (3.14)
  end for
  Select the point x_{max} producing the largest gain in the objective function (Equation 3.14)
  $B = B \cup \{x_{max}\}$
  $U_t = U_t \backslash \{x_{max}\}$
  Evaluate the current score $S(B)$
  $S^{new}B(i) = S(B) - \lambda_3 * |B|$
end for
Batch Size $m = argmax(S^{new}(B))$
Point Set $P = B(1:m)$
return m and P

Similar to the previous formulation, solving a single optimization problem yields the batch size and the specific points to be selected for batch query. The time complexity is $O(n^2)$ (similar to the state-of-the-art static BMAL algorithms), where n is the number of unlabeled instances.

3.4 Using the Proposed Frameworks for Static BMAL

It is to be noted that the proposed frameworks can be used for batch mode active learning in cases where the batch size is specified. If the batch size is fixed, there is no need to balance the computation cost against the classification performance. Thus, the penalty terms from the objective functions are dropped and a constraint is imposed on the batch size. For example, for the gradient descent based method, the following problem is solved

for static batch mode active learning where the penalty term on the batch size is dropped from the objective and an equality constraint is appended on the batch size m:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda_1 \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1})$$

s.t.:

$$0 \leq M_j \leq 1, \quad \forall j \quad \text{and} \quad \sum_{j=1}^{|U_t|} M_j = m.$$

An analogous strategy is applied for static BMAL using the submodular optimization framework:

$$\max_{B \subseteq U_t : |B| = m} S(B)$$

To achieve this, the outer loop in Algorithm 3 is run from 1 to the desired batch size m and the set B returns the optimum set of points after the loop ends on line 13.

3.5 Experiments and Results

We conducted extensive experiments to depict the efficacy of the proposed dynamic batch mode active learning algorithms. The cost parameter (λ_2 for the SGD algorithm and λ_3 for the sub-modularity based framework) were selected to be 1 (we also empirically demonstrate the effect of this parameter). The other weight parameters were selected to be 1 using cross-validation. Gaussian Mixture Models (GMMs) were used as the classifier in our experiments because of their success in face recognition [231]. The parameters of each Gaussian were trained using the Expectation Maximization (EM) algorithm [232]. For the sake of fair run-time comparison, all the algorithms were implemented in MATLAB on a quad-core Intel processor with 2.66 GHz CPU and 8 GB RAM.

Datasets and Feature Extraction

We used two challenging biometric datasets for our experiments: (1) The VidTIMIT dataset [233], which contains video recordings of subjects reciting short sentences un-

der unconstrained natural conditions and (2) the MOBIO dataset [234], which was recently created for the MOBIO (Mobile Biometry) challenge to test state-of-the-art face and speech recognition algorithms. It contains recordings of subjects under challenging real world conditions, captured using a hand-held device. Sample images from these datasets are shown in Figure 3.1. Our purpose was to test the performance of active learning and so, for the MOBIO dataset, we did not follow the protocols specified in the actual challenge, which were intended for person recognition. Both these datasets contain video recordings of subjects under natural conditions where there is a redundancy of information and are therefore appropriate to test active learning algorithms. The face images in the video frames were automatically detected and cropped to 128 by 128. The Discrete Cosine Transform (DCT) feature was used in all our experiments (for details about the feature extraction process, please refer [235]).

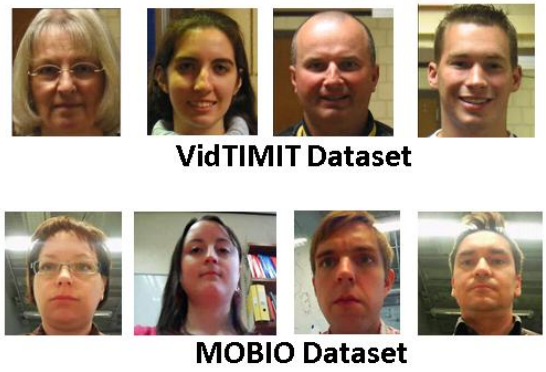


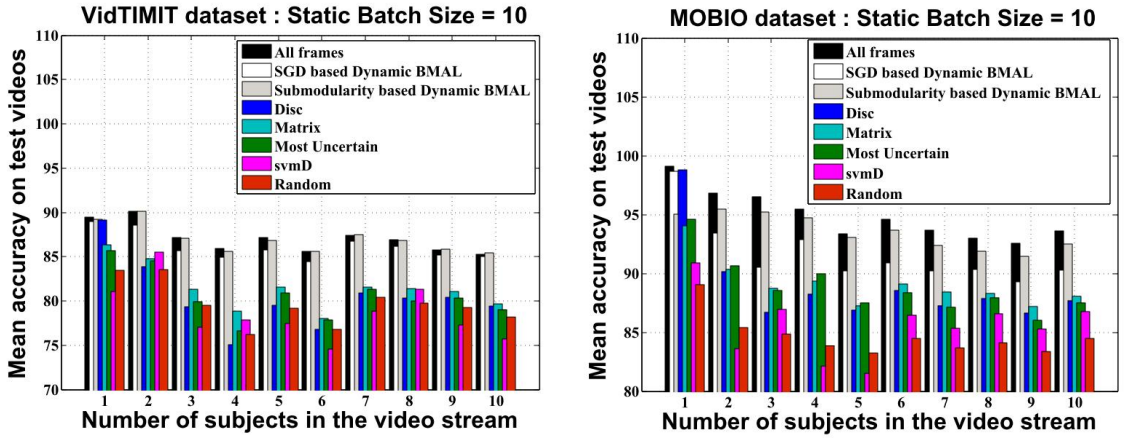
Figure 3.1: Sample images from the VidTIMIT and MOBIO datasets

Experiment 1: Dynamic vs. Static BMAL

The purpose of this experiment was to demonstrate the efficacy of dynamic batch selection over static selection (where the batch size needs to be specified in advance) in applications like face recognition. 25 subjects were randomly selected from each dataset. A classifier was induced with 10 training images of each of the 25 subjects. Unlabeled video streams (each containing about 100 frames) were then presented to the learner. To demonstrate the

generalizibility with different subject combinations, the number of subjects in each unlabeled stream was varied between 1 and 10 (selected randoly from the set of 25). For each stream, the batch size and the specific image samples were selected simultaneously using the proposed optimization strategies. The classifier was updated with the selected images and tested on test videos containing the corresponding subject(s) as in the unlabeled videos.

To illustrate the usefulness of dynamic batch size selection, the accuracy of the proposed techniques was compared against the case when *all the frames* in the unlabeled video were used for learning (this is assumed to be an estimate for the best achievable performance, as there is no better way to quantify the same for a given video stream) and also against the following static BMAL algorithms: (1) *Disc*, a discriminative batch mode active learning strategy, proposed by Guo and Schuurmans [59], (2) *Matrix*, that queries a batch of data samples by maximizing the mutual information between the labeled and unlabeled sets [217], (3) *Most Uncertain*, where the top k uncertain points were queried from the unlabeled video, k being the batch size, (4) *svmD* which incorporates diversity in active learning using SVMs, as proposed by Brinker [203] and (5) *Random*, where a batch of points is queried at random. The Disc and the Matrix approaches have been shown to be the state-of-the-art BMAL techniques [217]. The static batch selection techniques require the batch size to be specified in advance; the static batch size was selected as 10 (the effect of this parameter is studied later). The results are shown in Figure 3.2 and are averaged over 10 trials to rule out effects of randomness. The x axis denotes the number of subjects in the video stream and the y axis denotes the accuracy on test videos containing the corresponding number of subjects. We see that, in both datasets, the accuracy obtained with dynamic batch selection matches the best achievable accuracy more closely than any of the static batch selection algorithms, including the state-of-the-art schemes.



(a) VidTIMIT dataset (b) MOBIO dataset

Figure 3.2: Dynamic vs Static BMAL on the VidTIMIT and MOBIO datasets (static batch size = 10). Best viewed in color.

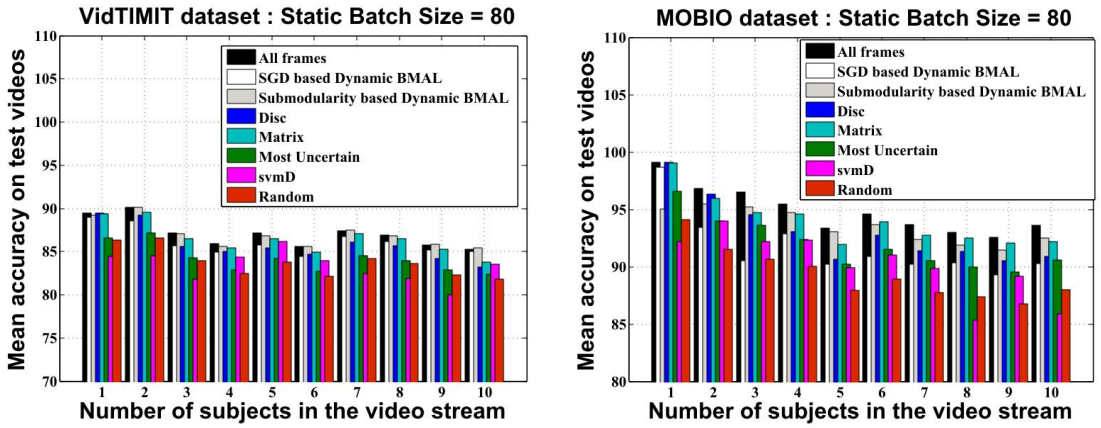
In general, we can expect that if we select a greater number of images from an unlabeled set, the updated learner will perform better on a test set containing the same subjects. Thus, if we select a higher value of the batch size in a static BMAL learner, then static selection is expected to perform better than in Figure 3.2. This is depicted in Figure 3.3 where the static batch size was taken as 80 instead of 10. We see that the static BMAL schemes perform much better than before and the state-of-the-art techniques marginally outweigh dynamic batch selection in terms of classification accuracy.

No. of subjects	1	2	3	4	5	6	7	8	9	10
VidTIMIT (PBS)	18.7	12.4	17.9	25.6	16.5	24.2	22.3	22.5	19.9	24.5
VidTIMIT (LCR)	61.3%	67.6%	62.1%	54.4%	63.5%	55.8%	57.7%	57.5%	60.1%	55.5%
MOBIO (PBS)	15.5	12.2	12.7	15.1	10.9	10.9	10.4	9.9	11.7	11.6
MOBIO (LCR)	64.5%	67.8%	67.3%	64.9%	69.1%	69.1%	69.6%	70.1%	68.3%	68.4%

Table 3.1: Mean predicted batch size (PBS) and percent labeling cost reduction (LCR) using SGD based dynamic selection against static selection with batch size 80.

No. of subjects	1	2	3	4	5	6	7	8	9	10
VidTIMIT (PBS)	29.8	54.2	56.8	60.9	51.8	59.2	61.7	56.4	54.1	50.9
VidTIMIT (LCR)	50.2%	25.8%	23.2%	19.1%	28.2%	20.8%	18.3%	23.6%	25.9%	29.1%
MOBIO (PBS)	19.3	17.9	21.4	21.2	21.1	20.9	22.0	18.4	21.8	21.4
MOBIO (LCR)	60.7%	62.1%	58.6%	58.8%	58.9%	59.1%	58.0%	61.6%	58.2%	58.6%

Table 3.2: Mean predicted batch size (PBS) and percent labeling cost reduction (LCR) using submodularity based dynamic selection against static selection with batch size 80.



(a) VidTIMIT dataset

(b) MOBIO dataset

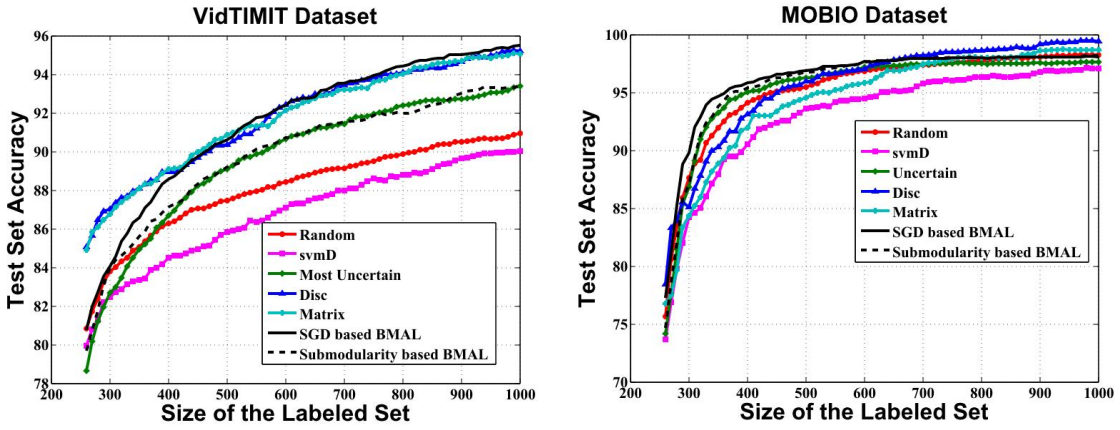
Figure 3.3: Dynamic vs Static BMAL on the VidTIMIT and MOBIO datasets (static batch size = 80). Best viewed in color.

However, to achieve this performance, the static selection required a significantly greater number of images to be labeled than dynamic selection. Table 3.1 shows the mean predicted batch size and mean percentage reduction in the number of images that had to be labeled using SGD optimization based dynamic selection against static selection with batch size 80. It is evident that for both the datasets, the static framework required a much greater number of images to be labeled to marginally outweigh dynamic selection. The same conclusion is evident in Table 3.2 which depicts the analogous values for the submodular optimization based dynamic BMAL framework. Hence, by selecting a number at random, the static batch selection strategies can sometimes query too few points leading to poor generalization power of the updated learner, while in some cases it can entail considerable labeling cost to attain a marginal improvement in accuracy. The dynamic selection strategies, on the other hand, compute the batch size by striking a balance between the uncertainty of the learner on the images in the unlabeled video and the cost of labeling the images, and thus provide a more concrete basis to decide the batch size.

Experiment 2: Batch Selection Criteria of BMAL Algorithms For a Given Batch Size

The purpose of this experiment was to analyze the efficacies of the batch selection criteria of BMAL algorithms to study their usefulness in real world settings. Since the objective was to study the batch selection criteria, the batch size was decided in advance. For the proposed algorithms, the static versions were used (since the batch size was pre-specified), as described in Section 3.4. Similar to Experiment 1, the proposed approaches were compared against the two state-of-the-art BMAL techiques: Disc and Matrix, and the three heuristic techniques: Most Uncertain, svmD and Random. A classifier was induced with 10 training images of each of 25 randomly chosen subjects. Unlabeled video streams (each containing about 250 frames) were then presented to the classifier sequentially. The images in the video streams were randomly chosen from all 25 subjects and did not have any particular proportion of subjects in them, to mimic general real-world conditions. A batch of 10 images was queried from each video stream (that is, the batch size was fixed at 10 for each unlabeled video). After each batch selection, the selected images were appended to the training set, the classifier updated and then tested on a test video containing about 5000 images spanning all the 25 subjects. The goal was to study the increment in accuracy on the test set with increasing size of the training set. The results (averaged over 5 random runs) are presented in Figure 3.4, where the x axis denotes the size of the labeled set and the y axis denotes the accuracy on the test set.

It is evident that the proposed SGD and submodularity based techniques perform much better than svmD and Random sampling. The Most Uncertain depicts the best performance among the heuristic techniques. We also note that the proposed algorithms demonstrate comparable performance as Disc and Matrix, the state-of-the-art BMAL schemes (in fact, they marginally outperform Matrix on the MOBIO dataset). Thus, the proposed algorithms succeed in selecting the salient and prototypical data points



(a) VidTIMIT dataset (b) MOBIO dataset

Figure 3.4: Batch Mode Active Learning on the VidTIMIT and MOBIO datasets (Best viewed in color).

	VidTIMIT	**MOBIO**
SGD Dynamic BMAL	71.02	82.45
Submodular Dynamic BMAL	3.03	5.27
Disc	112.78	122.45
Matrix	38.22	36.79
svmD	1.05	1.17
Most Uncertain	1.23	1.92
Random	0.01	0.01

Table 3.3: Average time taken (in seconds) to query a batch of 10 images from an unlabeled video with 250 images.

and require similar level of manual labeling effort as the state-of-the-art, to attain a given level of generalization accuracy. We also note that the stochastic gradient descent based scheme performs better than the sub-modular BMAL technique for both the datasets. This can be attributed to the fact that the gradient descent strategy selects unlabeled points for manual annotation by directly optimizing the performance with respect to the future learner (the learner trained on the current training set together with the newly selected batch); it therefore has greater efficiency in deciding the set of points which can furnish maximal information. The submodular technique, on the other hand, uses the uncertainty of the current model together with a redundancy-based batch selection criterion and does not involve a "look-ahead" strategy using the future learner.

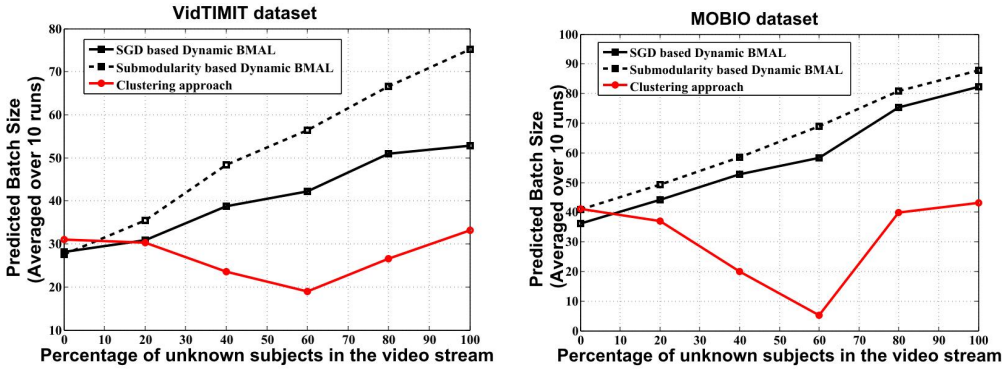
However, since the stochastic gradient descent based BMAL strategy involves classifier retraining in each iteration (due to the involvement of the future learner), its running time is significantly higher than the heuristic methods (as evident in Table 3.3). The submodular framework, on the other hand, is solved using a greedy algorithm (and is devoid of model retraining) and involves much lesser computational overhead, as depicted in the run time values. Thus, depending on the requirements of a particular application, an appropriate scheme can be adopted. While the heuristic techniques (svmD, Random and Most Uncertain) depict promising running time values, their active learning performances are worse than those of the proposed algorithms (Figure 3.4).

Experiment 3: Performance of the Proposed Dynamic BMAL with Varying Complexities of a Video Stream

In a real world scenario, video streams have varying levels of complexities, in terms of presence of unknown subjects (not present in the training set), unknown expressions, head poses, changing illumination among others. To study the performance of the proposed frameworks under such settings, we performed an experiment for dynamic batch selection with varying complexities of a video stream. We compared our approaches against this heuristic clustering based dynamic batch selection scheme (as described in Section 3.1) to study their efficacy in selecting a proper batch size. This was done since no other adaptive batch selection scheme has been proposed till date, to the best of our knowledge.

25 subjects from each dataset were selected and divided into two groups - a "known" group containing 20 subjects and an "unknown" group containing the remaining 5 subjects. A classifier was induced with 10 training images of each of the known subjects. Unlabeled video streams were then presented to the learner and the batch size decided by the optimization and the clustering schemes were noted. The proportion of unknown subjects in the unlabeled video was gradually increased from 0% (where all the subjects

in the unlabeled video were from the training set) to 100% (where none of the subjects in the unlabeled video were present in the training set) in steps of 20%. Thus, the classifier was exposed to different video streams of varying levels of new information. However, the learner was not given any information about the composition of the video streams. Also, the size of each video stream was kept the same (approximately 100 frames) to facilitate fair comparison.



(a) Experiment with unknown subjects from the VidTIMIT dataset

(b) Experiment with unknown subjects from the MOBIO dataset

Figure 3.5: Study of the Proposed Dynamic Batch Selection Frameworks with Varying Complexities of a Video Stream

The results of the aforementioned experiment (averaged over 10 trials to rule out the effects of randomness) are shown in Figure 3.5. The x-axis denotes the percentage of atypical images in the unlabeled pool and the y-axis denotes the batch size predicted using both the proposed and clustering-based strategies. We note that in both the experiments, as the proportion of salient images in the unlabeled stream increases, the uncertainty term outweighs the annotation cost term in the objective functions and the proposed algorithms decide on a larger batch size. This matches our intuition because, with growing percentages of atypical images in the video stream, the confidence of the learner on those images decreases and so it needs to query more images to attain good generalization capability. The clustering based scheme, on the other hand, does not consider the training set in deciding the batch size and so, it fails to reflect the uncertainty of the classifier. The batch

size, therefore, does not bear any specific trend to the percentage of atypical images in the unlabeled set. Thus, while the clustering scheme decides the number of points to be queried based on a score computed from the spatial distribution of the unlabeled points, the optimization based techniques provide a better criteria to decide the batch size by considering the data typicalness with respect to the training set together with the labeling cost.

Proportion of new identities	0%	20%	40%	60%	80%	100%
SGD Dynamic BMAL	96.5%	89%	82%	75%	87.1%	81.3%
Submodular Dynamic BMAL	95.1%	92.2%	91.9%	89%	88.9%	89.5%
Clustering approach	95.9%	85.6%	81.4%	70.6%	79.7%	74.6%

Table 3.4: Test set accuracies using Proposed and Clustering based Dynamic BMAL on the VidTIMIT dataset with increasing proportions of new identities.

Proportion of new identities	0%	20%	40%	60%	80%	100%
SGD Dynamic BMAL	86%	73.5%	75.7%	78.3%	83.3%	87.4%
Submodular Dynamic BMAL	87.1%	79.9%	81.6%	85.3%	86.7%	90.5%
Clustering approach	72.2%	68.1%	53.3%	57.5%	55.9%	56.2%

Table 3.5: Test set accuracies using Proposed and Clustering based Dynamic BMAL on the MOBIO dataset with increasing proportions of new identities.

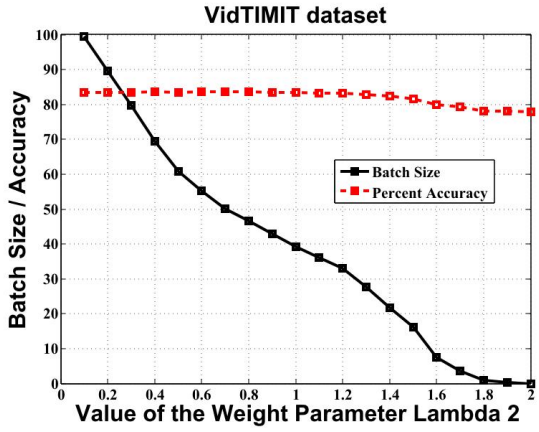
Besides the predicted batch size, it is equally important to analyze the accuracy obtained on test sets with similar compositions as the unlabeled videos. In case of the clustering technique, the gradient descent based approach (Section 3.4) was used for batch query once the batch size was determined [236]. Tables 3.4 and 3.5 show the accuracies obtained on test videos from the VidTIMIT and MOBIO datasets using the optimization and clustering based strategies. The proposed techniques appropriately reflect the uncertainty of the learner and query points accordingly, while the clustering based scheme is a heuristic approach to decide the batch size. Thus, while the optimization based techniques consistently deliver high accuracy values on test videos, the accuracy obtained from the clustering scheme is erratic and inconsistent with varying proportions of new identities in the unlabeled stream. This is more accentuated in the MOBIO dataset. We also note that

the submodularity based technique depicts better accuracy than the SGD based method for both the datasets. However, a comparison of the accuracies of the two dynamic batch selection techniques will not be fair here, as their selected batch sizes are different (as evident from Figure 3.5), unlike the previous experiment, where the batch size was kept constant to facilitate fair comparison. The important thing to note in this experiment is the fact that for both the proposed dynamic batch selection algorithms, the predicted batch size appropriately reflects the complexity of the data.

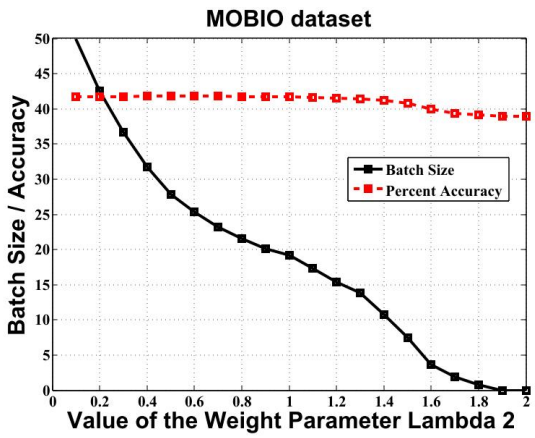
Experiment 4: Effect of the Cost Parameter

In the experiments described above, the value of the cost parameter (λ_2 for the SGD based method and λ_3 for the sub-modularity based method) was taken as 1. Here, we study the effect of this parameter on the batch size and the accuracy. As in the previous experiment, the training set consisted of 250 images and the test set had 5000 images spanning all subjects. An unlabeled video steam (with 250 frames) was then presented for dynamic batch selection and the selected images were appended to the training set (note that in this case, we are not interested in studying the growth in accuracy with increasing size of the training set, hence we focus on the accuracy obtained after dynamic batch selection from a single unlabeled video).

Figure 3.6 shows the results (averaged over 20 different unlabeled video streams) of the SGD based algorithm, where the weight parameter λ_2 was varied between 0 and 2. We note that, an increase in the value of the cost parameter leads to a reduction of the predicted batch size and also the generalization accuracy. This corroborates our intuition as an increase in the labeling cost per sample restricts the number of unlabeled samples that can be purchased for labeling, which also degrades the accuracy on the same test set. Our observation revealed that the difference in accuracy for $\lambda_2 = 0$ and $\lambda_2 = 2$ was about 7%. A similar result was obtained for the parameter λ_3 in the sub-modularity based algorithm.



(a) Study of the Cost Parameter on the VidTIMIT dataset

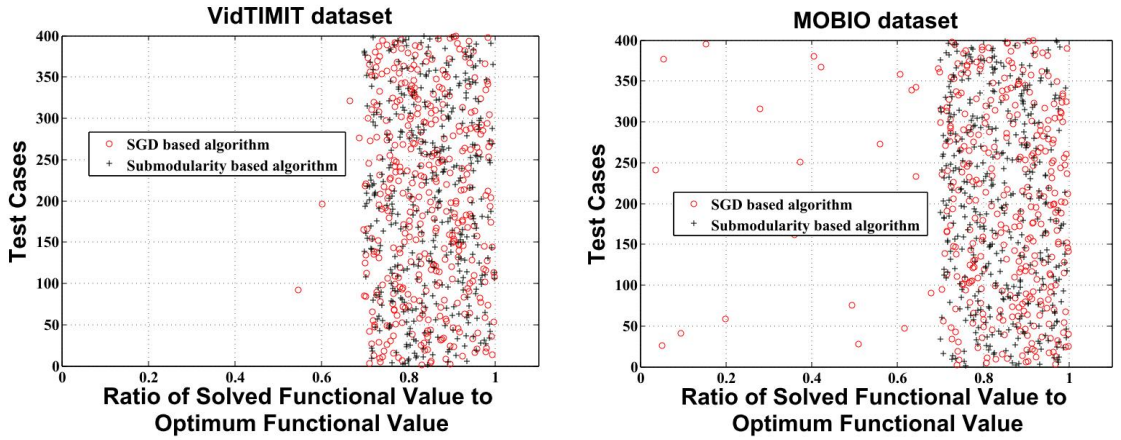


(b) Study of the Cost Parameter on the MOBIO dataset

Figure 3.6: Effect of the Cost Parameter in the SGD based Dynamic BMAL

Experiment 5: Study of Solution Quality

To solve the SGD based optimization problem for dynamic batch selection, the integer constraints in Equation (3.10) were relaxed into continuous constraints in Equation (3.11). Similarly, for the sub-modularity based approach, a greedy algorithm was used to solve the dynamic batch selection problem in Equation (3.15). Both these strategies lead to sub-optimal solutions and it is important to study the quality of the solutions obtained from the relaxations. To this end, 400 random unlabeled video streams were taken from the VidTIMIT and the MOBIO datasets and the relaxed batch selection algorithms were applied for dynamic batch selection. Also, an exhaustive search was performed to find the best solution for a given unlabeled stream by brute-force. The ratio $\frac{f(\widehat{x})}{f(x^*)}$ was computed for the 400 random samples, where $\widehat{x}$ is the solution obtained after relaxation, x^* is the optimal solution obtained by a brute-force search and f is the objective function to be maximized (Equation (3.10) for the SGD based approach and Equation (3.15) for the sub-modularity algorithm).



(a) Study of the Solution Quality on the VidTIMIT dataset

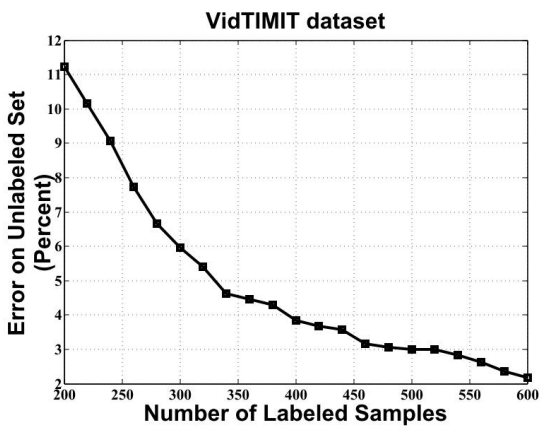
(b) Study of the Solution Quality on the MOBIO dataset

Figure 3.7: Validation of Solution Quality for the SGD and Submodularity based Dynamic BMAL

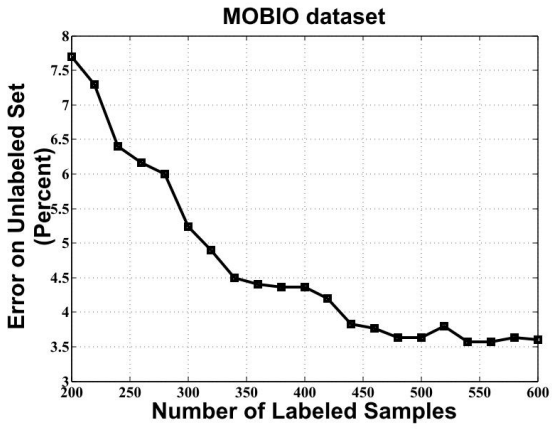
The results are presented in Figure 3.7 and depict the fact that the aforementioned ratio is very close to 1 (greater than 0.8 for most of the test cases). Thus, the functional value attained by solving the relaxation is very close to the optimal functional value. The results lead to the conclusion that both the relaxations produce high quality solutions of the corresponding optimization problems. However, we also note that for the MOBIO dataset, the SGD algorithm sometimes yielded poor solutions (where the ratio is less than 0.3); this is mostly because of bad starting points of the stochastic gradient descent algorithm, which led to bad local optima.

Experiment 6: GTAM Algorithm for Label Prediction

In this experiment, we validate the efficacy of the graph based transductive algorithm (GTAM) proposed by Wang *et al.* [228] to assign labels to the current batch of unlabeled samples in order to estimate the future classifier w^{t+1} in the iterations of the SGD algorithm. The performance of GTAM was studied on a test set of 1000 images with different sizes of the training set ranging from 200 to 600.



(a) Study of the Efficacy of GTAM on the VidTIMIT dataset



(b) Study of the Efficacy of GTAM on the MOBIO dataset

Figure 3.8: Validation of the Efficacy of GTAM

The results are reported in Figure 3.8, which plots the test error against different training set sizes. We note that with only 200 labeled samples, the GTAM algorithm produces a generalization error of about 10% and it reduces further with increasing sizes of the labeled set. This corroborates that the GTAM algorithm is effective in accurately assigning labels to unlabeled samples and thus provides a good approximation of the future classifier w^{t+1} in the Quasi Newton iterations of the SGD based dynamic BMAL algorithm.

3.6 Discussions

In this chapter, we proposed two novel approaches of dynamic batch mode active learning, which adaptively select the batch size and the specific data samples for manual annotation based on the level of complexity of a data stream and the cost of annotation of each unlabeled data sample. Unlike the previously proposed approaches of BMAL, which need the batch size as an input, our framework incorporates the labeling cost in the batch selection criterion and computes the batch size automatically. The batch size and selection criteria are integrated into a single formulation and solving a single optimization problem

yields the desired batch size and the specific samples for query. The frameworks were validated on the face recognition application using two challenging biometric datasets. Our results corroborated the efficacies of the approaches against static BMAL in terms of dynamically identifying the batch size for a given data stream based on its complexity level and the labeling cost of the images. The proposed algorithms also depicted comparable performance against the state-of-the-art static BMAL techniques, when the batch size was pre-specified. We further note that for a given batch size, the gradient descent based scheme has a better label complexity than the sub-modularity approach, but the latter outweighs the former in terms of computation time. Thus, based on the requirements of a given application, an appropriate technique can be selected. Moreover, the algorithms are flexible and the specific terms in the objective function can be modified based on the requirements of a particular application. We also empirically established that our algorithms yield high quality solutions of the relaxations of the corresponding NP-hard problems.

The proposed frameworks can also be used for dynamic batch selection through a single formulation in problems where multiple sources of information are available, such as video containing both face and speech data of an individual or multiple image features extracted from a given face image. Learning from multiple sources can be superior to learning from a single source, if the sources are used appropriately [237]. Let U_{t1} and U_{t2} denote the unlabeled data streams from the two sources of information. Then, the objective functions can be modified by appending relevant terms from the two sources, together with a penalty on the batch size. In case of the stochastic gradient descent method, the following selection criterion can be used for dynamic BMAL:

$$\max_{M} \sum_{j \in U_{t1}} \rho_j M_j - \sum_{j \in U_{t1}} (1 - M_j) S(y|x_j, w^{t+1}) + \sum_{j \in U_{t2}} \rho_j M_j - \sum_{j \in U_{t2}} (1 - M_j) S(y|x_j, w^{t+1}) - \sum_{j} M_j$$

This can be solved in a similar way using the Quasi Newton method. Further, let x_{1i} and x_{2i} denote the feature representations from the two sources of information and let E_1, D_1 and E_2, D_2 be the entropy and the distance functions for the two sources, as defined in Section 3.3. The submodular technique can be adapted for dynamic batch selection from two sources using the following score function:

$$S^{new}(B) = \sum_{x_{1i} \in B} \{\rho(x_{1i}) + E(x_{1i}) + D(x_{1i})\} + \sum_{x_{2i} \in B} \{\rho(x_{2i}) + E(x_{2i}) + D(x_{2i})\} - |B|$$

This can be solved in an analogous way as Algorithm 3, where the submodular and non-decreasing score function is defined as:

$$S(B) = \sum_{x_{1i} \in B} \{\rho(x_{1i}) + E(x_{1i}) + D(x_{i1})\} + \sum_{x_{2i} \in B} \{\rho(x_{2i}) + E(x_{2i}) + D(x_{2i})\}$$

Moreover, if contextual information is available (eg. location of a subject, at home or in office), the same approach can be used to construct a prior probability vector depicting the chances of seeing particular acquaintances in a given context. The entropy term can then be computed on the posterior probabilities obtained by multiplying the likelihood values returned by the classifier with the context aware prior. Thus, subjects not expected in a given context (eg. a home acquaintance in an office setting) will have low priors and consequently, the corresponding posteriors will not contribute much in the entropy calculation. The frameworks can therefore be extended to context-aware adaptive batch selection.

Chapter 4

BATCH MODE ACTIVE LEARNING FOR FUZZY LABEL PROBLEMS : AN ANALYSIS WITH FACIAL EXPRESSION RECOGNITION

Facial expressions play a pivotal role in effective social interactions. Expressions have been recognized as one of the most powerful and immediate means for human beings to communicate their emotions, intentions and opinions to each other. Recent advancements in human-computer interaction (HCI) have spurred active research in the field of automated recognition of facial expressions ([238], [239]). Computer systems endowed with this capability have a wide range of applications including assistive technologies, security, law enforcement, psychiatry and telecommunications among others. In our earlier research [1], (as mentioned in Section 1.1) we have developed a social interaction assistant system intended to aid visually challenged individuals to interact with their sighted peers. The system consists of a pair of glasses with a small camera mounted on the nose-bridge, which the user wears (as shown in Figure 4.1). Efficient analysis of the incoming video stream using computer vision algorithms enables the blind individual better understand and interpret the surroundings. In such an application, accurate estimation of the facial expressions of the subjects in the video stream can enable the visually impaired user judge their emotional states, which is important for effective social interactions.

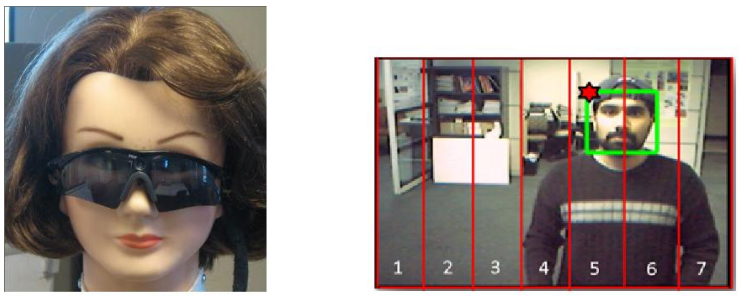


Figure 4.1: The Social Interaction Assistant for individuals with visual impairments [1]

Ekman [240, 241] gave evidence about the universality of facial expressions and proposed six basic expressions that convey human emotion - anger, disgust, fear, happy, sad and surprise. However, these categories have an inherent vagueness in their definitions as there is no clear distinction between the class boundaries. Hence, it is more appropriate to consider a particular data sample as belonging to different classes with varying degrees. Therefore, expression recognition is often treated as a problem involving fuzzy class labels. Moreover, it has been shown that such fuzzy label assignment allows a better representation of each point with respect to the different classes, which can improve the accuracy of the classification model [242, 243]. In this chapter, we propose a novel framework for batch mode active learning in the fuzzy label context. To the best of our knowledge, this is the first attempt to develop such an algorithm for fuzzy label classification problems. We propose an optimization-based strategy for batch selection from an unlabeled set and exploit the properties of sub-modular functions to derive an efficient solution with provable performance guarantees. Although validated only on facial expression data in this work, the proposed framework is generic and can be used in any application involving fuzzy labels (document classification [244] or image segmentation [245], for instance). Our BMAL framework for fuzzy label problems is based on the notion of fuzzy sets and membership functions. We briefly introduce the basic concepts in the following section.

4.1 Fuzzy Sets and Membership Functions

Fuzzy sets are sets whose elements have varying degrees of memberships [246]. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition an element either belongs or does not belong to the set. In contrast, fuzzy set theory permits each element to possess varying degrees of membership, which is quantified by the membership function of the fuzzy set. The membership function out-

puts a value in the real unit interval $[0,1]$ for each element in the set, denoting its level of membership to the set. Fuzzy sets generalize classical sets, since the indicator functions of classical sets are special cases of the membership functions of fuzzy sets, if the latter only take values 0 or 1. In fuzzy set theory, classical bivalent sets are usually called crisp sets. The fuzzy set theory can be used in a wide range of domains in which information is incomplete or imprecise, such as bioinformatics.

Formally, a fuzzy set is a pair (A,μ_A), where A is a set and μ_A is the membership function of the set, that is, $\mu_A : A \rightarrow [0,1]$ [247]. For each element $x \in A$, $\mu_A(x)$ is called the grade or membership of x in the set. The element x is *not included* in the set if $\mu_A(x) = 0$ and *fully included* if $\mu_A(x) = 1$. These cases resemble the classical crisp sets. x is called a fuzzy member if $0 < \mu_A(x) < 1$. The set $\{x \in A | \mu_A(x) > 0\}$ is called the support of (A,μ_A) and the set $\{x \in A | \mu_A(x) = 1\}$ is called its kernel. In the context of fuzzy classification, the trained fuzzy classifier for each class represents the membership function for that class and outputs the membership value of a given point with respect to the class in question.

4.2 Batch Mode Active Learning for Fuzzy Label Problems

Consider a batch mode active learning problem, where we are given a training set L_t and an unlabeled set U_t at time t. Both the training and the unlabeled sets have fuzzy class labels. However, only the labels of the training points are known. Let w^t be the fuzzy classification model trained on L_t. The objective is to select a batch A (containing k points) from U_t in such a way that the future learner w^{t+1}, trained on $L_t \bigcup A$, has maximum generalization capability. Let Y denote the set of possible classes in the problem.

The fundamental idea of batch mode active learning is to identify the salient examples for manual annotation. An example could be considered useful for annotation if the current classification model has a high level of uncertainty in classifying it. Thus, uncertainty or entropy of an unlabeled point is a measure of informativeness of that point. The

entropy of a point in the fuzzy theory domain can be defined using the concept of membership functions. For a fuzzy set A with membership function μ_A, the fuzzy entropy of a point x for $|Y|$ membership functions is computed as follows [248] (here, $\mu_i(x)$ denotes the degree of membership of point x with respect to class i):

$$H = -\sum_{i=1}^{|Y|} \{\mu_i(x)\log(\mu_i(x)) + (1-\mu_i(x))\log(1-\mu_i(x))\} \tag{4.1}$$

However, merely considering uncertainty as the batch selection criterion ignores the redundancy among the selected data points (which results from very similar and nearly duplicate samples). If two points separately furnish valuable information, but they furnish the same/overlapping information, then the knowledge gained by querying both the points is not substantial. Thus, a metric to quantify the diversity among the selected data samples is critical in formulating the batch selection criterion. In this work, the redundancy of an unlabeled example was quantified as the minimum Euclidean distance of this point from the already selected batch. A greater value of this minimum distance depicts a more promising point from the diversity perspective. We would like to select a batch of unlabeled points where each point furnishes useful, but distinctly unique information. To this end, we formulate an objective function denoting the score of a set of points A as follows:

$$S(A) = \sum_{x_i \in A} \{E(x_i) + \lambda D(x_i)\} \tag{4.2}$$

where $E(x_i)$ is the fuzzy entropy of the unlabeled point x_i computed using the membership functions as described in Equation (4.1) and

$$D(x_i) = \min_{x_j \in A: j<i} EuclidDist(x_i, x_j) \tag{4.3}$$

λ is a tradeoff parameter controlling the relative importance of the uncertainty and distance terms. By definition, we have $S(\phi) = 0$. Since the goal is to select a batch of points with high aggregate uncertainty scores and high distance among them, the optimal batch selection condition can be expressed as the following optimization problem:

$$\max_{A \subset U_t : |A| = k} S(A) \tag{4.4}$$

Due to the exponential nature of the search space, exhaustive search techniques are not feasible. In the following sections, we derive an efficient strategy to solve the above optimization problem.

4.3 Submodularity of the Objective Function

Let U be a finite set and let $A \subseteq B \subseteq U$ be two subsets of U. Consider an element $x \in U \setminus B$. A function $f : 2^U \rightarrow \Re$ is submodular if

$$f(A \cup \{x\}) - f(A) \geq f(B \cup \{x\}) - f(B) \tag{4.5}$$

That is, a function is submodular if adding an element to a set increases the functional value by at least as much as adding the same element to its superset. This property is called the diminishing returns property [229, 230].

Lemma 3. *The score function $S(A)$ for batch selection, as defined in Equation (4.2), is a submodular set function.*

Proof. Let A and B be two sets formed by selecting unlabeled points from U_t, such that $A \subseteq B \subseteq U_t$ and consider an unselected instance $x \in U_t \setminus B$. The increment in the value of the objective function achieved by appending x to the already selected set A is given by

$$S(A \cup \{x\}) - S(A) = E(x) + \lambda \min_{x_j \in A} EuclidDist(x, x_j)$$

Similarly, the increment obtained by appending x to the already selected set B is:

$$S(B \cup \{x\}) - S(B) = E(x) + \lambda \min_{x_j \in B} EuclidDist(x, x_j)$$

Since, $A \subseteq B$, the minimum distance of a point x will always be greater for the set A as there may exist some point x_j in the superset B which is closer to x than any element in its subset A. Hence,

$$\min_{x_j \in A} EuclidDist(x, x_j) \geq \min_{x_j \in B} EuclidDist(x, x_j)$$

Thus, we have,

$$S(A \cup \{x\}) - S(A) \geq S(B \cup \{x\}) - S(B)$$

This completes the proof of the lemma.

□

Lemma 4. *The score function $S(A)$ is a monotonically non-decreasing function.*

Proof. Let A denote the currently selected set of points and consider an element $x \in U_t \backslash A$, U_t being the unlabeled pool. If x is added to the current set, the value of the objective function changes by $E(x) + \lambda \min_{x_j \in A} EuclidDist(x, x_j)$. Both the entropy and distance are positive quantities and hence,

$$S(A \cup \{x\}) \geq S(A)$$

This completes the proof.

□

4.4 Greedy Solution with Performance Guarantees

The problem of maximizing a submodular function, as given in Equation (4.4), is NP-hard. However, Nemhauser *et al.* [229] established that for a function S, which is submodular and non-decreasing, with $S(\phi) = 0$, a greedy algorithm provides an efficient solution with near-optimal results. The rationale of the greedy algorithm is to iteratively select points by maximizing the incremental gain in the objective function at each step. Let A_g be

the solution set obtained by the greedy algorithm and A_{max} be the best possible solution set that can be achieved with this optimization formulation. Then, the functional value obtained using the greedy solution can be related to the best achievable value through the following inequality (proved by Nemhauser *et al.* [229]):

$$S(A_g) \geq (1 - \frac{1}{e})S(A_{max}) \tag{4.6}$$

That is, the greedy algorithm is guaranteed to find a solution which achieves at least a constant fraction $(1 - \frac{1}{e}) \approx 63\%$ of the optimal solution. Moreover, no other approximation algorithm can give a better performance guarantee unless $P = NP$ [229]. Based on this result, we propose an efficient algorithm for active batch selection. The following pseudocode outlines our algorithm:

Algorithm 4 Batch Mode Active Learning for Fuzzy Classification

Require: Training set L_t, Unlabeled set U_t and batch size k

```
Train a classifier w^t on the training set L_t
A = {φ}
for i = 1 → k do
    for all x ∈ U_t \ A do
        A_temp = A ∪ {x}
        Compute S(A_temp) as in Equation (4.2)
    end for
    Select the point x_max producing the largest gain in the objective function (Equation 4.2)
    A = A ∪ {x_max}
    U_t = U_t \ {x_max}
end for
return A
```

We note that to compute the entropy term in the objective function, we merely need to apply the trained model on the unlabeled points to derive the class membership values of the points. This does not require classifier retraining in each step (unlike [59]) and hence is computationally very efficient. Evidently, when the fuzzy entropy is replaced by the regular Shannon entropy, this formulation can be applied to problems beyond fuzzy

classification. This corroborates the generalizibility of the framework in efficiently selecting a batch of promising data samples, with provable performance guarantees.

4.5 Experiments and Results

In this section, we present the empirical results of the proposed fuzzy BMAL algorithm.

Datasets and Feature Extraction

We used the MMI and the MindReading datasets in our experiments. The MMI dataset contains videos of subjects exhibiting various expressions and is extensively used in expression recognition research [249]. The MindReading is a computer based guide to emotions primarily collected to help individuals diagnosed with autism recognize facial expressions of emotion [250]. Both these datasets contain videos of subjects under challenging real world conditions and thereby represent an appropriate ground to test our algorithms for active learning in facial expression recognition. Videos containing the six basic expressions for 30 subjects were selected from the databases. Relevant frames around the peak of the expression were extracted from each video. Automated facial detection [251] was applied to crop the faces. The images were subsampled to 96×96 and filtered using a Gabor filter bank of 4 orientations and 4 spatial frequencies [252]. The output from the filter was subsampled to 16×16 pixels. Every frame from the video generated 16 Gabor outputs of size 16×16 pixels. These were concatenated into a single vector of 4096 dimensions. PCA was applied to reduce the dimensionality to 100 retaining about 98% of the variance.

Classification Model

Fuzzy neural network was used for classification (due to its efficacy in fuzzy classification [253]). It allows effective representation of the membership degrees of every data point relative to all the classes. The output layer contained 6 nodes which represented the

membership values of each point with respect to the 6 expression classes. A single hidden layer of 50 nodes was used and the model was trained using the standard backpropagation algorithm.

Experimental Setup

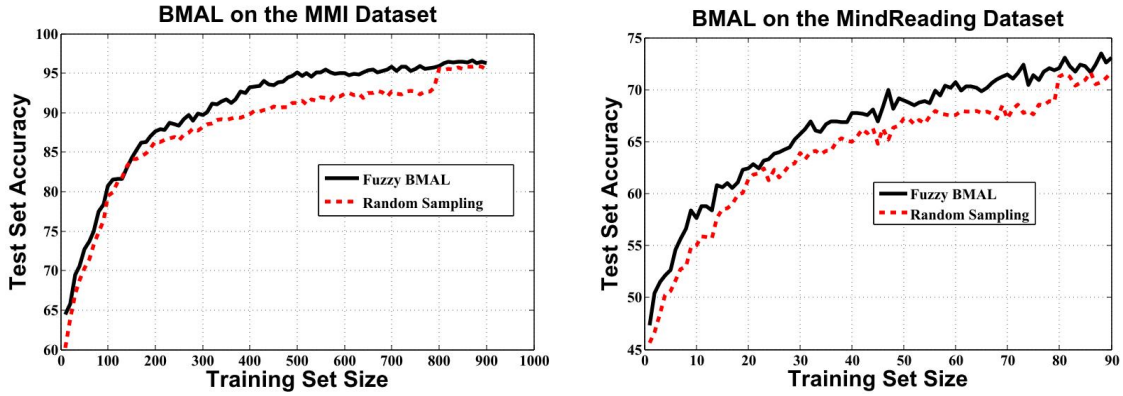
The active learning process started with 60 labeled instances, 10 from each of the 6 expression categories. The learner was exposed to an unlabeled pool of 1000 points and a test set of about 2000 points was set aside to judge the generalization capability of the model. For a fixed batch size k, the algorithm iteratively selects k points from the unlabeled pool to be labeled each time. After batch selection, the selected points were removed from the unlabeled pool and appended to the training set. The goal was to study the improvement in performance on the test set as the newly selected instances are added to the training set (this setup is similar to previous work [59]).

Ground Truth

An image bank containing about 100 images from each expression class was isolated for data fuzzification. The class labels of the points in the training and unlabeled sets were fuzzified using robust measures of inter and intra class dispersions. For details about this method, please refer [253]. The label of each training and unlabeled point was represented using continuous values in the $[0, 1]$ interval denoting its proximity value to the corresponding class centers in the image bank. To predict the class label of a test point, the trained fuzzy model was applied to derive the fuzzy membership values of the different classes and the output was de-fuzzified by predicting the class corresponding to the maximum value. (The purpose of the image bank is to get ground truth fuzzy labels for a given unlabeled point; it plays the role of a human oracle who is entrusted with the task of assigning ground truth labels to the selected points during active learning).

Experiment 1: Fuzzy BMAL vs. Random Sampling

In this experiment, we studied the performance of the proposed fuzzy BMAL technique against Random Sampling, where a batch of points was selected at random from the unlabeled pool (this was used as a baseline for comparison, since no other fuzzy BMAL schemes have been proposed in the literature). The batch size k was taken as 10 in this experiment (similar to [59]).



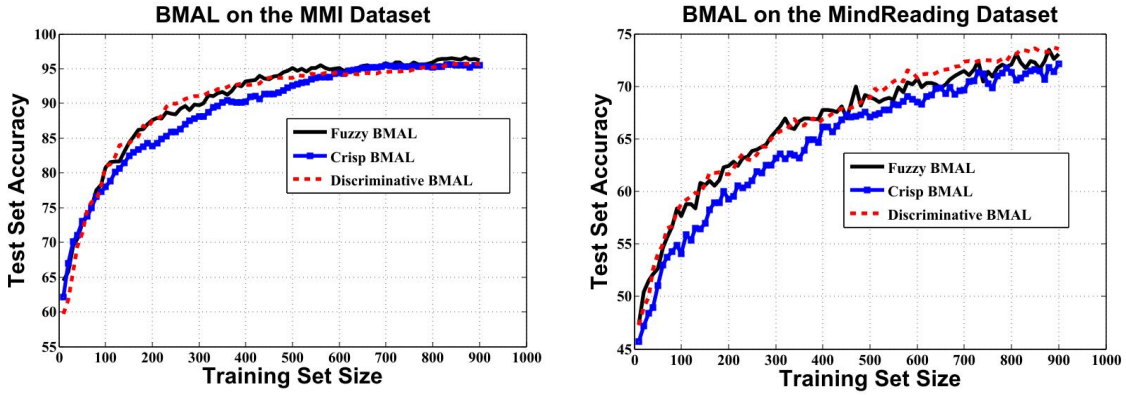
(a) The MMI Dataset (b) The MindReading Dataset

Figure 4.2: Comparison of Fuzzy BMAL against Random Sampling (Best viewed in color)

The results are shown in Figure 4.2. Each graph represents the average performance over 5 random runs with different permutations of the training, testing and unlabeled sets. The x axis represents the size of the training set and the y axis depicts the accuracy obtained on the test set. We note that the proposed fuzzy BMAL algorithm outperforms random sampling as its accuracy on the test set grows at a faster rate. The label complexity (number of data points that need to be labeled to achieve a given level of accuracy) is much less in case of the proposed method.

Experiment 2: Fuzzy BMAL vs. Crisp BMAL

The purpose of this experiment was to present a comparative study of the proposed fuzzy BMAL technique against crisp BMAL. We used two crisp batch mode active learning algorithms in this work - (i) the discriminative batch mode active learning algorithm proposed by Guo and Schuurmans [59], which has been shown to be the best performing crisp BMAL technique till date [217] (the basic idea of this framework was to select a batch of points which maximized the log-likelihoods of the selected points with respect to their assigned class labels and minimized the entropy of the unselected points in the unlabeled pool with respect to the future learner, as detailed in Section 2.3) and (ii) the crisp version of the proposed approach, where the fuzzy entropy term in the objective function was replaced with the Shannon entropy and the crisp labels of the training and unlabeled sets were retained (they were not fuzzified using the image bank).



(a) The MMI Dataset (b) The MindReading Dataset

Figure 4.3: Comparison of Fuzzy BMAL against Crisp BMAL and the Discriminative BMAL algorithm (Best viewed in color)

Figure 4.3 shows the performance (averaged over 5 runs) for the same experimental setup as before. We note that the proposed fuzzy BMAL approach achieves comparable performance to the state-of-the-art algorithm. However, the optimization problem in the discriminative algorithm is non-convex and needs to be solved using iterative techniques

like Quasi-Newton. It also requires the classification model to be retrained in every iteration during batch selection. Thus, it involves severe computational overhead which adversely affects the run time. Table 4.1 presents a comparison of the average time taken to query a batch of 10 points from the unlabeled pool. It is evident that our algorithm surpasses the discriminative approach by a considerable margin in terms of the running time. The proposed scheme thus achieves comparable performance to the state-of-the-art BMAL technique at a significantly lower running time. We also note that although the crisp version of the proposed method achieves comparable performance as the fuzzy version, the fuzzy version has a much lower label complexity, which is the primary goal in active learning. These results show tremendous promise in using fuzzy theory concepts for batch mode active learning in problems like expression recognition, where there is an inherent imprecision and vagueness in the class label definitions.

Dataset	Fuzzy BMAL	Discriminative BMAL
MMI	0.9075	555.1616
MindReading	0.6890	509.7487

Table 4.1: Average time (in seconds) taken by the Fuzzy and Discriminative Batch Mode Active Learning techniques to query a batch of 10 points from the unlabeled pool.

4.6 Discussions

In this chapter, we proposed a novel batch mode active learning framework for fuzzy label classification problems. Our batch selection criterion is based on the uncertainty and redundancy furnished by a set of points. We exploited the theory of submodular functions to derive a greedy algorithm to solve the optimization problem driving the active learning process. The proposed framework is efficient in terms of accuracy and computation time and also has strong performance guarantees. Although validated only on fuzzy label problems in this chapter, the proposed framework is generic and can be easily extended to crisp label classification problems. The results in the previous section demonstrate that the proposed approach has tremendous potential in reducing human annotation effort in

real world fuzzy classification problems like facial expression recognition and can also be scaled to large datasets commonly encountered in today's digital world.

Chapter 5

ACTIVE BATCH SELECTION VIA CONVEX RELAXATIONS WITH GUARANTEED SOLUTION BOUNDS

In this chapter, we attempt to derive novel batch mode active learning algorithms which provide strong theoretical guarantees on the quality of the obtained solutions. State-of-the-art BMAL techniques typically formulate the batch selection task as an NP-hard integer programming problem. Convex relaxations of the NP-hard problems are then solved to select an appropriate batch of unlabeled samples [59, 217]. Even though such techniques depict promising empirical performance, no formal mathematical guarantee has been established on the solution qualities of the convex relaxations. In this work, we first formulate the batch selection as an NP-hard integer quadratic programming (IQP) problem. We then propose a linear programming (LP) based relaxation and show that the batch selection task reduces to a score ranking problem; we hence call this method *BatchRank*. We also propose a semi-definite programming (SDP) based relaxation and use randomization techniques to solve the BMAL problem; we therefore name this algorithm *BatchRand*. Further, we derive a deterministic bound on the quality of the first relaxation and a probabilistic bound on the quality of the second. To the best of our knowledge, this is the first research effort to provide concrete mathematical guarantees on the solution quality of the batch mode active learning problem. (We note here that the purpose of this work is not to derive a bound on the number of queries required to achieve a given generalization error in active learning. This problem has been extensively studied in the literature [73, 78, 254]. Our objective is to derive mathematical guarantees on the solution qualities of the convex relaxations of the batch mode active learning problem, which, to the best of our knowledge, has not been addressed till date). We also empirically validate that the proposed

algorithms deliver high quality solutions and are robust to label noise and population imbalance.

5.1 The Proposed Batch Mode Active Learning Formulation

Consider a batch mode active learning problem, where we are given a training set L_t and an unlabeled set U_t at time t. Let w^t be the classifier trained on L_t. The objective is to select a batch B (containing k points) from U_t in such a way that the future learner w^{t+1}, trained on $L_t \bigcup B$, has maximum generalization capability. Let Y denote the set of possible classes in the problem. We quantify the quality of a batch of selected samples based on their informativeness and diversity, that is, we would like to select a batch of samples such that each point furnishes valuable information and the selected samples have minimal redundancy among them.

Formally, we compute an information vector $c \in \Re^{|U_t|\times 1}$ where $c(i)$ denotes the information furnished by the point x_i in the unlabeled pool. The uncertainty of the trained model w^t on a point x_i is used as a measure of the informativeness of x_i; higher uncertainty values denote higher degrees of information. The uncertainty of an unlabeled point x_i (that is, $c(i)$) is computed as the entropy $S(y|x_i, w^t)$ of the distribution $P(y|x_i, w^t)$, such that

$$c(i) = S(y|x_i, w^t) = -\sum_{y \in Y} P(y|x_i, w^t) \log P(y|x_i, w^t) \tag{5.1}$$

In addition to c, a divergence matrix $R \in \Re^{|U_t|\times|U_t|}$ is also defined whose $(i, j)^{th}$ entry is a measure of redundancy between unlabeled points x_i and x_j (higher the value of R_{ij}, lower the redundancy). The divergence measure between two points is an estimate of the amount of information overlap between the points, which is captured by the symmetric Kullback Leibler divergence. Let p^i and p^j denote the vectors of posterior probabilities of two points x_i and x_j in the unlabeled pool with respect to all the classes. Then, the $(i, j)^{th}$

entry in R equals the symmetric KL divergence between these two vectors (for details, please refer [255]):

$$R(i,j) = \sum_{y=1}^{|Y|} (p_y^i - p_y^j) \log \frac{p_y^i}{p_y^j} \tag{5.2}$$

By definition, all the entries in c and R are non-negative, that is, $c_i \geq 0$ and $r_{ij} \geq 0$. Also, $R_{ii} = 0, \forall i$. Given c and R, our objective is to select a batch of points having high information scores and high divergence (or minimal redundancy) among them. For notational simplicity, we combine c and R into a single matrix $D \in \Re^{|U_t| \times |U_t|}$ as follows:

$$D(i,j) = \begin{cases} R(i,j), & \text{if } i \neq j \\ \lambda . c(i), & \text{if } i = j \end{cases} \tag{5.3}$$

where each entry in the matrix D is non-negative, that is $d_{ij} \geq 0, \forall i, j$. λ is a trade-off parameter. (We note that the matrix D can be defined suitably based on the application at hand. Since defining the most suitable batch selection criterion for a given application is not the focus of this work, we proceed with the criterion based on entropy and redundancy and explain our framework.)

We now formulate the batch selection task as an explicit mathematical optimization problem, where the objective is to select a batch of points with high aggregate uncertainty scores and high divergences among them. Specifically, we define a binary vector m with $|U_t|$ entries ($m \in \{0,1\}^{|U_t| \times 1}$) where each entry m_i denotes whether the corresponding unlabeled point x_i will be included in the batch ($m_i = 1$) or not ($m_i = 0$). Our batch selection criterion (with given batch size k) can thus be expressed using the following integer quadratic programming (IQP) problem:

$$\max_m m^T D m$$

$$\text{s.t. } m_i \in \{0,1\}, \forall i \text{ and } \sum_{i=1}^{|U_t|} m_i = k \tag{5.4}$$

The binary constraint on m_i makes this IQP problem NP-hard. We now propose two convex relaxations to solve this NP-hard problem.

5.2 BatchRank : Convex Relaxation I

We first show that the IQP in Equation (5.4) is equivalent to an Integer Linear Programming (ILP) problem in the following Lemma.

Lemma 5. *The Integer Quadratic Programming batch mode active learning formulation in Equation (5.4) can be simplified into an equivalent Integer Linear Programming (ILP) problem.*

Proof. We introduce a binary matrix $Z = (z_{ij})$ with $z_{ij} = m_i.m_j$. Thus, the optimization problem in (5.4) reduces to:

$$\max_{m,Z} \sum_{i,j} d_{ij} z_{ij}$$
$$\text{s.t. } z_{ij} = m_i m_j, \sum_{i=1}^{|U_t|} m_i = k, \quad and \quad m_i \in \{0,1\}, \forall i \tag{5.5}$$

The quadratic equality constraint $z_{ij} = m_i m_j$ makes this problem difficult to solve. Interestingly, we can show that this quadratic constraint, in fact, allows itself to be represented as a simpler linear inequality $-m_i - m_j + 2z_{ij} \leq 0, \forall i, j$. This ensures that the value of z_{ij} is 0 if either m_i or m_j (or both) is equal to 0. When both m_i and m_j are equal to 1, z_{ij} is free to be either 0 or 1. However, the maximization criterion in (5.5) forces the value of z_{ij} to be 1 since $d_{ij} \geq 0$. Hence, the problem can now be written as:

$$\max_{m,Z} \sum_{i,j} d_{ij} z_{ij}$$
$$\text{s.t. } -m_i - m_j + 2z_{ij} \leq 0, \forall i, j$$
$$and \sum_{i=1}^{|U_t|} m_i = k, m_i, z_{ij} \in \{0,1\}, \forall i, j \tag{5.6}$$

This is an integer LP problem, proving Lemma 5. □

Although a global maximum exists for the ILP, it is computationally expensive to compute. To solve such an ILP, a standard approach is to employ the LP relaxation.

Lemma 6. *The convex LP relaxation of the above ILP (Equation 5.6) in Lemma* 5 *is equivalent to a ranking formulation based on the entries in the matrix D.*

Proof. We consider the following linear program relaxation:

$$\max_{m,Z} \sum_{i,j} d_{ij} z_{ij}$$

$$\text{s.t.} \quad -m_i - m_j + 2z_{ij} \leq 0, \forall i, j, \quad \sum_{i=1}^{|U_t|} m_i = k$$

$$\textit{and} \quad m_i, z_{ij} \in [0,1], \forall i, j \tag{5.7}$$

Since this is a maximization problem with $d_{ij} \geq 0$, at optimality, $z_{ij} = \frac{m_i + m_j}{2}$ (from the inequality constraint $-m_i - m_j + 2z_{ij} \leq 0$). Hence, (5.7) is equivalent to:

$$\max_{m} \frac{1}{2} \sum_{i,j} d_{ij}(m_i + m_j)$$

$$\text{s.t.} \quad \sum_{i=1}^{|U_t|} m_i = k \quad \textit{and} \quad m_i \in [0,1], \forall i \tag{5.8}$$

This formulation admits an analytical (as well as an integer) solution for m by a simple ranking based on the entries in the matrix D. The objective in (5.8) can be written as $\sum_{i,j} d_{ij} m_i + \sum_{i,j} d_{ij} m_j$. Since the matrix D is symmetric, the maximization problem essentially becomes equivalent to ranking the column sums of D (hence the name BatchRank). This proves Lemma 6. □

The pseudo-code for the BatchRank algorithm is given below. The complexity of the algorithm is $O(n^2)$, where n is the number of unlabeled samples.

Algorithm 5 BatchRank algorithm for Batch Mode Active Learning

Require: Training set L_t, Unlabeled set U_t and batch size k

1: Train a classifier w^t on the training set L_t
2: Compute information vector c (Equation 5.1) and the divergence matrix R (Equation 5.2) using w^t
3: Compute the matrix D, as described in Equation (5.3)
4: Compute a vector $v \in \Re^{|U_t| \times 1}$ containing the column sums of D
5: Identify the k largest entries in v and select the corresponding unlabeled points from U_t in the batch

Solution Bound of BatchRank

In this section, we prove a bound on the solution to the convex LP relaxation in (5.8) with respect to the solution of the original NP-hard integer quadratic programming problem. To this end, we transform the original maximization problem in Equation (5.4) into an equivalent minimization problem through the following objective function:

$$f(m) = ||D||_1 - m^T Dm \tag{5.9}$$

where $||D||_1 = \sum_{i,j} d_{ij}$. We note that since $||D||_1$ is constant for a given matrix D, maximizing $m^T Dm$ as in Equation (5.4) is equivalent to minimizing the function $f(.)$ defined above, that is, maximizing $m^T Dm$ and minimizing $f(.)$ as defined above will fetch exactly the same solution to the variable m. Since we are interested in the solution quality of m, we prove an upper bound on the minimization problem in Equation (5.9). Since the solution to m is the same, it is essentially equivalent to proving a bound on the original maximization problem in Equation (5.4). The main result is summarized in the following theorem:

Theorem 1. *Let m^* and $\widehat{m}$ be optimal solutions of (5.4) and (5.8) respectively. Then,*

$$f(\widehat{m}) \leq 2f(m^*)$$

Proof. The optimization in (5.8) is an LP relaxation of the quadratic formulation in (5.4) and thus the objective value of (5.8) is larger than that of (5.4). That is,

$$m^{*T}Dm^* \leq \frac{1}{2}\sum_{i,j} d_{ij}(\widehat{m}_i + \widehat{m}_j)$$

$$= \frac{1}{2}\sum_{i,j:\widehat{m}_i+\widehat{m}_j=1} d_{ij} + \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} \tag{5.10}$$

Since all entries in D are non-negative, the following holds:

$$\begin{aligned} ||D||_1 &= \sum_{ij} d_{ij} \\ &= \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} + \sum_{i,j:\widehat{m}_i+\widehat{m}_j=1} d_{ij} + \sum_{i,j:\widehat{m}_i+\widehat{m}_j=0} d_{ij} \\ &\geq \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} + \sum_{i,j:\widehat{m}_i+\widehat{m}_j=1} d_{ij} \end{aligned}$$

Thus,

$$\sum_{i,j:\widehat{m}_i+\widehat{m}_j=1} d_{ij} \leq ||D||_1 - \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} \tag{5.11}$$

Combining the above two, we have

$$\begin{aligned} f(m^*) &= ||D||_1 - m^{*T}Dm^* \\ &\geq ||D||_1 - \frac{1}{2}\sum_{i,j:\widehat{m}_i+\widehat{m}_j=1} d_{ij} - \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} \\ &\geq ||D||_1 - \frac{1}{2}\left(||D||_1 - \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij}\right) \\ &\quad - \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij} \\ &= \frac{1}{2}\left(||D||_1 - \sum_{i,j:\widehat{m}_i+\widehat{m}_j=2} d_{ij}\right) = \frac{1}{2}f(\widehat{m}) \end{aligned}$$

The first inequality follows from Equation (5.10) and the second from Equation (5.11). The last equality is true because in the evaluation of $m^T Dm$, only the indices

where both m_i and m_j are 1 will survive, others will vanish. This completes the proof of the theorem. We thus note that the convex relaxation of the original NP-hard problem in BatchRank has a guaranteed bound on the solution quality.

□

5.3 BatchRand: Convex Relaxation II

In this section, we attempt to further improve the theoretical guarantee of the solution quality of the NP hard batch selection problem. We therefore propose a second relaxation based on a randomized approximation algorithm (hence the name BatchRand) and show that it achieves a probabilistic bound of 87.856%. Starting with Equation (5.4), we first make the following variable transformation:

$$y_i = 2(m_i - \frac{1}{2}) \Rightarrow m_i = \frac{y_i + 1}{2}$$
$$\Rightarrow \sum_{i=1}^{n} y_i = 2\sum_{i=1}^{n} m_i - n = 2k - n \triangleq p$$

where $n = |U_t|$ is the number of unlabeled data samples in the unlabeled pool. The entire optimization problem in Equation (5.4) is now rewritten in terms of the new variable y (ignoring the constant $\frac{1}{4}$):

$$\max_{y} \sum_{i,j} d_{ij}(y_i + 1)(y_j + 1)$$
$$\text{s.t. } y_i \in \{-1, 1\}, \forall i \text{ and } \sum_{i=1}^{|U_t|} y_i = p \tag{5.12}$$

Multi-dimensional Relaxation

Since solving the integer quadratic program in Equation (5.12) is NP hard, we consider relaxations of the constraints. Specifically, we follow the strategy proposed by Goemans and Williamson [256], where each variable y_i is relaxed to a multidimensional vector v_i belonging to $\Re^n$ of unit Euclidean norm, instead of a one dimensional scalar variable. In other words, we assume that each vector v_i belongs to the n-dimensional unit sphere S_n.

The relaxation of the NP hard problem in Equation (5.12) is therefore given by (ignoring the constant 1):

$$\max_{v} \sum_{i,j} d_{ij}(v_i^T v_j + v_0^T v_i + v_0^T v_j) \tag{5.13}$$

$$\text{s.t.} \;\; v_i \in S_n, \forall i \;\; \text{and} \;\; \sum_{i=1}^{|U_t|} v_i = p \tag{5.14}$$

where v_0 is a vector of n-dimensions with all entries 1. Once we solve for the vectors v from this formulation (we present the solution details below), we select a random unit vector r that is uniformly distributed on the unit sphere and find the dot product of r with every vector v_i. We then select the set of unlabeled points whose corresponding v vectors yield a positive dot product with the random unit vector r, as in [256]. We note that the number of positive dot products may not exactly equal the desired batch size k. However, due to the equality constraint in Equation (5.14), the *expected* number of positive dot products equals the batch size k.

Semi-Definite Programming (SDP) Relaxation

Using the decomposition $Y = B^T B$, we note that any positive semi-definite (psd) matrix with diagonal entries 1 corresponds to a set of unit vectors v_i if we correspond the vector v_i to the i^{th} column of the matrix B. Then, $y_{ij} = v_i v_j$, which accounts for the term $v_i^T v_j$ in the objective function of Equation (5.13). However, to incorporate the terms $v_0^T v_i$ and $v_0^T v_j$ in the objective, the matrix Y is decomposed as

$$Y = \begin{pmatrix} v_0^T \\ B^T \end{pmatrix} \begin{pmatrix} v_0 & B \end{pmatrix}$$

where $B = [v_1\, v_2\, \ldots v_n]$. We can therefore rewrite the entire relaxation in terms of the defined matrix Y (in the previous equation) as follows:

$$\max_{Y} \sum_{i,j} d_{i,j}(Y_{i+1,j+1} + Y_{1,i+1} + Y_{1,j+1})$$

$$\text{s.t.} \quad Y_{ii} = 1, \text{for i = 2 to n+1},$$
$$\sum_{j=2}^{n+1} Y_{1j} = p, \quad Y_{11} = n$$
$$\text{and} \quad Y \succeq 0 \quad (\text{Y is psd})$$

This is a semi-definite programming (SDP) problem and can be solved using existing software packages like SeDuMi. The pseudocode of the BatchRand algorithm is provided below. The complexity of the algorithm is $O(n^3)$, n being the number of unlabeled data samples.

Algorithm 6 BatchRand algorithm for Batch Mode Active Learning

Require: Training set L_t, Unlabeled set U_t and batch size k

1: Train a classifier w^t on the training set L_t
2: Compute information vector c (Equation 5.1) and the divergence matrix R (Equation 5.2) using w^t
3: Compute the matrix D, as described in Equation (5.3)
4: Solve the optimization problem (Equation 5.13) to yield a set of vectors v
5: Select a random unit vector r and evaluate the dot product of r with each vector v_i
6: Select the set of unlabeled points $S = \{i | v_i.r \geq 0\}$ for manual annotation

Probabilistic Solution Guarantee

We first rewrite the objective in Equation (5.13) in a simplified form as follows:

$$\sum_{i,j} d_{ij}(v_i^T v_j + v_0^T v_i + v_0^T v_j) = \sum_{i,j} \widehat{d_{ij}}(v_i^T v_j)$$

where v_0 is the vector obtained from the first row of the decomposed matrix after solving the SDP problem and $\widehat{d_{ij}}$ is obtained from d_{ij}, to simplify the representation. The main result regarding the solution bound of the BatchRand algorithm is summarized in the following theorem:

Theorem 2. *Let W denote the value of the objective function produced using the BatchRand algorithm and $\mathbb{E}(W)$ denote its expectation. Also, let $\widehat{D}_{total}$ denote the sum of all entries in the matrix $\widehat{D}$. Then,the following bound holds:*

$$[\mathbb{E}(W) + \widehat{D}_{total}] \geq 0.87856 \left[\sum \widehat{d_{ij}} v_i v_j + \widehat{D}_{total}\right]$$

Proof. Consider a random unit vector r. By the linearity of expectation, the expectation of the value of the objective function is given by:

$$\begin{aligned}\mathbb{E}(W) &= \sum \widehat{d}_{ij}[1.Pr(sgn(v_i.r) = sgn(v_j.r)) \\ &\quad + (-1).Pr(sgn(v_i.r) \neq sgn(v_j.r))] \\ &= \sum \widehat{d}_{ij}[1 - 2.Pr(sgn(v_i.r) \neq sgn(v_j.r))] \\ &= \sum \widehat{d}_{ij}\left[1 - 2\frac{arccos(v_i.v_j)}{\pi}\right]\end{aligned}$$

where $sgn(x) = 1$ if $x \geq 0$ and -1 otherwise. The last equality follows from the result proved by Goemans and Williamson [256]. Further, it can be shown that for $-1 \leq z \leq 1$, $1 - \frac{arccos(z)}{\pi} \geq \alpha.\frac{1}{2}(1+z)$, where

$$\alpha = \min_{0 \leq \theta \leq \pi} \frac{2}{\pi}.\frac{\theta}{1-\cos\theta} \geq 0.87856$$

(the proof of the above inequality can be found in [256]). That is, $1 - 2\frac{arccos(z)}{\pi} \geq \alpha(1+z) - 1$. Since in our formulation, the $v_i s$ are all unit vectors, we have $-1 \leq z = v_i v_j \leq 1, \ \ \forall i, j$. Therefore, $1 - 2\frac{arccos(v_i v_j)}{\pi} \geq \alpha(1 + v_i v_j) - 1$.

Combining all of the above, we get

$$\begin{aligned}\mathbb{E}(W) &\geq \sum \widehat{d}_{ij}[\alpha(1 + v_i v_j) - 1] \\ &= \alpha \sum \widehat{d}_{ij} v_i v_j + (\alpha - 1)\widehat{D}_{total} \\ &\Rightarrow [\mathbb{E}(W) + \widehat{D}_{total}] \geq \alpha \left[\sum \widehat{d}_{ij} v_i v_j + \widehat{D}_{total}\right]\end{aligned}$$

which proves the theorem.

□

5.4 Computational Considerations

We note that the time complexity of BatchRank is $O(n^2)$ and that of BatchRand is $O(n^3)$. This may limit the scalability of the algorithms to very large datasets. To overcome this, we used a sub-sampling strategy in our empirical study, where the current classifier was

applied to all the unlabeled data samples and the batch selection was restricted to the top p uncertain samples (given by the entropy values of the current model). The value of p was taken as 400 in our experiments and can be suitably selected based on a given application. In future, we plan to investigate other sub-sampling strategies; for instance, another sub-sampling approach may be to perform a k-means clustering on the unlabeled data with $k = p$, take the p cluster centers as the sub-sampled pool and restrict the batch selection to this subset. This method will select the representative samples from the unlabeled set into the sub-sampled pool, as opposed to selecting the uncertain samples, as in our experiments. Comparing the two approaches is an interesting direction of future research.

5.5 Experiments and Results

In this section, we empirically study the performance of the proposed BatchRank and BatchRand algorithms.

Datasets and Feature Extraction

The datasets used in our experiments are detailed below:

UCI datasets: We used 9 datasets (binary and multi-class) from the UCI Machine Learning Repository [257] as benchmarks to validate our algorithms.

Face Recognition datasets: We also used two challenging biometric datasets in our experiments: (1) The VidTIMIT dataset [233], which contains video recordings of subjects reciting short sentences under unconstrained natural conditions and (2) the MOBIO dataset [234], which was recently created for the MOBIO (Mobile Biometry) challenge to test state-of-the-art face and speech recognition algorithms. It contains recordings of subjects under challenging real world conditions, captured using a hand-held device. (Our purpose was to test the performance of active learning and so, for the MOBIO dataset,

we did not follow the protocols specified in the actual challenge, which were intended for person recognition.) Both these datasets contain video recordings of subjects under natural conditions where there is a redundancy of information and are therefore appropriate to test active learning algorithms. The face images in the video frames were automatically detected and cropped to 128 by 128. The Discrete Cosine Transform (DCT) feature was extracted from the face images [235] and PCA was used to reduce the dimension to 100, retaining about 99% of the variance.

Facial Expression Recognition datasets: We further used two challenging facial expression recognition datasets - the MMI and the MindReading datasets, to test our algorithms. The MMI dataset contains videos of subjects exhibiting various expressions and is extensively used in expression recognition research [249]. The MindReading is a computer based guide to emotions primarily collected to help individuals diagnosed with autism recognize facial expressions of emotion [250]. Both these datasets contain videos of subjects under challenging real world conditions and thereby represent an appropriate ground to test our algorithms for active learning in facial expression recognition. Videos containing the six basic expressions for 30 subjects were selected from the databases. Relevant frames around the peak of the expression were extracted from each video. Automated facial detection [251] was applied to crop the faces. The Gabor filter was applied to the images for feature extraction ([252]) and PCA was used to reduce the dimensionality to 100, retaining about 98% of the variance.

Multi-Label datasets: In addition, we also validated our algorithms on two multi-label datasets - the Scene and the Yeast. These are widely used in multi-label learning research [258].

The value of the weight parameter λ was selected as 1 based on preliminary experiments.

Competing Algorithms and Experiment Set-up

We compared the proposed algorithms against the following batch mode active learning strategies proposed in the literature: (1) *Random*, where a batch of points is queried at random from the unlabeled set (this is used for baseline comparison), (2) *Most Uncertain*, where the top k uncertain points are queried from the unlabeled set, k being the batch size, (3) *Fisher*, that selects a batch of samples for manual annotation by maximizing the Fisher information of the classification model (proposed by Hoi *et al.* [95]), (4) *Disc*, a discriminative strategy that selects a batch of points by optimizing the performance of the future learner, proposed by Guo and Schuurmans [59] and (5) *Matrix*, that queries a batch of unlabeled points by maximizing the mutual information between the labeled and unlabeled points [217]. The Disc and Matrix approaches have been shown to be the state-of-the-art BMAL schemes [217].

For each dataset, we started with an initial labeled training set, an unlabeled pool and a test set. For a fixed batch size k, each algorithm repeatedly selected k instances from the unlabeled pool to be labeled each time (as mentioned in Section 5.3, the number of points queried by BatchRand may not be exactly k, this algorithm was therefore used first to note the exact number of samples queried in each iteration; these values were then used in the other algorithms to query the same number of points in the corresponding iteration, for fair comparison). After each batch selection, the selected points were removed from the unlabeled pool and appended to the training set. The goal was to study the improvement in performance on the test set with an increasing size of the training set (this experimental setup is similar to earlier work [59],[217]). All the results were averaged over 10 runs to rule out the effects of randomness. The sub-sampling strategy, mentioned in Section 5.4, was used for the BatchRand and BatchRank algorithms when the size of the unlabeled set was more than 400. Logistic Regression (LR) was used as the base classifier (similar to

[59]). The training, unlabeled and test splits for each dataset are summarized in Table 5.1. The algorithms were implemented in MATLAB on a quad-core Intel processor with 2.66 GHz CPU and 8 GB RAM.

Dataset	Classes	Dimensionality	Training	Unlabeled	Testing
Breast Cancer	2	30	10	259	300
Heart	2	13	4	120	146
Musk	2	166	2	500	490
Spect	2	22	7	110	150
Wine	3	13	3	87	88
Waveform	3	20	9	1000	500
Vehicles	4	18	16	500	330
Image Segmentation	7	19	35	300	2000
Handwritten Digits	10	64	50	1000	2751
VidTIMIT	25	100	250	1000	4500
MOBIO	25	100	50	1000	4500
MindReading	6	100	50	1000	1511
MMI	6	100	50	1000	1785
Scene	6	294	10	350	2047
Yeast	14	103	10	600	1807

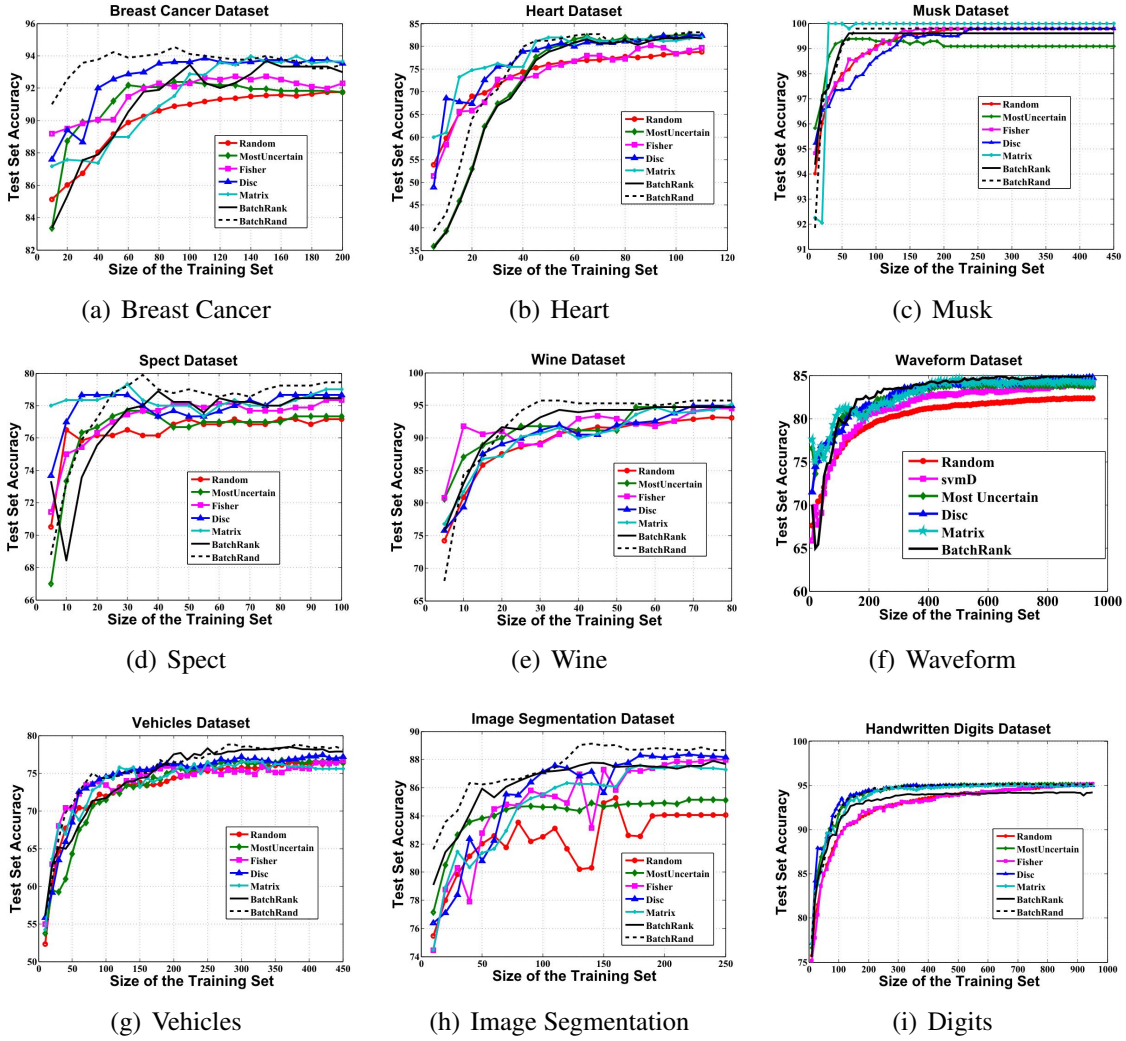
Table 5.1: Dataset Details

Experiment 1: Batch Mode Active Learning on Binary and Multi-class Datasets

The results on the UCI datasets are reported in Figure 5.1. In each graph, the x axis denotes the size of the labeled training set and the y axis denotes the accuracy obtained on the test set. As mentioned earlier, the objective was to study the growth in accuracy on the test set as more and more points are queried from the unlabeled set.

From the results, it is evident that BatchRand and BatchRank outperform Random sampling on all the datasets, as the accuracy grows at a faster rate with increasing size of the labeled set. This shows that the proposed approaches succeed in selecting the salient and prototypical samples from the unlabeled data population and attain a given level of accuracy with the least number of labeled examples. The Most Uncertain and Fisher methods perform better than random sampling, but are not as good as the proposed algorithms. The proposed frameworks consistently depict comparable performance to Disc and Matrix, the state-of-the-art BMAL techniques. Also, we note that the BatchRand approach performs

better than BatchRank. This corroborates our intuition as the BatchRand method offers a much better expected performance guarantee than BatchRank.



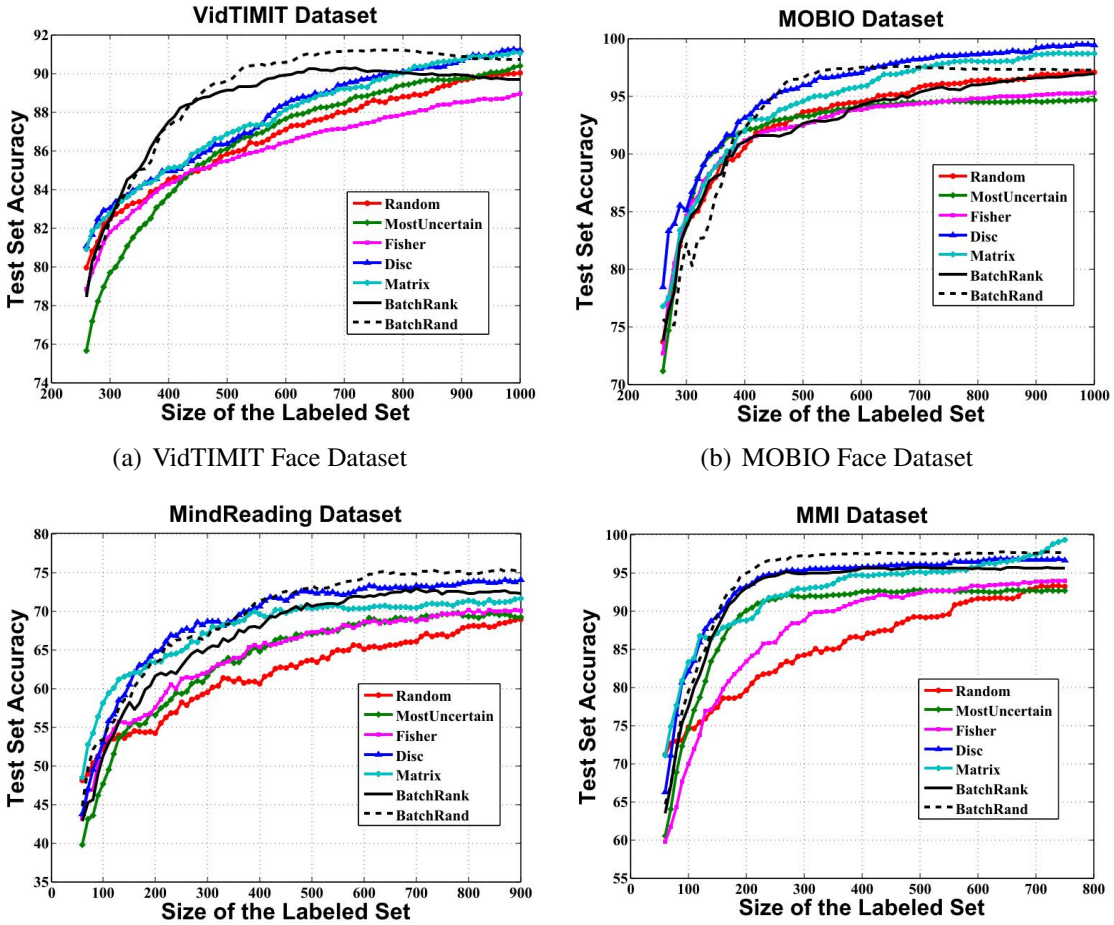
(a) Breast Cancer (b) Heart (c) Musk

(d) Spect (e) Wine (f) Waveform

(g) Vehicles (h) Image Segmentation (i) Digits

Figure 5.1: Batch Mode Active Learning on the UCI datasets. (Best viewed in color.)

Figure 5.2 depicts the results on the face recognition and facial expression recognition datasets. We note that our algorithms once again depict comparable performance as the state-of-the-art (the BatchRand method in fact, marginally outweighs Disc and Matrix on the VidTIMIT and MindReading datasets). We also note that the random sampling method may sometimes depict good performance, as in the VidTIMIT and MOBIO datasets. However, it is not consistent and performs poorly in the other datasets.



(a) VidTIMIT Face Dataset

(b) MOBIO Face Dataset

(c) MindReading Expression Dataset

(d) MMI Expression Dataset

Figure 5.2: Batch Mode Active Learning Face Recognition and Facial Expression Recognition datasets. (Best viewed in color.)

Experiment 2: Batch Mode Active Learning on Multi-Label Datasets

Multi-label learning is a generalization of conventional single-label learning, where each data sample can have multiple labels associated with it. Manual annotation of samples is even more difficult in a multi-label application, as the user has to scan through all the possible labels to decide the label set of a particular example. Thus, batch mode active learning is of paramount importance in such settings. The proposed BatchRank and BatchRand frameworks are flexible and can be extended to multi-label contexts. We demonstrate their performance on two benchmark multi-label datasets in this section. For

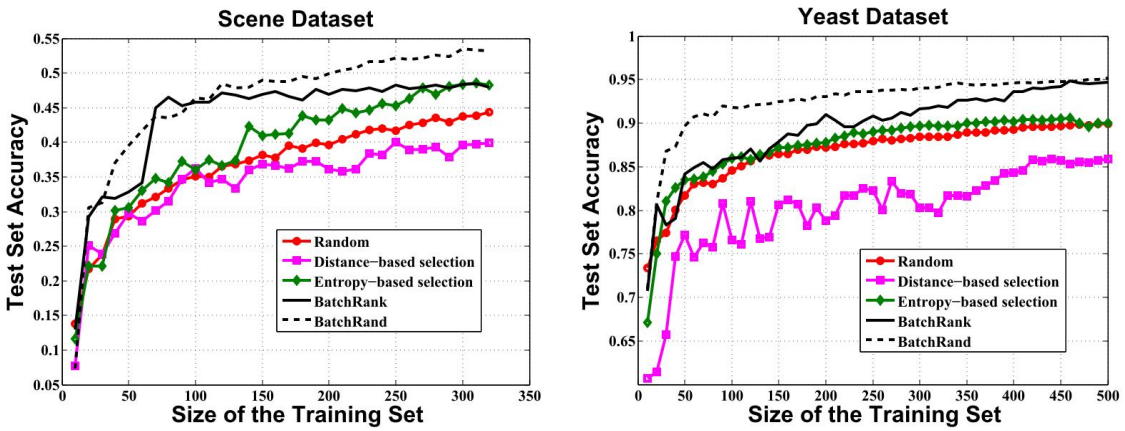
this purpose, the entropy term in the matrix D was modified and was computed as the average entropy over the individual classes:

$$S(Y|x_i, w^t) = \frac{1}{|Y|} \sum_{j=1}^{|Y|} [p_j^i \log p_j^i + (1 - p_j^i) \log(1 - p_j^i)]$$

The proposed approaches were compared against (i) *Random Selection*, (ii) *Distance based Selection*, in which an SVM was trained for each possible class and the k closest unlabeled points (k being the batch size) to all the hyperplanes in the feature space were selected for annotation, (as in [48]) and (iii) *Entropy based Selection*, where the entropy was computed for every unlabeled instance and the top k points were queried based on the entropy ranking (as in [47]). A polynomial kernel SVM was used as the base classifier because of its established performance in multi-label learning [48]. We used the Scene and the Yeast datasets in our experiments. The natural scene dataset contains 2407 natural images belonging to one or more of the six natural scene categories including beach, sunset, foliage, field, mountain and urban. The images are first converted into CIE Luv color space and then the first and second color moments (mean and variance) are extracted over a 7×7 grid on the image, resulting in a 294 dimensional feature vector [259]. The Yeast dataset consists of micro-array expression data and phylogenetic profiles with 2417 genes. Each gene in the set belongs to one or more of the 14 different functional classes and is represented as a 103 dimensional feature vector. Further details about this biological dataset can be found in [260]. The results on these two datasets are shown in Figure 5.3 and corroborate the conclusions drawn in the previous experiments, with BatchRand depicting the best performance.

Experiment 3: Run Time Analysis

In this section, we perform an analysis of the computation time of each of the batch mode active learning algorithms. Table 5.2 (binary and multi-class datasets) and Table 5.3 (multi-label datasets) report the average time taken to query a batch of samples from



(a) Multi-label Scene dataset (b) Multi-label Yeast dataset

Figure 5.3: Multi-label Batch Mode Active Learning on the Scene and Yeast datasets. (Best viewed in color.)

Dataset	US	R	MU	F	D	M	BRank	BRand
Wine	87	0.01	0.06	0.49	6.92	1.02	0.22	0.86
Spect	110	0.01	0.11	0.78	1.18	2.34	0.21	3.91
Heart	120	0.01	0.08	0.33	5.21	1.80	0.24	0.37
Breast Cancer	259	0.01	0.23	0.46	16.23	5.78	0.40	0.49
Image Segmentation	300	0.01	0.25	0.64	12.09	3.61	0.69	1.44
Musk	500	0.01	0.12	1.03	96.14	35.51	1.35	2.17
Vehicles	500	0.01	0.18	0.97	27.04	9.87	1.39	2.94
Waveform	1000	0.01	0.58	1.03	296.43	122.34	3.57	11.38
Handwritten Digits	1000	0.01	0.53	1.56	977.28	234.78	7.35	19.47
VidTIMIT	1000	0.01	1.78	3.92	923.65	171.88	14.27	26.08
MOBIO	1000	0.01	1.18	2.37	757.43	204.57	11.92	24.69
MindReading	1000	0.01	1.48	1.78	88.71	12.23	6.19	13.32
MMI	1000	0.01	0.44	1.86	73.94	129.77	6.43	11.94

Table 5.2: Time taken (in seconds) to query a batch of samples from an unlabeled set. (Binary and Multi-class Datasets).

an unlabeled set for all the algorithms (here, US: Unlabeled Set Size, R: Random, MU: MostUncertain, F: Fisher, D: Discriminative, M: Matrix Partition, BRank: BatchRank, BRand: BatchRand). We note that random selection and uncertainty based selection are the most efficient in terms for running time. The Fisher information based selection framework also has low computation time. BatchRank and BatchRand surpass Disc by a substantial margin in terms of running time. The improvement is more prominent in case of the larger datasets (VidTIMIT, MOBIO and Handwritten Digits) thus demonstrating their

Dataset	US	R	B	MU	BRank	BRand
Scene	350	0.01	0.72	1.39	0.77	2.28
Yeast	600	0.01	0.54	1.89	2.58	5.74

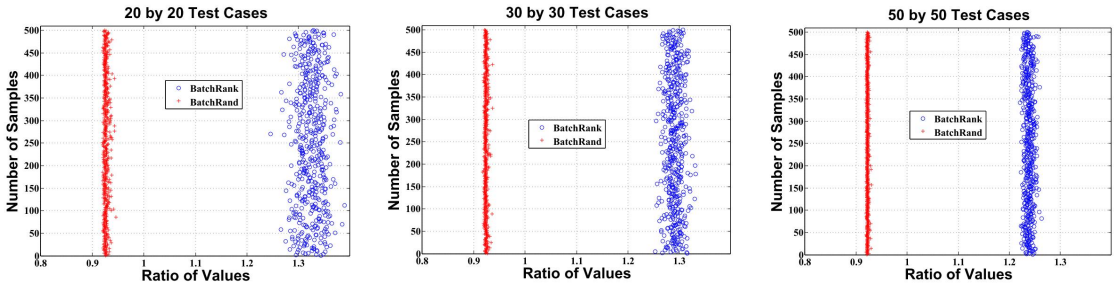
Table 5.3: Time taken (in seconds) to query a batch of samples from an unlabeled set (Multi-label Datasets).

potential for large scale learning. The Disc algorithm involves intensive computation and classifier retraining as part of the optimization process, which adversely affects its running time. The Matrix algorithm is better than Disc (as observed in [217]) but is slower as compared to BatchRank and BatchRand. The results unanimously point to the conclusion that the proposed approaches deliver comparable performance as the state-of-the-art BMAL algorithms at a significantly lower computation time.

Experiment 4: Validation of the Solution Bounds

To empirically validate the solution bounds of the proposed algorithms, we generated random symmetric matrices D, with $d_{ij} \geq 0$. We derived the optimal solutions m^* and $\widehat{m}$ by solving the integer programming problem in (5.4) and the relaxed formulation in (5.8) respectively for the BatchRank algorithm. Note that the ILP formulation in (5.4) can be solved exactly for small-scale problems. The ratio $\frac{f(\widehat{m})}{f(m^*)}$ was computed for the given matrix D and for a specific batch size k, where $f(.)$ was the function defined in Equation (5.9). This was necessary because the solution bound was proved on the objective $f(.)$ and not on the original objective defined in Equation (5.4). We also computed the ratio $\frac{E(W)+\widehat{D}_{total}}{\sum \widehat{d}_{ij}v_iv_j+\widehat{D}_{total}}$ as described in Theorem 2 for the BatchRand algorithm. Figure 5.4 shows sample results obtained on 3 test cases with different matrix dimensions (the batch size k was taken as 10 in all cases). Each graph shows the ratio $\frac{f(\widehat{m})}{f(m^*)}$ and $\frac{E(W)+\widehat{D}_{total}}{\sum \widehat{d}_{ij}v_iv_j+\widehat{D}_{total}}$ for 500 different matrices of the same dimension. For the BatchRank framework, it is evident that the ratio of the functional values is less than 2 in all cases, which validates the bound established in Section 5.2. Moreover, in all the test cases, the ratio is only slightly greater

than 1 (in the range 1.2 to 1.4), which shows that the functional value obtained using the proposed method is very close to that obtained using the optimal solution. In case of BatchRand, we note that the ratio is always greater than 87% validating the bound proved in Theorem 2. Further, we note that the ratio is actually very close to 93% in most cases depicting that the solutions obtained in practise are much better than the theoretical guarantee. We therefore conclude that both BatchRank and BatchRand produce high quality approximations of the corresponding relaxations.



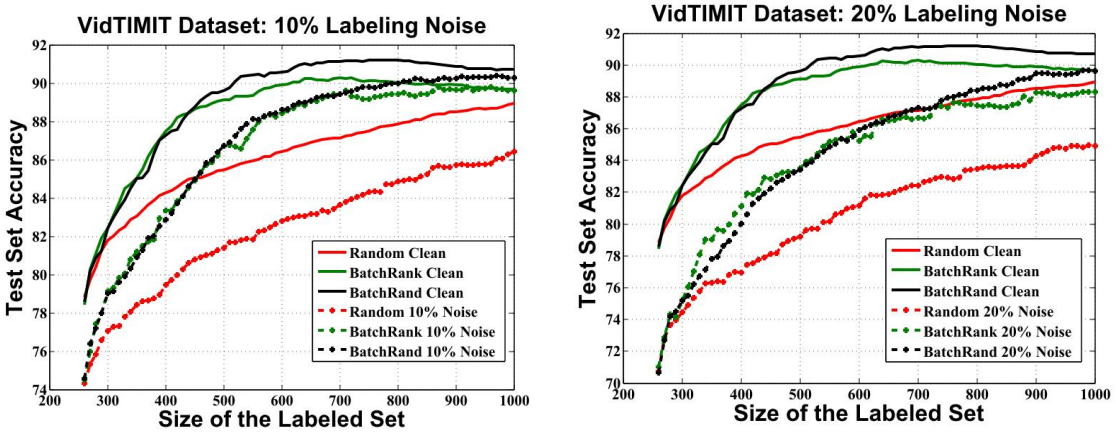
(a) 20×20 Test Cases (b) 30×30 Test Cases (c) 50×50 Test Cases

Figure 5.4: Validation of Solution Bounds of BatchRank and BatchRand. (Best viewed in color.)

Experiment 5: Noise Sensitivity

In many real-world scenarios, the labels of the data samples are noisy, either due to errors in data collection, or even because of human annotation errors. In this section, we study the label-noise sensitivity of the BatchRank and BatchRand algorithms. To simulate the situation, we artificially imparted stochastic labeling noise to the unlabeled samples. An $n\%$ noise implies that the samples are randomly given an incorrect label with a probability of $n\%$. The labels of the test set were kept unchanged and the algorithms were run on clean as well as noisy data. We compared the proposed algorithms against random selection on the VidTIMIT dataset.

The results are depicted in Figure 5.5, which plots the active learning curves for BatchRank, BatchRand and Random sampling with 10% and 20% labeling noise. As expected, the classification accuracy reduces in the presence of noise. But, from Figure 5.5(a), we note that, even with 10% labeling noise, the accuracy of both BatchRank and BatchRand drops only marginally as compared to the values on clean data and the final accuracy matches very closely to that obtained using clean data. Further, both the methods outperform random sampling on the same amount of noisy data and even on clean data. Even with 20% labeling noise, the final accuracy values of BatchRank and BatchRand are very close to those obtained using clean data. These results corroborate the fact that the proposed algorithms are robust to a significant amount of labeling noise.



(a) 10 percent label noise

(b) 20 percent label noise

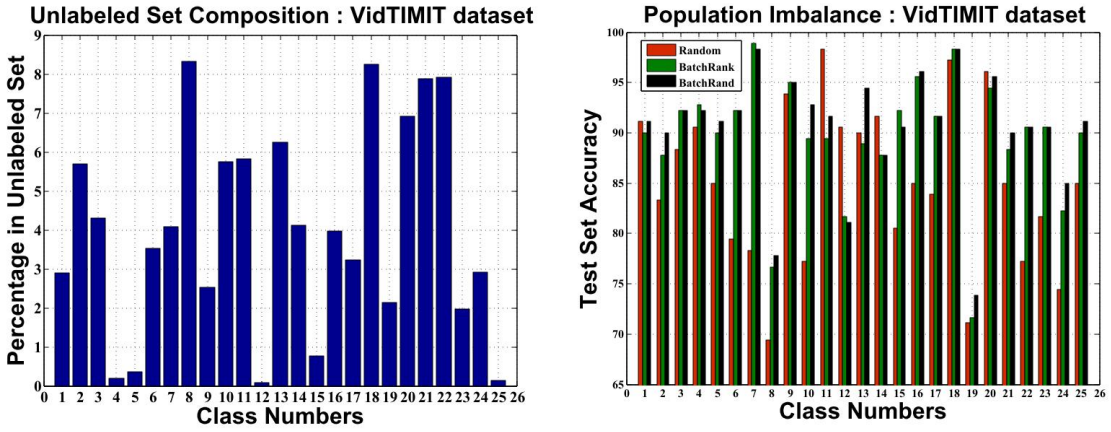
Figure 5.5: Noise Sensitivity of BatchRank and BatchRand on the VidTIMIT dataset. (Best viewed in color.)

A possible reason for this is the fact that both the BatchRank and BatchRand algorithms select unlabeled samples for query which are the most uncertain (and also diverse) with respect to the current classification model. Considering a binary classification problem, the uncertain samples typically lie close to the decision hyperplane. If the user gives an incorrect label of such an unlabeled sample, it will have minimal effect on the orientation of the decision boundary. In contrast, an incorrect label of a sample deep inside the

region of a certain class can have much more severe effects on the decision boundary. The same argument can be extended to a multi-class classification problem with C classes by treating it as C binary classification problems using the one-vs-rest approach. Thus, the batch selection criterion of the proposed frameworks endow them with the capability to counter label noise and deliver high classification accuracy.

Experiment 6: Population Imbalance

In real-world data, there is often a large variation in the number of samples belonging to the different classes. In this section, we study the performance of BatchRank and BatchRand to counter population imbalance. We once again present the results on the VidTIMIT dataset and compare the proposed algorithms against random selection. To simulate the real world scenario, an unlabeled pool of samples, with largely varying number of instances per class, was presented to the active learner for batch selection. The test set was kept unchanged. Figure 5.6(a) depicts the percentage of images of each of the 25 classes that were present in the unlabeled pool.

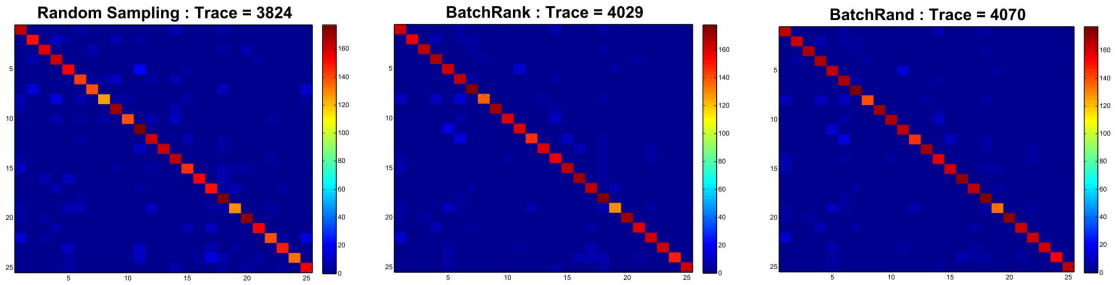


(a) Percentage of images of each class in the unlabeled set

(b) Classwise Accuracy

Figure 5.6: Population Imbalance : VidTIMIT dataset. (Best viewed in color.)

Random selection suffers since it does not integrate the composition of the unlabeled pool relative to the current training set in the batch selection process. In contrast, the proposed frameworks accurately identify the salient instances from the unlabeled set based on the uncertainty and divergence criteria. Thus, regardless of the composition of the unlabeled set, they append useful information to the underlying classification model and consistently deliver high accuracy on the individual classes. This is evident from Figure 5.6(b) which plots the accuracy of each class for random sampling and the active learning methods. We see that the proposed algorithms depict better performance than random selection in 20 of the 25 classes. They also comprehensively outperform random sampling in the overall classification accuracy, as evident from the confusion matrices in Figure 5.7. We note that BatchRank and BatchRand lead to much lower confusion among the classes and beat random sampling by about 6%. This emaphasizes their robustness to population imbalance.



(a) Random Selection (b) BatchRank (c) BatchRand

Figure 5.7: Population Imbalance : Confusion matrices for Random Selection, BatchRank and BatchRand (Max trace = 4500) : VidTIMIT dataset.

Experiment 7: Visual Demonstration

In this section, we present a visual demonstration of the efficacy of the proposed algorithms in selecting salient images from video streams containing redundant data. To this end, we applied BatchRand and BatchRank on the video "si1909" of subject "fadg0" from

the VidTIMIT dataset and the algorithms were required to select a batch of 10 frames from the video (containing about 117 frames). The images selected by BatchRand, BatchRank and random sampling are shown in Figures 5.8, 5.9 and 5.10 respectively. It is evident that random selection captured a much lesser variation in facial appearances as compared to the proposed techniques.

Figure 5.8: Batch of images selected using BatchRand

Figure 5.9: Batch of images selected using BatchRank

Figure 5.10: Batch of images selected using Random Sampling

To study the performance more objectively, the total pairwise distance was computed between all the selected images for these methods. Our preliminary experiments

confirmed the efficacy of the DCT feature in ensuring varied facial appearances are at a larger distance than similar looking faces. Thus, a higher value of the total pairwise distance would mean that images with multifarious appearances were selected. Selection of redundant images will reduce the value of the total pairwise distance. The results of pairwise distance computation are presented in Figure 5.11. We note that BatchRank and BatchRand selected a more diverse set of images compared to random sampling, as depicted by the total pairwise distance values.

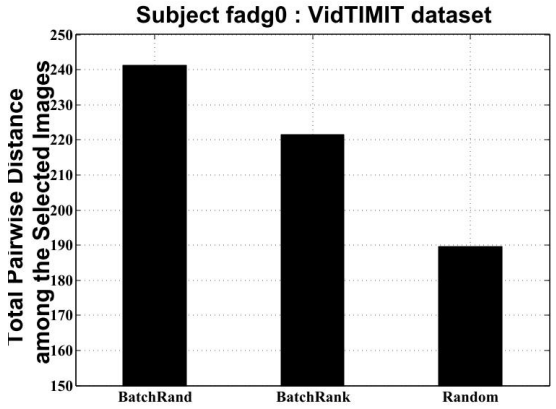


Figure 5.11: Total pairwise distance among the selected images

5.6 Discussions

In this chapter, we proposed two novel batch mode active learning algorithms called BatchRank and BatchRand. Starting with an NP-hard optimization problem, we derived two convex relaxations and established bounds on the solution quality of each relaxation. Our empirical results on several challenging binary, multi-class and multi-label datasets corroborate the fact that the proposed approaches perform at par with the state-of-the-art BMAL techniques and also deliver high quality solutions. Our results also verified the proved theoretical bounds. We further demonstrated the robustness of the proposed algorithms to real-world issues like label noise and population imbalance.

Chapter 6

ACTIVE MATRIX COMPLETION

Recovering a matrix from a sampling of its entries is a problem of rapidly growing interest and has been studied under the name of *matrix completion*. It occurs in many areas of engineering and applied science. However, considering the enormity of data in the modern era, manually completing all the entries in a matrix will be an expensive process in terms of time, labor and human expertise. It is therefore natural to extend the idea of active learning to the matrix completion problem. To this end, appropriately identifying a subset of missing entries (for manual annotation) in an incomplete matrix is critically important; this can potentially lead to better reconstructions of the incomplete matrix with minimal human effort. In this chapter, we propose novel algorithms to address this issue. Since the query locations are actively selected by the algorithms, we refer to these methods as *active matrix completion* algorithms. To the best of our knowledge, this is the first effort to develop algorithms which can incorporate human supervision in actively completing a matrix. Such a framework will also be immensely useful in problems like multi-class active learning, multi-label active learning, transductive active learning and active feature acquisition, where the data can be represented as a matrix of observations and selective human supervision can be exploited to solve the problem in question.

6.1 Background

The data collected in most modern applications are mostly structured in the form of matrices (for instance, a grayscale image is a matrix of pixel intensity values; in a recommendation system, the data is represented in the form of a matrix, where each row is a user, each column is an object and the corresponding entry represents the rating given by the particu-

lar user to that object). The problem of matrix completion, or reconstructing a matrix from a set of partially observed entries, is of immense practical importance. Missing values occur due to a number of reasons, including ignorance of the concerned individual about the value of the entry in question or unwillingness of a subject to divulge some sensitive information. It is a recurrent problem in applications like machine learning [218], computer vision and graphics [219], recommendation systems [220], clinical research [261], DNA microarray analysis [262] and climatic data analysis [263] among others. Missing data in any of these applications can bias results, reduce generalizability and lead to erroneous conclusions. The ultimate consequence of missing data is distortion from the truth, reducing the validity of study results. For instance, in a hypothetical study of the course of dementia, persons who become unable to follow directions may not complete formal cognitive testing and will have missing test scores. Over time, as those who are unable to complete the tests do not contribute data, the mean and range of cognitive test scores will appear better than they really are, which can mislead data analysts. Such problems have motivated active research in the area of matrix completion to estimate the missing entries in a matrix from a subset of observations. Several matrix completion algorithms have been proposed over the last couple of decades to address this practical and useful challenge [264, 265, 266, 267].

Of late, there has been a growing interest in the development of machine learning algorithms with "humans-in-the-loop", (such as active learning) which have shown tremendous promise in the development of reliable classification and regression models. It is therefore logical to conceive of the development of matrix completion algorithms with selective human supervision. This can potentially lead to better reconstruction of the partially observed data matrix. As an exemplar application, consider a movie recommendation system where the data is organized in the form of a matrix with rows denoting users, columns denoting movies and entries in the matrix representing the rating given by

a particular user to a particular movie. In such a system, vendors provide recommendations based on the users' preferences. However, each user rates only a few movies and such a matrix happens to be extremely sparse. To infer the preference of a user on an unrated movie (in order to provide an efficient recommendation), an accurate reconstruction of the sparse matrix is imperative. A framework which can identify the missing entries in the matrix having maximal prediction uncertainty will be of immense use to facilitate completion of the matrix in such a situation. Given a set of such entries in the matrix, the corresponding users can be requested to provide ratings for the corresponding movies. Knowledge of these uncertain entries can enable a better reconstruction of the ratings matrix, which in turn, can lead to a better inference regarding the preference of a user on an unrated movie.

In this chapter, we propose novel algorithms to integrate the intelligence of human oracles in the matrix completion problem. We develop frameworks to quantify the uncertainty of prediction of each missing entry in the matrix, which can be used to identify the indices which need to be labeled manually. The frameworks are generic and can be used in conjunction with any popular matrix completion algorithm (EM/SVD and others). Further, we demonstrate how the proposed frameworks can be adapted to different variants of active learning (multi-class, multi-label, transductive active learning, active feature acquisition) to select a set of exemplar unlabeled data instances for manual annotation. We hope that this work will serve as a motivation for the development of matrix completion algorithms with "human-in-the-loop" and their adaptations in other interactive problem settings. We first present a brief survey of matrix completion algorithms followed by our active matrix completion frameworks.

6.2 Matrix Completion : A Brief Survey

Most matrix completion algorithms usually assume that the underlying data matrix has low rank. The problem of low rank matrix completion has been addressed in the context of machine learning [218], computer vision [219] and recommendation systems [220], among others. Low rank matrix completion is typically posed as the following optimization problem:

$$\begin{aligned} &\min_{X} \frac{1}{2}||X-M||_{\Omega}^{2} \\ &s.t.\, \text{rank}(X) \leq r \end{aligned} \tag{6.1}$$

where $X, M \in \Re^{p \times q}$ and the elements of M in the set Ω are given while the remaining elements are missing. Fazel *et al.* [268] heuristically used the matrix trace norm to approximate the rank of matrices. Later, Candes and Recht [269, 270] theoretically justified the usage of trace norm as an approximation of the matrix rank. Srebro *et al.* [271] employed second order conic programming (SOCP) to formulate a trace norm related problem in matrix factorization. However, the SOCP based approach was computationally intensive and did not scale to large matrices. To address this issue, Ma *et al.* [267] applied the fixed point and Bregman iterative methods to solve the rank minimization problem for large scale matrices. Cai *et al.* [266] proposed a singular value thresholding algorithm to address the same problem of large scale matrix completion. Candes and Tao [264] presented optimality results quantifying the minimum number of entries needed to recover a matrix of rank r exactly by any method. Candes and Recht [269] proved that most low rank matrices can be exactly recovered from most sets of sampled entries, even though these sets have small cardinality. They also proved that this can be achieved by solving a convex optimization problem. Recht [265] further improved on these results to provide a bound on the number of entries required to reconstruct a low rank matrix which is optimal upto a small numeric constant and one logarithmic factor. Hastie *et al.* [272] proposed three

methods for matrix completion - one based on singular value decomposition (SVD), one on nearest neighbor averaging and a third on repeated regressions. Schneider [263] proposed a method to impute the missing values using the Expectation Maximization (EM) algorithm. Liu *et al.* [273] extended the idea of matrix completion to tensors and proposed a method for estimating missing values in visual data. Yi *et al.* [274] proposed a matrix completion based approach for crowdclustering, where m users were asked to complete the similarity matrix, the set of entries which were not agreed upon by the majority were treated as missing and were imputed using a matrix completion algorithm. This framework, however, was not focused on actively selecting a subset of uncertain entries in the matrix to query their ground truth values.

The fundamental idea behind the proposed algorithms is to compute a measure of uncertainty of prediction of every missing entry in the incomplete data matrix. The top uncertain entries can then be queried for manual annotation. We present three strategies to quantify the uncertainty of prediction and consequently, three active matrix completion algorithms. We first present the mathematical details of these algorithms; we then present methodologies to improve the computational efficiency of our frameworks, so as to make them scalable to large datasets.

6.3 Active Matrix Completion using the Conditional Gaussian Distribution

This method treats each row (or column, as the case may be) of the data matrix as a particular case (or sample).For each case, it is assumed that the set of missing entries conditioned on the set of observed entries follows a multivariate normal distribution. A well-known result from statistical learning theory enables us to compute the mean and the covariance matrix of this conditional distribution. The overarching idea is to impute the missing entries of each case with the conditional mean vector while the diagonal elements of the covariance matrix of the conditional distribution quantifies the variance (uncertainty) as-

sociated with each imputation. The top k uncertain entries across the entire matrix (where k is the batch size or the allowable number of entries that can be queried) are then selected for manual annotation. Our algorithm is based on the GLasso and MissGlasso frameworks that have been proposed for sparse inverse covariance estimation of the multivariate normal distribution in the presence of missing entries. We now present the details of these methods.

GLasso

Consider a variable X (of dimensionality p) which follows the multi-variate normal distribution with mean vector μ and covariance matrix Σ that is, $X \sim N(\mu, \Sigma)$. The problem of Graphical Lasso (or GLasso) [275] is to estimate the parameters μ and Σ from a complete random sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$. Let K denote the concentration matrix, $K = \Sigma^{-1}$. A typical approach is to minimize the negative ℓ_1-penalized log-likelihood [275]:

$$-\ell(\mu, K; x) + \lambda ||K||_1 = -\frac{n}{2} \log |K| + \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T K (x_i - \mu) + \lambda ||K||_1 \tag{6.2}$$

where K is a positive semidefinite matrix, $\lambda > 0$ is a tuning parameter and $||K||_1 = \sum_{j,j'=1}^{p} |K_{jj'}|$. The minimizer $\hat{K}$ can be obtained by solving the following optimization problem:

$$\hat{K} = \underset{K \succ 0}{\arg\min} (-\log |K| + tr(KS) + \rho ||K||_1) \tag{6.3}$$

where $S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$ and $\rho = \frac{2\lambda}{n}$. Friedman *et al.* [276] proposed an efficient algorithm called GLasso to solve the optimization problem (6.3).

MissGlasso

This algorithm was proposed by Stadler and Buhlmann [277] to estimate the mean and covariance matrix of the multi-variate normal model in the presence of missing entries.

Let $\mathbf{x} = (\mathbf{x}_{obs}, \mathbf{x}_{mis})$ denote the set of observed values and missing values respectively of a random sample $\mathbf{x}$ of size n. Further, let $\mathbf{x}_{obs} = (x_{obs,1}, x_{obs,2}, \ldots, x_{obs,n})$ where $x_{obs,i}$ represents the set of observed variables for case $i, i = 1, \ldots n$. The likelihood function is now based on the observed indices for each case and is summed over all the cases:

$$\ell(\mu, \Sigma; \mathbf{x}_{obs}) = -\frac{1}{2} \sum_{i=1}^{n} (\log |\Sigma_{obs,i}| + (x_{obs,i} - \mu_{obs,i})^T (\Sigma_{obs,i})^{-1} (x_{obs,i} - \mu_{obs,i})) \quad (6.4)$$

where $\mu_{obs,i}$ and $\Sigma_{obs,i}$ are the mean and covariance matrix of the observed components of X for case i. Equivalently, in terms of K:

$$\ell(\mu, K; \mathbf{x}_{obs}) = -\frac{1}{2} \sum_{i=1}^{n} (\log |(K^{-1})_{obs,i}| + (x_{obs,i} - \mu_{obs,i})^T (K^{-1}_{obs,i})^{-1} (x_{obs,i} - \mu_{obs,i})) \quad (6.5)$$

Similar to Equation (6.2), the inference for μ and K are now based on the sum of the observed log-likelihood over all the cases, together with an ℓ_1 penalty on the concentration matrix K:

$$\hat{\mu}, \hat{K} = \underset{(\mu, K): K \succ 0}{\arg\min} \; -\ell(\mu, K; \mathbf{x}_{obs}) + \lambda ||K||_1 \quad (6.6)$$

This problem can be solved using a well-known theorem on partitioned Gaussians, which can be stated as follows [232]. Consider a joint Gaussian distribution $N(x|\mu, \Sigma)$ with $K = \Sigma^{-1}$ and consider a partition

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}, \quad K = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}$$

Then, x_a conditioned on x_b will also follow a normal distribution and its mean and covariance can be expressed in terms of the known parameters [232], that is $X_{a|b} \sim N(\mu_{a|b}, \Sigma_{a|b})$ where

$$\mu_{a|b} = \mu_a - K_{aa}^{-1} K_{ab}(x_b - \mu_b) \tag{6.7}$$

$$\Sigma_{a|b} = K_{aa}^{-1} \tag{6.8}$$

Assuming that for each case in the data, the set of missing entries conditioned on the set of observed entries follows a multivariate normal distribution, the problem in Equation (6.6) can be solved using the results stated above together with the Expectation Maximization (EM) algorithm. The complete data x follows a multivariate normal distribution which belongs to the exponential family with sufficient statistics

$$T_1 = x^T.1 = \left(\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \ldots, \sum_{i=1}^{n} x_{ip} \right)$$

and

$$T_2 = x^T x$$

The log-likelihood for the complete data, given in Equation (6.2) can be expressed in terms of the sufficient statistics T_1 and T_2 as follows:

$$-\ell(\mu, K; x) + \lambda ||K||_1 = -\frac{n}{2} \log |K| + \frac{n}{2} \mu^T K \mu - \mu^T K T_1 + \frac{1}{2} tr(K T_2) + \lambda ||K||_1 \tag{6.9}$$

which is linear in T_1 and T_2. The expected complete penalized log-likelihood is denoted by:

$$Q(\mu, K | \mu^{'}, K^{'}) = -\mathbb{E}[\ell(\mu, K; x) | x_{obs}, \mu^{'}, K^{'}] + \lambda ||K||_1$$

The EM algorithm works by iterating between the E step and the M step. Let (μ^m, K^m) denote the parameter values in iteration m.

E step: In the expectation step, the expected value of the complete penalized log-likelihood is computed. As the complete penalized log-likelihood is linear in terms of the sufficient statistics T_1 and T_2, the E step essentially consists of calculating the expected

values of the parameters T_1 and T_2 based on the current observations together with the current values of the mean vector and the covariance matrix:

$$T_1^{m+1} = \mathbb{E}[T_1 | x_{obs}, \mu^m, K^m]$$
$$T_2^{m+1} = \mathbb{E}[T_2 | x_{obs}, \mu^m, K^m]$$

To evaluate these, we need to compute the conditional expectation of x_{ij} and $x_{ij}x_{ij'}$ for $i = 1, \ldots, n$ and $j, j' = 1, \ldots, p$. Assuming for each case, the set of missing entries conditioned on the set of observed entries is normally distributed, Equation (6.7) gives

$$\mathbb{E}[x_{ij} | x_{obs,i}, \mu^m, K^m] = \begin{cases} x_{ij} & \text{if } x_{ij} \text{ is observed} \\ c_j & \text{if } x_{ij} \text{ is missing} \end{cases}$$

where the vector c is defined as (from Equation (6.7))

$$c = \mu^m_{mis} - (K^m_{mis,mis})^{-1} K^m_{mis,obs}(x_{obs,i} - \mu^m_{obs})$$

Here, $K_{mis,mis}$ is the sub-matrix of K with rows and columns corresponding to the missing entries for case i. $K_{mis,obs}$ is defined analogously. Similarly,

$$\mathbb{E}[x_{ij}x_{ij'} | x_{obs,i}, \mu^m, K^m] = \begin{cases} x_{ij}x_{ij'} & \text{if } x_{ij} \text{ and } x_{ij'} \text{ are observed} \\ x_{ij}c_{j'} & \text{if } x_{ij} \text{ observed } x_{ij'} \text{ missing} \\ (K^m_{mis,mis})^{-1}_{jj'} + c_j c_{j'} & \text{if } x_{ij}, x_{ij'} \text{ missing} \end{cases}$$

M step: In the maximization step, the expected log-likelihood for the complete dataset is maximized to obtain the update equations for the mean vector and the covariance matrix. Differentiating Equation (6.9) with respect to μ and equating it to 0, we get the update equation for μ as:

$$\mu^{m+1} = \frac{1}{n} T_1^{m+1}$$

Also, from a re-organization of the terms in Equation (6.9), it is evident that the covariance matrix can be updated by solving the following optimization problem:

$$K^{m+1} = \arg\min_{K \succ 0} \left(-\log|K| + tr(KS^{(m+1)}) + \frac{2\lambda}{n}||K||_1 \right)$$

where $S^{m+1} = \frac{1}{n}T_2^{m+1} - \mu^{m+1}(\mu^{m+1})^T$. The M step therefore reduces to a standard GLasso problem, as discussed in Section 6.3. The E step and the M step are iterated until convergence. Once the final matrix K is obtained, the covariance matrix corresponding to the missing entries for a particular case (or data sample) i can be derived as the inverse of the sub-matrix of K with rows and columns corresponding to the missing variables for case i. The diagonal elements of the covariance matrix quantify the variance of prediction of each of the missing entries for that case. Thus, it is possible to estimate the variance (uncertainty) of prediction of each of the missing entries (on a case-by-case basis) in the partially observed matrix. The missing entries are ranked in descending order of their uncertainties and the top k entries (k being the required batch size) are queried for manual annotation.

6.4 Active Matrix Completion using Query by Committee (QBC)

The QBC framework quantifies the prediction uncertainty based on the level of disagreement among an ensemble of matrix completion algorithms. Specifically, a committee of matrix completion algorithms are applied on the partially observed data matrix to impute the missing values. The variance of prediction (among the committee members) of each missing entry is taken as a measure of uncertainty of that entry. The top k uncertain entries are then queried for manual annotation. We used the following three commonly used matrix completion algorithms as members of our committee:

***k*-NN:** The k nearest neighbors (k-NN) method identifies the k most similar features to the current one with a missing value and uses the average of these k nearest neighbors as a guess for the missing one [272].

EM: This method imputes the missing values using the Expectation Maximization (EM) algorithm [263]. An iteration of the EM algorithm involves two steps. In the E step, the mean and covariance matrix are estimated from the data matrix (with the missing entries filled with zeros or estimates from the previous M step); in the M step, the missing value of each data column is filled in with their conditional expectation values based on the available entries and the estimated mean and covariance. The mean and the covariance are re-estimated based on the newly filled matrix and the process is iterated until convergence.

SVD: Singular value decomposition (SVD) is a standard method for matrix completion based on low-rank approximation [272]. In this method, some initial guesses are first provided to the missing data values. SVD is then applied to obtain a low rank approximation of the filled-in data matrix. The missing values are then updated based on their corresponding values in the low rank estimation. SVD is applied to the updated matrix again and the process is iterated until convergence.

6.5 Active Matrix Completion using Committee Stability

This approach is similar to the QBC strategy. However, instead of different matrix completion techniques, we used the same SVD based imputation algorithm with different values of the rank parameter to form the committee. The uncertainty of prediction of each missing entry was computed as the variance of the values from the committee members for that entry, as before. We refer to this method as the stability based active matrix completion algorithm since it essentially measures the regularity of prediction of a particular entry from an ensemble of predictors (similar to the QBC framework).

A general pseudo-code of the three active matrix completion algorithms is presented in Algorithm 7.

Algorithm 7 Active Matrix Completion

Require: A partially observed matrix $M \in \Re^{p \times q}$, set Ω of observed indices, batch size k and number of iterations n

1: **for** *rounds* $= 1 \rightarrow n$ **do**
2: Complete the partially observed matrix and compute the variance of prediction of every missing entry (as detailed in Sections 6.3, 6.4 and 6.5)
3: Sort the missing entries in descending order of their uncertainty (variance) values
4: Query the ground truth values of the top k uncertain entries from human oracles
5: Update the matrix with the newly acquired entries
6: Complete the matrix using any standard matrix completion algorithm
7: Compute the reconstruction error
8: **end for**
9: **return** The completed matrix after n iterations

6.6 Computational Considerations

In Section 6.3, we noted that the active matrix completion algorithm based on Conditional Gaussian distribution involves estimation of the inverse covariance matrix K (in the M step of the EM algorithm). Also, both the QBC and the Stability based algorithms involve low rank matrix completion using the singular value decomposition technique. These can adversely affect the computation time for large scale matrices. In this section, we present two efficient algorithms - one for inverse covariance estimation and the other for low rank matrix completion, to speed up the computations in our algorithms.

Efficient Inverse Covariance Estimation

Sparse inverse covariance estimation is typically achieved using the GLasso algorithm [276], which solves the following problem:

$$\widehat{K} = \underset{K \succ 0}{\arg\min}(-\log|K| + tr(KS) + \lambda||K||_1) \tag{6.10}$$

where S is the sample covariance matrix of dimension $p \times p$. To address the scalability issue, we used the thresholding strategy, proposed by Mazumder and Hastie [278], for

large scale graphical lasso. The authors presented a novel property characterizing the family of solutions to the graphical lasso problem in Equation (6.10) as a function of the regularization parameter λ, which states that the vertex partition induced by the connected components of the non-zero pattern of the estimated concentration matrix (at λ) and the thresholded sample covariance matrix S (at λ) are exactly equal. Specifically, the sparsity pattern of the solution $\widehat{K}^{(\lambda)}$ to (6.10) gives rise to the symmetric edge matrix/skeleton $\in \{0,1\}^{p\times p}$ defined by:

$$E1_{ij}^{(\lambda)} = \begin{cases} 1 & \text{if } \widehat{K}_{ij}^{(\lambda)} \neq 0, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

This defines a symmetric graph $G_1^{(\lambda)} = (V, E1^{(\lambda)})$, and suppose that it admits a decomposition into $m_1(\lambda)$ connected components:

$$G_1^{(\lambda)} = \cup_{l=1}^{m_1(\lambda)} G_{1l}^{(\lambda)} \tag{6.11}$$

where $G_{1l}^{(\lambda)} = (\widehat{V}_l^{(\lambda)}, E1_l^{(\lambda)})$. Now, a thresholding on the entries of the sample covariance matrix S (for a given λ) is performed to obtain a graph edge skeleton $E_2^{(\lambda)} \in \{0,1\}^{p\times p}$ defined by:

$$E2_{ij}^{(\lambda)} = \begin{cases} 1 & \text{if } |S_{ij}| > \lambda, i \neq j \\ 0 & \text{otherwise} \end{cases}$$

The symmetric matrix $E2^{(\lambda)}$ defines a symmetric graph on the nodes $V = \{1,\ldots,p\}$ given by $G_2^{(\lambda)} = (V, E2^{(\lambda)})$. The graph $G_2^{(\lambda)}$ admits a decomposition into connected components given by:

$$G_2^{(\lambda)} = \cup_{l=1}^{m_2(\lambda)} G_{2l}^{(\lambda)} \tag{6.12}$$

where $G_{2l}^{(\lambda)} = (V_l^{(\lambda)}, E2_l^{(\lambda)})$ are the components of the graph $G_2^{(\lambda)}$. Mazumder and Hastie [278] proved that the vertex partition of the connected components of (6.12) is exactly

equal to that of (6.11). This implies that the optimization problem in (6.10) completely separates into $m_2(\lambda)$ separate sub-problems of the same form as (6.10). The sub-problems have size equal to the number of nodes in each component $p_i = |V_i|, i = 1, \ldots, m_2(\lambda)$. Also, the cost of computing the connected components of the thresholded sample covariance graph (6.12) is much smaller than the cost of fitting graphical models (6.10). Further, the computations pertaining to the covariance graph can be done off-line and are amenable to parallel processing. This property of thresholding a graph into connected components facilitates efficient computation of the inverse covariance matrix for large scale data. For more details, please refer to [278].

Efficient Low Rank Matrix Completion

A standard approach to solve the low rank matrix completion problem (Equation (6.1)) is to relax the rank function to its convex surrogate trace norm, which is then solved using singular value decomposition (SVD) operations. To avoid the computational overhead of SVD, some researchers proposed to sacrifice the convexity by local searching. These methods only require light-weighted matrix computations and are scalable. In this work, we used the low rank factorization LMaFit proposed in [279]. The main idea of the method is presented as follows. For a low rank matrix $\mathbf{X} \in \mathscr{R}^{m,n}$ of rank r, where $r < \min(m,n)$, we can represent the matrix by $\mathbf{X} = \mathbf{UV}$ where $\mathbf{U} \in \mathscr{R}^{m,r}$ and $\mathbf{V} \in \mathscr{R}^{r,n}$, we plugin this product form of $\mathbf{X}$ into the objective and solve the following problem:

$$(\mathbf{U}^*, \mathbf{V}^*, \mathbf{Z}^*) = \underset{\mathbf{U},\mathbf{V},\mathbf{Z}}{\arg\min} \|\mathbf{UV} - \mathbf{Z}\|_F^2 \quad \text{s.t.} \mathscr{P}_\Omega(\mathbf{Z}) = \mathscr{P}_\Omega(\mathbf{X}_0)$$

where $\mathbf{Z} \in \mathscr{R}^{m,n}$ is an auxiliary matrix and the required low rank matrix is given by $\mathbf{X}^* = \mathbf{U}^*\mathbf{V}^*$. The formulation can be directly solved by the block coordinate descent algorithm. In each step of the algorithm, two least squares problems need to be solved, which can be effectively reduced to a single least squares problem [279].

Algorithm 8 Block coordinate descent for efficient low rank matrix completion

Ensure: $\mathbf{U}^*, \mathbf{V}^*, \mathbf{Z}^*, \Omega, \mathbf{X}_0$
Require: $\mathbf{V}_0, \mathbf{Z}_0$
  $\mathbf{V}_- = \mathbf{V}_0, \mathbf{X}_- = \mathbf{X}_0$
  while true **do**
    perform QR decomposition on $\mathbf{Z}_-\mathbf{V}_-^T$ and let $\mathbf{Q}$ be its orthogonal basis.
    $\mathbf{U}_+ = \mathbf{Q}$
    $\mathbf{V}_+ = \mathbf{Q}\mathbf{Z}_-$
    $\mathbf{Z}_+ = \mathbf{U}_+\mathbf{V}_+ + \mathscr{P}_\Omega(\mathbf{X}_0 - \mathbf{U}_+\mathbf{V}_+)$
    if convergence **break**
    return $\mathbf{U}_- = \mathbf{U}_+, \mathbf{V}_- = \mathbf{V}_+, \mathbf{Z}_- = \mathbf{Z}_+$
  end while
  return $\mathbf{U}^* = \mathbf{U}_+, \mathbf{V}^* = \mathbf{V}_+, \mathbf{Z}^* = \mathbf{Z}_+$

The improved algorithm is outlined in Algorithm 8. Since QR decomposition is very cheap for the matrices $Z_-\mathbf{V}_-^T$, the major computational cost in Algorithm 8 is the computation of $\mathbf{U}_+\mathbf{V}_+$ for computing $\mathbf{Z}_+ = \mathbf{U}_+\mathbf{V}_+ + \mathscr{P}_\Omega(\mathbf{X}_0 - \mathbf{U}_+\mathbf{V}_+)$. Note that $\mathbf{Z}_+$ is a sparse + low rank structure, thus there is no need to compute the multiplication operation, but directly work on the sparse + low rank structure whenever $\mathbf{Z}_+$ is used. This reduces the computational complexity and makes the method scalable.

6.7 Experiments and Results

In this section, we study the performance of the proposed active matrix completion algorithms. We started with a given data matrix and manually deleted a certain percentage (ranging from 40% to 98%) of the entries at random. The active matrix completion algorithms (refered to as Conditional Gaussian, QBC and Stability for the methods proposed in Sections 6.3, 6.4 and 6.5 respectively) were then applied to query a fixed number of entries in each iteration. After the batch query, the selected locations in the matrix were annotated using a human oracle (we simulated this by supplying the ground truth value of the batch of selected indices). The matrix was then completed using any standard matrix completion algorithm (we used the SVD based completion algorithm [272] in our work).

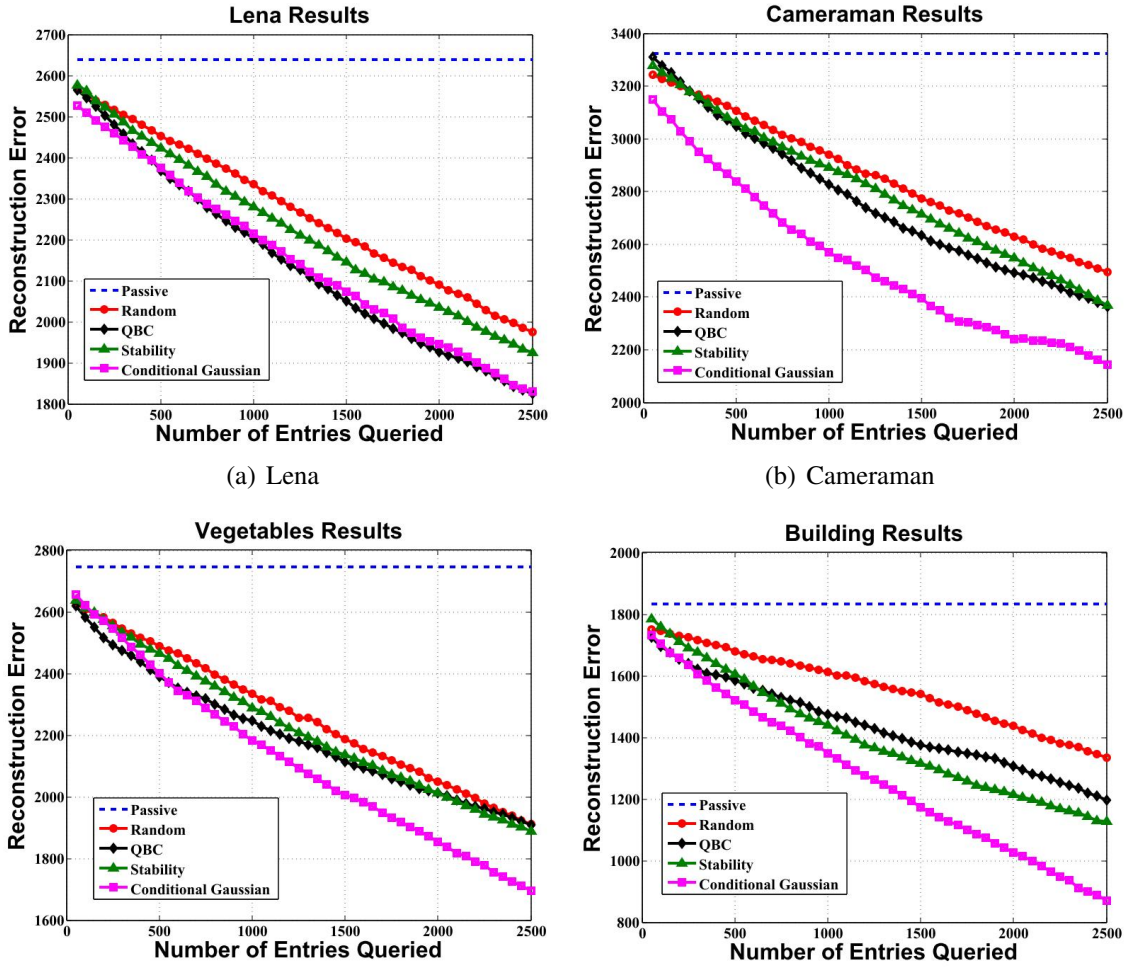
The reconstruction error was then computed as the Frobenius norm of the error matrix (the difference between the original data matrix and the predicted matrix, normalized by the number of indices predicted). The process was then repeated and the reduction in the reconstruction error with increasing number of iterations (equivalently, with increasing number of observed entries) was noted. We compared our results against the case where there was no human intervention and the matrix was merely completed using the SVD algorithm (referred to as *passive* completion in our experiments). We also compared our approaches against the case where the query locations for manual annotation were selected at random.

Experiment 1: Image Datasets

Figure 6.1: GrayScale images used in our experiments

We first conducted experiments on images (represented as grayscale matrices of size 256×256). We used four commonly used images in computer vision research - the Lena, Cameraman, Vegetables and Building images for our study. These images are shown in Figure 6.1. The degree of sparsity (percentage of missing entries to begin with) was set to 60%. The batch size (number of entries to be queried in each iteration) was set to 50 and the process was repeated for 50 iterations. The results were averaged over 5 runs (where the specific positions of the missing entries in the starting matrix were randomly permuted) to rule out the effects of randomness. Since these datasets are relatively small in size, we used the standard SVD method here (and not the algorithm detailed in Section

6.6). For the inverse covariance estimation, the efficient approach proposed by Mazumder and Hastie [278] was used. The results are presented in Figure 6.2.



(a) Lena (b) Cameraman

(c) Vegetables (d) Buildings

Figure 6.2: Active Matrix Completion on Image Datasets (Best viewed in color). Degree of Sparsity = 60%

The x axis denotes the number of queried entries and the y axis denotes the reconstruction error. The objective was to study the rate of decrease of the error as more and more entries in the matrix were labeled. The dotted horizontal line depicts the scenario when no human being is involved in the matrix completion process - the matrix is merely completed using SVD imputation (note that since there is no human labelers for this method, the number of observed entries remain the same, which leads to the same

error value from iteration to iteration; this line is plotted for comparison between active and passive completion). It is evident that active matrix completion results in enormous reduction in the reconstruction error. This corroborates the advantage of leveraging human intelligence in the matrix completion process. We further note that all the proposed algorithms outperform random sampling as the errors drop at a faster rate (with increasing number of observed entries) using the proposed methodologies. The method based on Conditional Gaussians depicts the best performance on these datasets. Further empirical studies (not presented here) revealed that the pattern of the graphs remained the same for different degrees of sparsity (20%, 40% and 80%) - the only difference being the absolute values of the reconstruction errors.

Experiment 2: Recommendation Systems

Recommendation systems is an important application of the proposed active matrix completion algorithms where requesting the rating of a particular object (movie, music, book) from a particular user makes intuitive sense. The proposed algorithms can be used to judiciously exploit human intelligence to facilitate a more accurate reconstruction, which in turn, can help making better recommendations to appropriate users. We used the following recommendation datasets in our experiments: (1) MovieLens 100k, (2) MovieLens 1M, (3) Netflix, all of which contain ratings given by a group of users on a set of movies, (4) Jester, containing the user ratings of jokes and (5) Dating Recommendation, containing anonymous ratings of profiles made by users. These datasets have been extensively used to validate recommendation prediction algorithms [280, 281]. In each dataset, rows represent users, columns represent items and the matrix entry denotes the rating given by a particular user for a particular item. The details of the datasets together with their sparsity levels or the percentage of missing entries, are reported in Table 6.1.

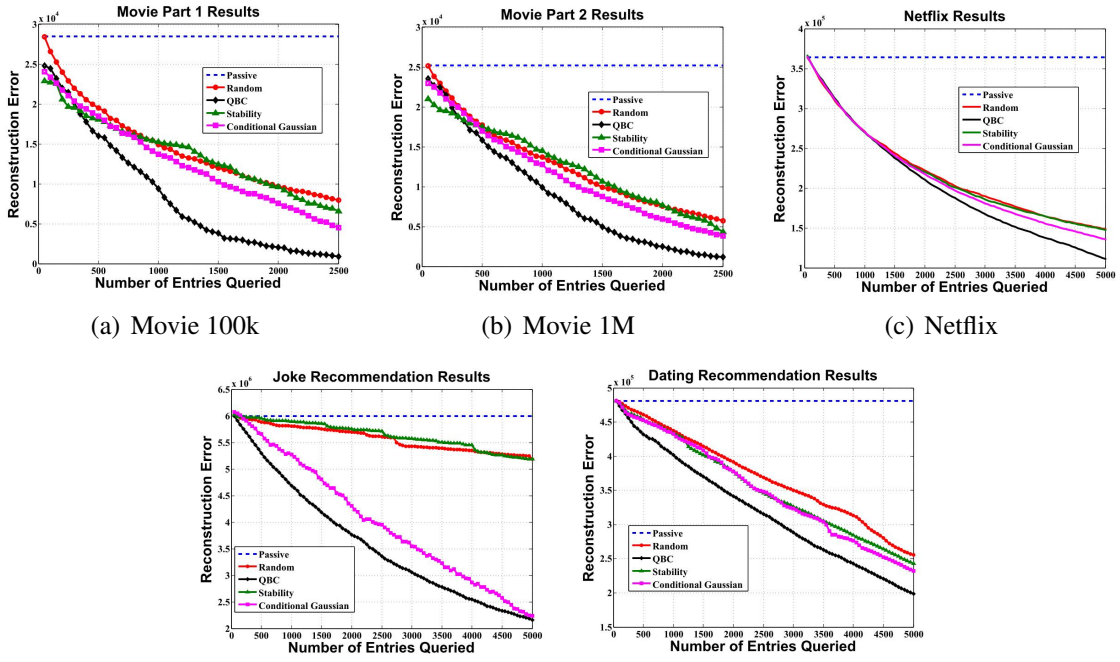
Dataset	Rows	Cols	Matrix Size	Sparsity Level
Movie 100k	943	1682	1.58 M	93.7%
Movie 1M	1000	2000	2 M	96.2%
Netflix	442	8307	3.67 M	97.6%
Jester	5000	100	0.5 M	28.12%
Dating	152	17906	2.7 M	98.2%

Table 6.1: Recommendation Datasets Details

These datasets inherently contain a lot of missing entries as it was not possible to get the ratings of each of the items from all the users. However, to test the performance of our algorithms from iteration to iteration, we need the ground truth values of all the entries in the matrix. To alleviate this issue, we focused only on the set of observed entries in the matrix and manually deleted 50% of the observed entries in each dataset. The active matrix completion algorithms were then run on the entire matrix and the prediction uncertainty values were ranked only on the set of entries which were manually deleted. After supplying the ground truth values of these indices, the matrix was completed and the error was measured only on the observed subset of the matrix. The batch size was set at 50 and the process was repeated for 100 iterations. As before, the results were averaged over 5 random runs. The efficient matrix completion algorithm (Section 6.6) was used in place of the standard SVD (for the QBC and the Stability based active matrix completion algorithms) for these datasets.

Figure 6.3 depicts the performance on these datasets. We once again note that the incorporation of human supervision significantly reduces the reconstruction error (as evident from the dashed line represeting passive completion and the solid lines representing active matrix completion). The QBC and the Conditional Gaussian based algorithms consistently depict good performance across all the datasets; QBC, in fact, marginally outweighs the Conditional Gaussian based framework and achieves the lowest reconstruction error after 100 iterations. The stability based method mostly demonstrates better performance than random sampling, except the Jester dataset, where it performs almost

at-par with random selection. From the results, we also note that random sampling can sometimes depict good performance (as in the Movie 100k dataset). However, it is not consistent across datasets in its performance.



(a) Movie 100k (b) Movie 1M (c) Netflix

(d) Jester (e) Dating

Figure 6.3: Active Matrix Completion on Recommendation Datasets (Best viewed in color)

Experiment 3: Computation Time Analysis

In this section we study the computation time of each of the active matrix completion algorithms to select a batch of entries from a matrix for manual annotation. The results are presented in Table 6.2. Other than Random Sampling, the QBC approach is the most efficient in terms of computation time and can handle a matrix with about 3.67 million entries in less than 2 minutes. The Stability and the Conditional Gaussian based frameworks also depict promising runtime values and can scale to a matrix with 3.67 million entries in approximately 5 minutes (for the dating recommendation dataset however, the Conditional Gaussian based approach took almost 15 minutes for each batch query; the reason

Datasets	Random	QBC	Stability	Conditional Gaussian
Lena	0.91	7.40	17.14	14.89
Cameraman	0.94	7.28	12.56	12.34
Vegetables	0.42	8.72	14.18	15.28
Building	0.89	17.50	17.33	17.09
Movie 100k	1.07	12.59	49.81	32.54
Movie 1M	1.76	17.18	56.77	138.62
Netflix	1.78	93.87	245.69	328.37
Jester	1.97	12.31	30.40	19.56
Dating	1.59	31.92	235.86	883.67

Table 6.2: Average time taken (seconds) to query a batch of indices from the matrix.

behind this needs to be investigated). From this table, it is evident that the QR decomposition based algorithm for efficient matrix completion and the connected components based graph-theoretic method for scalable inverse covariance estimation are effective in significantly reducing the computational overhead of the proposed frameworks. Thus, besides outweighing passive matrix completion and random sampling in terms of error reduction, the proposed active matrix completion algorithms are also efficient computationally and thus have the potential to scale to large datasets.

6.8 Extension to Active Learning Problem Settings

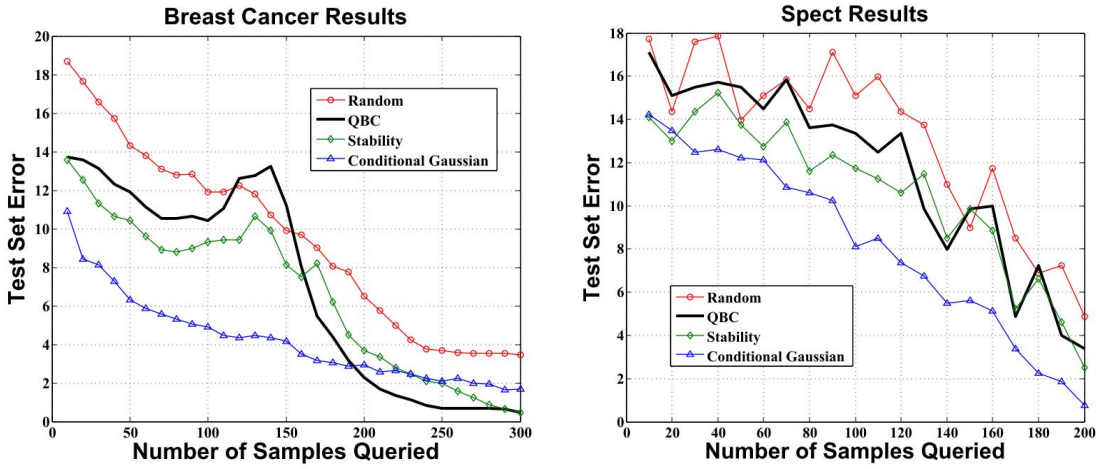
In this section, we demonstrate how the proposed approach can be extended to solve several variants of the active learning problem. We specifically focus on transductive active learning, multi-label active learning, active learning in regression and active feature acquisition. In addressing these problems, each column of the matrix is assumed to constitute a data point with feature values and class labels. Depending on the problem at hand, the active matrix completion frameworks are applied to query informative samples either from the features sub-matrix or from the labels sub-matrix. These features/labels are annotated manually and the reduction in generalization error is studied with increasing amount of information obtained.

Transductive Active Learning

The concept of transductive inference was introduced by Vapnik [282]. The formal problem setting for transductive inference is defined as follows: Given a set of ℓ training pairs $(x_1,y_1),(x_2,y_2),\ldots(x_\ell,y_\ell)$, where $x_i \in \Re^d$, $y_i \in \{-1,1\}$ and a sequence of k test vectors $x_{\ell+1},\ldots,x_{\ell+k}$, find among an admissible set of binary vectors $Y = \{y_{\ell+1},\ldots,y_{\ell+k}\}$, the one that classifies the test vectors with the least number of errors. It is assumed that $x_{\ell+1},\ldots,x_{\ell+k}$ are random i.i.d vectors drawn according to the same (unknown) distribution $P(x)$. The classifications y of the vectors x are defined by some (unknown) conditional probability function $P(y|x)$. A transductive learning problem is different from an inductive problem in at least two respects [283]. Firstly, the learning algorithm does not necessarily have to learn a general rule, it only needs to predict accurately for a finite number of test examples. Thus, it has the obvious advantage of not having to specify a particular learning model (and its parameters) apriori. Second, the test examples are known a priori and can be observed by the learning algorithm during training. This allows the learning algorithm to exploit any information that might be contained in the test examples. Transductive learning is therefore a particular case of semi-supervised learning, since it allows the learning algorithm to exploit the unlabeled examples in the test set. Popular learning algorithms based on transductive inference include transductive support vector machines (TSVMs) [284, 285, 286], the Conformal Predictions Theory [287, 288] and graph-based algorithms [228, 289, 290] among others.

In this section, we apply the active matrix completion algorithms to the problem of transductive active learning. The class labels are represented as $\{-1,1\}$; during prediction, the matrix is completed and if the value of a label entry is positive, it is discretized as 1 and if it negative, it is discretized as -1 (similar to the approach in [291]). The Breast Cancer and the Spect datasets (from the UCI Machine Learning Repository [257])

were used for this experiment. The Breast Cancer dataset has 569 samples with 30 attributes, where each patient is categorized as having or not having cancer. The features were computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. The Spect dataset describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. Each of the patients is classified into two categories: normal and abnormal. The database of 267 SPECT image sets (patients) was processed to extract features that summarize the original SPECT images. As a result, 44 continuous feature pattern was created for each patient. The pattern was further processed to obtain 22 binary feature patterns. 80% of the labels were deleted at random from the two datasets and were treated as unlabeled samples. The results (averaged over 3 runs to rule out the effects of randomness) are depicted in Figure 6.4. The x axis denotes the number of samples queried an the y axis denotes the percentage error on the test set. The batch size was taken as 10 for both the datasets and the process was repeated for 30 and 20 iterations for the two datasets respectively.



(a) Breast Cancer (b) Spect

Figure 6.4: Transductive Active Learning using Active Matrix Completion (Best viewed in color).

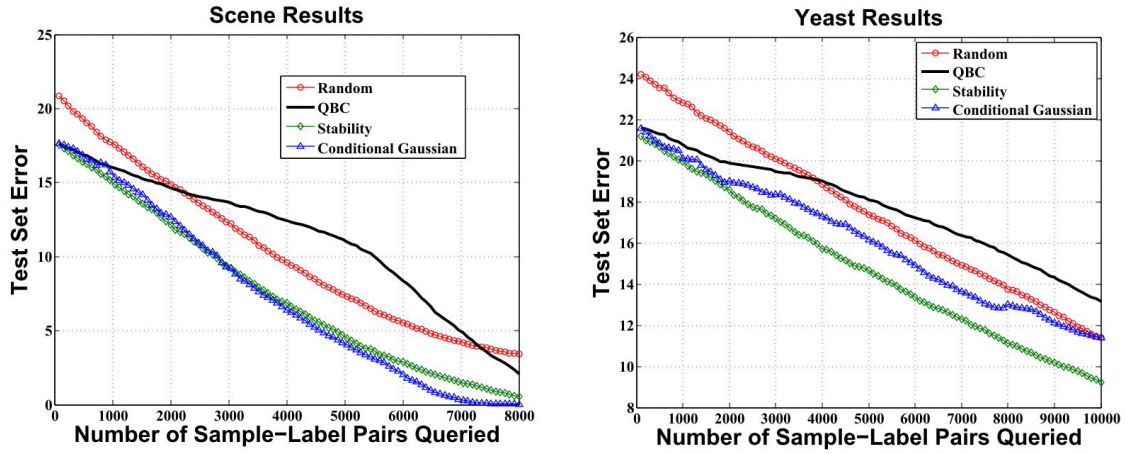
It is evident that the error decreases at a much faster rate for the active selection techniques as compared to random sampling. This certifies the fact that the active matrix completion techniques succeed in querying the salient and exemplar instances and inducing a reliable model with minimal human effort.

Multi-label Active Learning

Multi-label classification is a generalization of conventional classification problems, where each data sample can have multiple labels [45, 46]. Most of the previous work on multi-label active learning use the SVM classifier to quantify the uncertainty of an unlabeled sample (based on distance from the hyperplane or entropy) which is then used for batch selection [48, 50, 47, 49, 51]. However, all these methods request *all the labels* of an unlabeled sample once it is selected for query. In a multi-label learning problem, the labels share an inherent correlation and for each selected sample, only some effective labels need to be annotated while others can be inferred by exploring the label correlations. The contribution of each label in minimizing the classification error is different. Thus, it is important to develop multi-label active learning techniques which query specific sample-label pairs rather than all the labels of a given sample.

An effective way to exploit the label correlations is through low-rank matrix completion [291]. Minimizing the rank of the data matrix provides a natural way to exploit the dependencies among the labels of a multi-label sample. Thus, the proposed active matrix completion methodologies can be judiciously used to query the informative sample-label pairs and efficiently model the label correlations. The Scene (2407 samples, 6 classes, 14442 sample-label pairs) and the Yeast (2417 samples, 14 classes, 33838 sample-label pairs) multi-label datasets were used for this experiment. 60% of the sample-label pairs were deleted at random. The batch size was taken as 80 for the Scene dataset and 100 for the Yeast dataset and the process was iterated over 100 rounds. The results (averaged

over 3 random trials) are presented in Figure 6.5 where the x axis represents the number of sample-label pairs queried and the y axis denotes the test set error. They corroborate our previous findings, emphasizing the efficacy of the proposed algorithms for multi-label active learning.



(a) Scene (b) Yeast

Figure 6.5: Multi-Label Active Learning using Active Matrix Completion (Best viewed in color).

Active Learning in Regression

Contrary to classification problems, where the class labels are discrete, the labels in a regression problem are continuous. In this section, we depict the performance of our active matrix completion algorithms on regression problems (for our algorithms, the matrix entries being discrete or continuous is unimportant - so the exact same methods can be applied to regression settings). We used the FacePix and the AVEC datasets in our experiments.

The FacePix dataset [292] contains images of subjects under natural conditions with head poses ranging from -90^o to 90^o. Automatically determining the head pose of a subject is important in applications ranging from robotics to assistive technology. Since head pose is a continuous variable (and can assume any value in the range $[-90, 90]$), this

is a regression problem. Sample pose images from this dataset are shown in Figure 6.6. The Laplacian of Gaussian (LoG) feature was used in this experiment yielding vectors of dimension 2560 and PCA was used to reduce the dimension to 100.

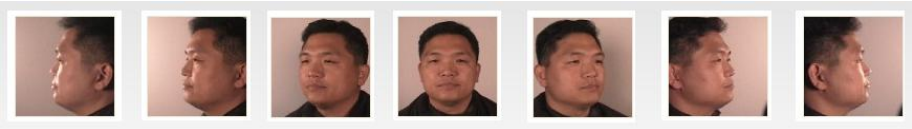
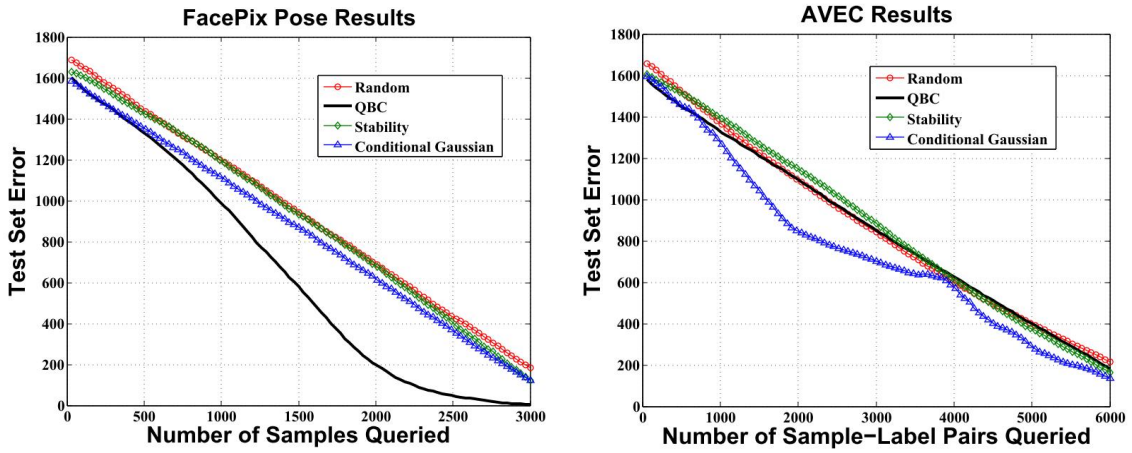


Figure 6.6: Pose images from the FacePix dataset

The Audio Visual Emotion Challenge (AVEC) [293] is a continuous human emotion recognition dataset that contains facial video data annotated with continuous emotion responses. The dataset contains Linear Binary Pattern (LBP) features [294] extracted for each video frame. Each video frame was divided into 10×10 image blocks and LBPs were calculated at each pixel in each block. A histogram of 59 LBPs was generated for each image block and all these 100 histograms were concatenated to a single feature vector of dimension 5900. Each video frame was annotated with four values, each corresponding to arousal, power, expectancy and valence. The values for arousal, power and valence lie between $[-0.6608, 0.6083]$, $[-0.4559, 0.7720]$ and $[-0.6332, 0.7774]$ respectively. But since expectancy values had a very different range of values between $[13.2320, 84.0560]$, they were normalized to lie between $[-0.7, 0.7]$. Each video was down-sampled by 7% to obtain a total of 2874 image samples. PCA was used to reduce the dimension from 5900 to 100 retaining about 98% of the variance. This is a regression problem with four labels, where the labels share an inherent correlation among them. Thus, the active matrix completion algorithms provide a natural way to exploit this correlation and query the informative sample-label pairs for manual annotation.

60% of the labels (for the FacePix Pose dataset) and sample-label pairs (for the AVEC dataset) were deleted at random and active learning was used to query the informative labels / sample-label pairs. The batch size was taken as 30 for FacePix and 60

for the AVEC dataset and the process was repeated over 100 iterations. The results (averaged over 3 random runs) are shown in Figure 6.7 where the x axis denotes the number of samples queried (for the FacePix dataset) / the number of sample-label pairs queried (for the AVEC dataset) and the y axis denotes the mean squared error on the test set. It is evident that all the active matrix completion algorithms result in a rapid decrease in the test error with increasing number of queries. We also note that random sampling depicts good performance on both these datasets.



(a) FacePix Pose (b) AVEC

Figure 6.7: Active Learning in Regression using Active Matrix Completion (Best viewed in color).

Active Feature Acquisition

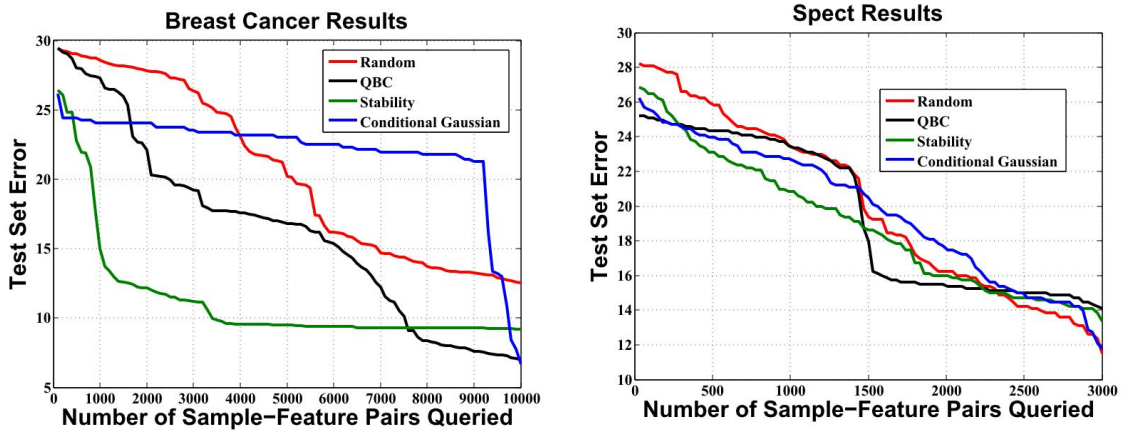
The incomplete-data problem, in which certain features are missing for particular data points, exists in a wide range of fields, including social sciences, computer vision, biological systems and remote sensing. In multi-sensor remote-sensing applications, incomplete data can result when only a subset of physical sensors (e.g. radar, infrared, acoustic) are deployed at certain regions. In a medical diagnosis application, the data for a particular patient may have missing features corresponding to the medical tests that could not be performed due to the associated time / costs. In such applications, it is possible to acquire

the missing data at a cost. In a medical diagnosis task, deciding which medical tests to administer would be equivalent to deciding which missing data values to acquire. Acquiring data is usually a time consuming, laborious task, which necessitates an active feature acquisition process. In contrast to conventional active learning, which selects the most beneficial samples to labels, this process would select the most informative features to acquire.

In this section, we demonstrate the usefulness of the proposed active matrix completion frameworks for active feature acquisition. The Breast Cancer and the Spect datasets were once again used for this experiment. Both these datasets represent natural settings for developing predictive models to classify a patient as normal or abnormal, where the acquisition of features has an associated cost. In our experiments, 60% of the feature values were deleted at random from the two datasets. The batch size was taken as 100 for Breast Cancer and 30 for the Spect dataset and the process was repeated over 100 iterations. The goal was to select the informative features so as to better diagnose patient as benign or malignant. We therefore study the decrease in generalization error with increasing number of features acquired. The averaged results (over 3 random trials) are depicted in Figure 6.8 and once again corroborate the efficacy of active sample selection over passive (random) selection. The QBC and Stability based methods particularly depict good performance.

6.9 Discussions

In this chapter, we presented novel algorithms to leverage the intelligence of human oracles to actively complete a partially observed data matrix. We presented two ensemble-based methods - Query-by-Committee and a Stability-based method and a strategy based on Conditional Gaussian distributions to compute the uncertainty of prediction of every missing entry in the matrix. The top k entries were then selected for manual annotation, where k is the desired batch size. Our results corroborated the advantage of active matrix



(a) Breast Cancer (b) Spect

Figure 6.8: Active Feature Acquisition using Active Matrix Completion (Best viewed in color).

completion over passive completion and also the efficacy of the algorithms in appropriately identifying a set of missing entries over random sampling in reducing the reconstruction error. To the best of our knowledge, this is the first research effort to intelligently exploit human supervision in the matrix completion problem, which is frequently encountered in data mining, machine learning and computer vision applications.

We also demonstrated how the proposed active matrix completion algorithms can be extended to solve several variants of the active learning problem, like transductive active learning, multi-label active learning, active learning in regression and active feature acquisition, to train a classification / regression model with minimal human effort. Our empirical analysis depicted tremendous promise in using the proposed methodologies to solve such problems. This corroborates the versatility of the framework in solving a variety of problems in collaborative filtering and active learning. We hope that this work will serve as a preliminary step in the development of matrix completion algorithms with "human-in-the-loop" and their adaptations in other interactive problem settings.

Chapter 7

GENERALIZATIONS AND EXTENSIONS

The fundamental rationale of the proposed batch mode active learning framework, as explained in Chapter 3, is based on selecting a set of m unlabeled points to ensure that the modified learner has low uncertainty on the unselected unlabeled points and also to select points that are diverse from the current set of labeled instances. The two conditions were combined into the following objective function, which drove the batch selection process:

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \tag{7.1}$$

s.t.:

$$M \in \{0,1\} \quad \text{and} \quad M^T \mathbf{1} = m$$

The indicator vector M was used to guide the point selection. The relaxed version of the above optimization problem was solved using the Quasi Newton method. In this chapter, we present a few variants of the batch selection criterion to corroborate the generalizibility of the framework. We also depict how the framework can be used for active batch selection when multiple sources of information are available. The flexibility of the approach is further certified by extending it to incorporate contextual information, which is frequently available in real world machine learning applications.

7.1 Varying the Criteria for Batch Selection

The active batch selection frameworks were mostly validated on the face based biometric recognition application. Considering the specific challenges of face-based biometrics, we would like to ensure that our learner, in addition to learning from the uncertain samples (as selected by the entropy term) also learns from the informative visages made briefly by

the subjects (e.g. a sudden smile or a sudden eyebrow raise). These images lie away from the main body of points, possibly in sparsely populated regions. To address this issue, it may be useful to design an objective function which selects samples from the sparsely populated regions of the unlabeled pool together with the samples that are uncertain for the current learner. Also, in certain situations, we may desire to place more weightage on the entropy term for batch selection so as to emphasize more on the uncertain unlabeled samples (as the diversity and density based terms involve distance based computations, which can be deceptive in high dimensional spaces). We present two objective functions to handle such situations.

Sparsity based Objective Function

We compute the sparsity of an unlabeled sample as its average Euclidean distance from the other unlabeled samples. Thus, samples that are located far away from the main body of points will have a high sparsity measure and vice versa. The purpose of this objective function is to ensure that the distance of each selected unlabeled point from the other unlabeled samples is maximum and the entropy of the updated classifier on the remaining points in the unlabeled pool is minimum. These conditions can be satisfied by the following definition:

$$\max_{M} \sum_{j \in U_t} D_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) S(y|x_j, w^{t+1}) \tag{7.2}$$

s.t.:

$$M \in \{0,1\} \text{ and } M^T \mathbf{1} = m \tag{7.3}$$

where D_j denotes the average Euclidean distance of the unlabeled point x_j from the other unlabeled samples and m is the user specified batch size (we focus on the static scenario

here; the extension to the dynamic scenario is trivial). M can be solved using the Quasi Newton framework to guide the point selection process.

Perplexity based Objective Function

Perplexity is a measure in information theory used to quantify how perplexed or confused a classifier is in predicting a test point. It is defined as 2 raised to the power of entropy [295]:

$$PPL = 2^S = 2^{-\sum_{y \in C} P(y) \log P(y)}$$

On similar lines as the diversity based selection criterion, we can define an objective function which selects points that are maximally diverse compared to the current training set and minimizes the perplexity of the points remaining in the unlabeled pool after batch selection. The perplexity term magnifies the value of entropy and dominates the point selection process.

$$\max_{M} \sum_{j \in U_t} \rho_j M_j - \lambda \sum_{j \in U_t} (1 - M_j) PPL(y|x_j, w^{t+1}) \tag{7.4}$$

s.t.:

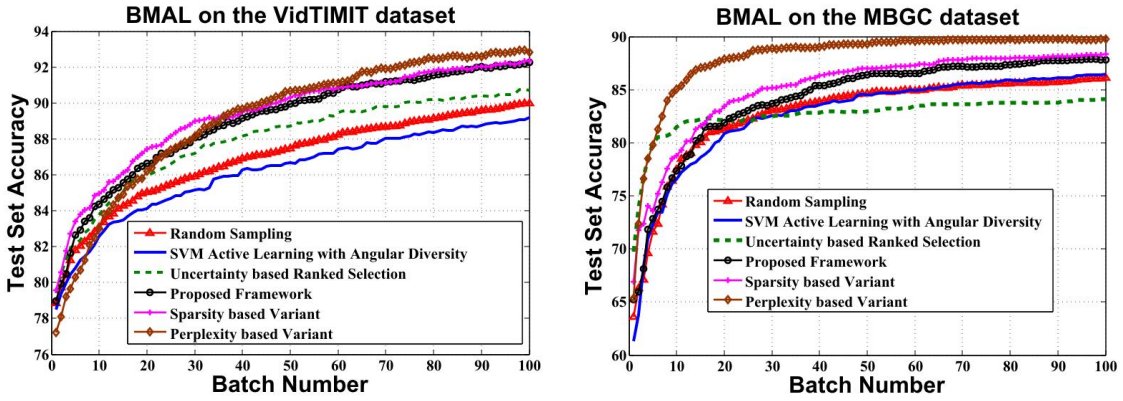
$$M \in \{0,1\} \text{ and } M^T \mathbf{1} = m \tag{7.5}$$

The selection vector M is solved using the Quasi Newton method, after relaxing constraint 7.5.

Comparison against the heuristic approaches

To depict their generalizibility, a training set was induced with 10 images of each of 25 subjects from the VidTIMIT and MBGC datasets. Unlabeled video streams, each containing 250 images were presented to the learner and a batch of 10 was queried from each. The growth in accuracy was studied on a test set containing 4500 images spanning all the

25 subjects. The training, test and the unlabeled sets were chosen at random - no particular proportion of subjects was maintained in the training and unlabeled sets. The results using the diversity, sparsity and perplexity based objective functions, together with the heuristic techniques (random sampling, SVM angular diversity based selection and uncertainty based sampling) are shown in Figure 7.1. It is noted that all the objective functions perform better than the heuristic BMAL techniques, corroborating their usefulness. This shows that the proposed framework is generalizible and by suitably choosing a batch selection criterion, it can substantially reduce human annotation effort as compared to the heuristic BMAL algorithms. The perplexity based variant depicts the best performance in both the datasets emphasizing the usefulness of the entropy based criterion in batch selection.



(a) Batch Mode Active Learning on the VidTIMIT dataset

(b) Batch Mode Active Learning on the MBGC dataset

Figure 7.1: Performances of different Batch Mode Active Learning schemes on the VidTIMIT and MBGC datasets (Best viewed in color).

7.2 Learning from Multiple Sources of Information

Most biometric systems used in real world applications are unimodal [296], that is they rely on a single modality to carry out the authentication / recognition task. Such systems suffer from a variety of problems:

- the data collected may be corrupted by noise
- a user may interact incorrectly with a sensor, for example, may provide an incorrect facial pose
- it is possible that a particular trait of two different persons are very similar
- a single trait may be subject to spoof attacks

Multimodal systems seek to alleviate this problem by consolidating evidence from multiple sensors. This can lead to better and reliable performance of the recognition / validation system. The individual pieces of information, being independent, are fairly robust to noise. Multimodal systems can be classified into 5 categories as shown in Figure 7.2 [297].

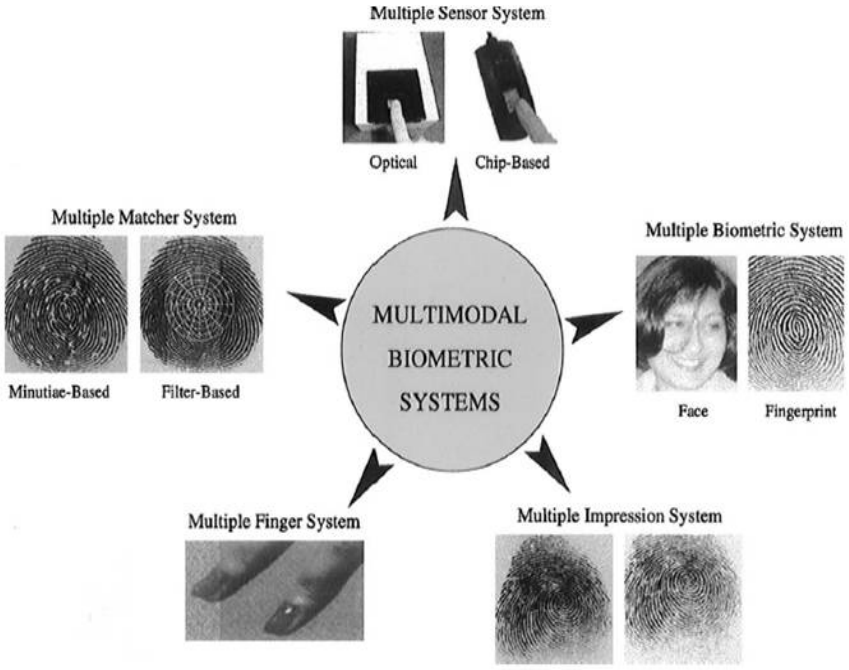


Figure 7.2: Categorization of approaches towards multimodal biometrics

When two modalities are used for person recognition, proper fusion schemes are required to combine the information provided by multiple modalities to perform effective recognition. In a multimodal biometric system, the information can be fused at different

levels. The various categories of fusion levels are summarized in Figure 8.4 (as illustrated in [298]). Dasarathy [299] categorized these approaches as data-level fusion (where data is combined), feature-level fusion (where features are extracted from the data in different modalities separately, which are then combined) and decision-level fusion (where the information is fused at the decision-making level). Over the last two decades, several approaches of multimodal fusion in biometrics have been explored [300] [301] [302] and it has been established that learning from mutiple sources can be superior to learning from a single source, if the sources are used appropriately [237]. In this section, we establish how the proposed batch mode active learning framework can be extended to integrate multiple sources of information.

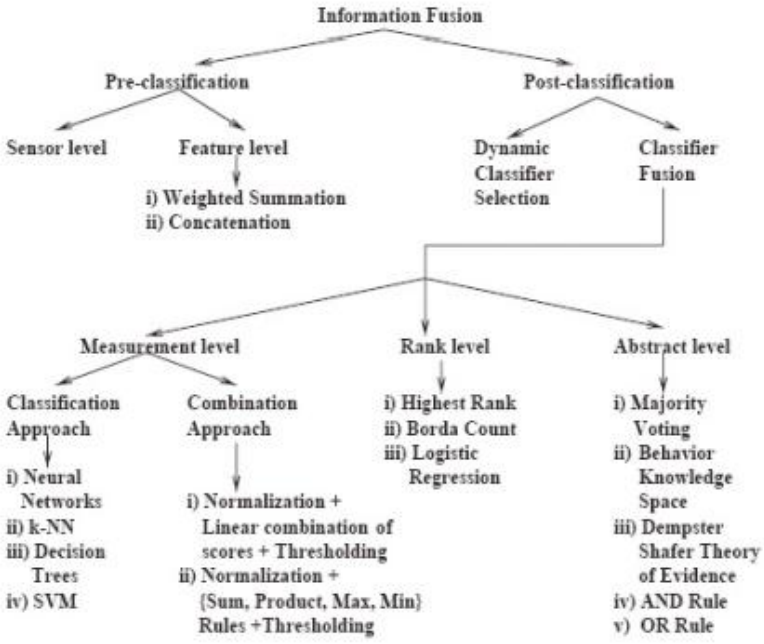


Figure 7.3: An overview of the approaches to information fusion

Batch Mode Active Learning from Multiple Sources of Information

Consider the case of learning from two sources of information, where we are given two training sets L_{t1} and L_{t2} and two unlabeled sets U_{t1} and U_{t2} corresponding to the two

sources. As before, we would like to sample points from each stream, not excluding the sparsely populated regions. Also, for each data point that is not selected in the batch, a condition can be imposed ensuring that the predictions from the two models obtained from each of the sources, agree to a large extent. A metric measuring the difference between two probability distributions can be used for this purpose. The symmetric Kullback Leibler divergence was used as a measure of difference in this work. The objective function guiding the point selection can therefore be modified as:

$$\max_{M} \sum_{j \in U_{t1}} \rho_j M_j + \sum_{j \in U_{t2}} \rho_j M_j - \lambda \sum_{j} (1 - M_j) KLD(y|x_j^{U_{t1}} w^{t+1}, y|x_j^{U_{t2}} w^{t+1}) \tag{7.6}$$

s.t.:

$$M \in \{0,1\} \quad \text{and} \quad M^T \mathbf{1} = m \tag{7.7}$$

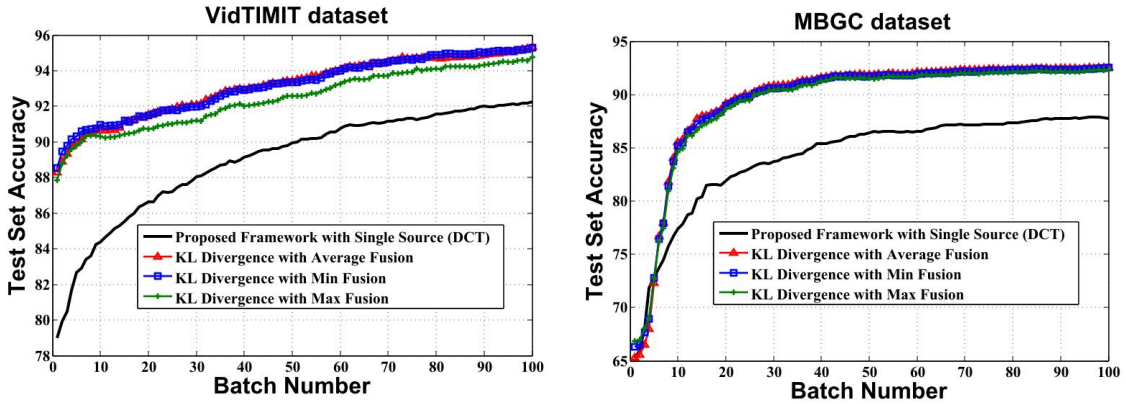
where

$$KLD(P,Q) = \sum_{i=1}^{n} (p_i - q_i) \log \frac{p_i}{q_i} \tag{7.8}$$

is the symmetric Kullback Leibler divergence between probability distributions P and Q [255]. The third term in Equation (7.6) denotes the Kullback Leibler divergence between the predicted probabilities of each unselected point from the two unlabeled sources U_{t1} and U_{t2}, which is to be minimized. The vector M, which governs the point selection, however, will remain the same across the sources of information. Together with constraints from Equation (7.7), the problem can be solved using the Quasi Newton method in a similar manner as before.

It is interesting to note that in case of a single source of information, Equation (7.6) reduces to Equation (7.1) in the original formulation. The second term in Equation 7.6 vanishes, and from Equation (7.8), we note that when Q is non-existent, the KLD term becomes $\sum_{i=1}^{n} p_i \log p_i$, which is the negative entropy of distribution P.

To demonstrate this idea, we represent multiple sources using multiple features from each face image. In addition to DCT, the Scale Invariant Feature Transform (SIFT) feature [303] was used as the second source of information. Two classifiers were trained separately on the two features. A batch of unlabeled points was selected according to Equation (7.6) and the corresponding classifiers were updated with the selected batch. They were then applied on separate test sets for each feature to yield two sets of class probabilities for each test point. Three simple fusion rules - average, minimum and maximum - were then applied to combine the two sets of probability values. The results on the VidTIMIT and the MBGC datasets are shown in Figure 7.4, where the results are compared with the proposed framework using only the DCT information.



(a) VidTIMIT dataset (b) MBGC dataset

Figure 7.4: Batch Mode Active Learning from multiple sources on the VidTIMIT and MBGC datasets.

For each fusion strategy, the performance tremendously improves when the two sources are used together for learning. This depicts the merit of the framework for effective batch selection in the presence of multiple sources of information. We also note that the three fusion strategies depict almost similar performance on both the datasets. Although validated using multiple image features in this work, the framework can be applied in conjunction with any two (or more) data sources. In the Social Interaction Assistant, for example, the same batch selection strategy can be used with the video and audio modali-

ties for robust person recognition. This will be taken up as part of our future work. We also intend to explore other fusion strategies like the Dempster-Shafer theory and the Dezert-Smarandache Theory.

7.3 Context Aware Learning

Context awareness has gained popularity as a mechanism to improve the performance of an application in ubiquitous computing environments. In a challenging problem like biometric / object recognition, contextual information can provide useful cues to increase the robustness of the recognition engine. Context has been defined in the machine learning literature in different ways and its precise definition is open to discussion. However, proper interpretation of context in a given application can improve the performance of the system and can also add practicality to the underlying computational framework. We briefly review a few context aware learning techniques that have been proposed in the literature.

Torralba [304] introduced a context aware fremwork for object detection which modeled the relationship between context and object properties based on the correlation between the statistics of low level features across the entire scene and the objects that it contained. The resulting scheme served as an effective procedure for object priming, context driven focus of attention and automatic scale selection on real world scenes. Strat and Fischler [305] used contextual information to improve the performance of object detection in complex outdoor scenes. Olson and Chun [306] argued that invariant spatial relationships of objects may provide a rich source of contextual information. The authors investigated whether both local context that surround a target and long range context that does not spatially coincide with a target can influence target localization. They concluded that implicit learning of spatial context was robust across noise and biased towards spatially grouped information. Song and Leung [307] developed an algorithm which exploited contextual information like the color and texture of clothes for robust person recognition.

Using context awareness in active learning is almost unexplored. Very recently, Kapoor *et al.* [155] incorporated *match* and *non-match* constraints in active learning for face recognition. However, this work was aimed at selecting one face image at a time (pool-based setting). In this section, we extend the proposed batch mode active learning framework to incorporate contextual information.

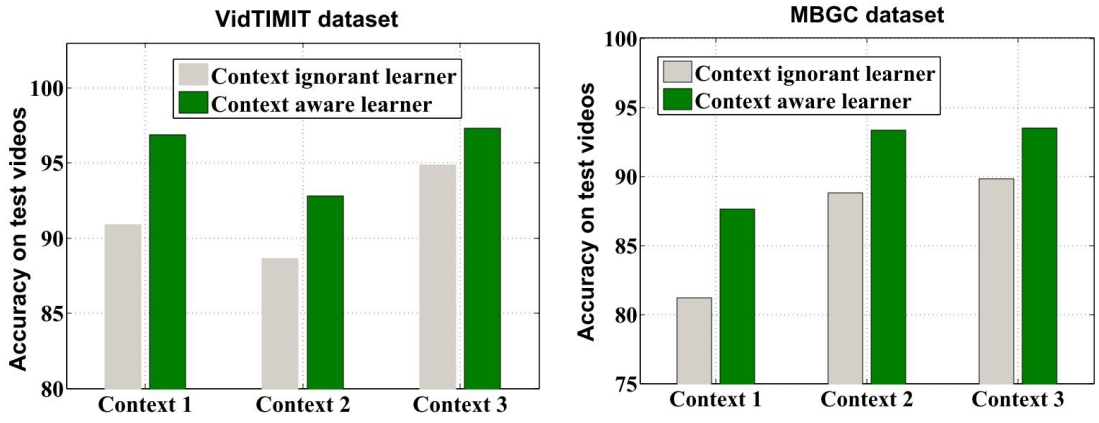
Batch Mode Active Learning Framework to Incorporate Contextual Information

In this work, context was defined as the *location of a user*, for the sake of simplicity without any loss in generality (similar to Dey *et al.* [308]). It was assumed that at any given location, the user is cognizant of the subjects to be expected in that location (for example, work acquaintances in an office setting or family members in a home setting). This was used to construct a prior probability vector depicting the chances of seeing each subject at a given location. In such a situation, a logical strategy for querying instances would be to guarantee that the images remaining in the unlabeled video after batch selection have low entropy with respect to the subjects expected in the given context. Thus, the performance score function can be modified to ensure that the entropy is computed only on the subjects that are present in a given video stream:

$$f(B) = \sum_{i \in B} \rho_i - \lambda \sum_{j \in U_t - B} S^{context}(y|x_j, w^{t+1})$$

Here, $S^{context}$ is the context aware entropy term. For each unlabeled image, this term was computed from the posterior probabilities, which in turn were obtained by multiplying the likelihoods returned by the trained GMM classifier with the context aware prior. Thus, subjects not expected in a given context will have low priors and consequently, the corresponding posteriors will not contribute much in computing $S^{context}$. To simulate this situation, three contexts were arbitrarily defined and 8 random subjects (chosen from the set of 25) were assigned to each context. BMAL was used to select batches of samples

from unlabeled video streams in each context. The updated classifiers were then tested on videos in the respective context. The context-ignorant learner was implemented using equal class priors in the entropy term. Note that this process generates different models for each of the contexts after applying the BMAL methodology.



(a) VidTIMIT dataset (b) MBGC dataset

Figure 7.5: Context Aware Learning on the VidTIMIT and MBGC datasets.

Figure 7.5 shows the accuracies obtained on the VidTIMIT and MBGC test videos (averaged over three trials in each context). It is noted that in each context, the context aware learner produces better accuracy on test videos than the context ignorant learner. Thus, incorporation of context in the formulation further helps in querying salient images. However, as mentioned earlier, the limitation of this approach is that a classifier trained in a given context will perform well only in that context. Nonetheless, many real world problems have, or can be reduced to, only a limited set of contexts. It is therefore possible to maintain an ensemble of context specific learners ensuring that a classifier trained in a given context has the highest weightage when tested in the same context. This makes the proposed approach feasible as well as meritorious.

7.4 Discussions

In this chapter, we demonstrated the generalizibility of the proposed batch mode active learning framework. We analyzed the performance using different variants of the objective function for batch selection. We also corroborated the flexibility of our framework by extending it to incorporate information from multiple sources and also contextual information, which are frequently available in biometrics based recognition systems. The results obtained speak of the potential of this method in being used for real world biometric recognition applications, like the Social Interaction Assistant.

From the results, it is evident that the density, diversity and perplexity based objective functions outweigh the heuristic BMAL techniques. However, given an unlabeled video stream, it becomes important to decide which one of them should be used to select images. This may be a difficult choice if no prior knowledge is available about the video stream in question. A possible solution may be to maintain an ensemble of objective functions and dynamically select one based on some criteria. Moreover, the Quasi Newton solution strategy requires the objective function to be differentiable. Thus, non-differentiable functions (for example, functions with a max/min sub-function) cannot be handled using this method. One may want to design an objective function having a term ensuring that for each image in the unlabeled video stream, the highest class probability predicted by the current classifier has to exceed a given threshold. A solution strategy involving Quasi Newton will not allow such a formulation. However, efforts have been made to optimize non-differentiable objective functions as described in [309]; such an approach will be explored in our future work to solve optimization problems with max / min functions.

Chapter 8

RELATED CONTRIBUTIONS

Active learning techniques primarily rely on the definition of a suitable *query function*, a function that queries each unlabeled point to decide on its appropriateness and relevance in training a classification model. Query functions in existing active learning techniques often select examples that have the most uncertainty [123], least confidence [103] or maximum disagreement among a committee of classifiers [63]. Most of the existing approaches have been based on inductive inference, where a general classifier function is learnt from existing training examples to predict the class labels of new examples. However, in recent years, there has been a growing interest on transductive inference [282], where the training examples are directly used to develop a reasoning to predict the labels of new examples. In this chapter, we propose a Generalized Query by Transduction approach for active learning in the online (stream-based) setting using p-values obtained from the Conformal Predictions (CP) framework. The main contributions of this work are two-fold. Firstly, we introduce the Generalized Query by Transduction (GQBT) approach for active learning using the theory of conformal predictions that can be used with any pattern classification algorithm in an online setting. Secondly, while most existing active learning approaches evaluate a single critereon (such as confidence, uncertainty or disagreement), there have been more recent efforts to combine multiple criteria (such as representativeness, informativeness and diversity by Shen *et al.* [223]) to select appropriate examples. We show how the proposed active learning approach can be used to combine multiple criteria for active learning. We demonstrate the improved performance of the proposed approach with commonly used datasets from the UCI Machine Learning repository, and apply the approach to face recognition to validate its applicability and performance in a

challenging real-world problem. We first present a background of the CP framework and then detail our online active learning algorithm.

8.1 Theory of Conformal Predictions

The theory of conformal predictions was recently developed by Vovk, Shafer and Gammerman [287, 310] based on the principles of algorithmic randomness, transductive inference and hypothesis testing. This theory is based on the relationship derived between transductive inference and the Kolmogorov complexity [311] of an i.i.d. (identically independently distributed) sequence of data instances. Consider the set of labeled data instances to be represented as the sequence $Z = ((x_1, y_1), \ldots, (x_n, y_n))$, where x_i is a data instance, and y_i is the corresponding class label. If $l(Z)$ is the length of this sequence, and $C(Z)$ is its Kolmogorov complexity (the length of the minimal description of Z using a universal description language), then:

$$\delta(Z) = l(Z) - C(Z) \tag{8.1}$$

where $\delta(Z)$ is called the *randomness deficiency* of the sequence Z. Intuitively, Equation (8.1) states that lower the value of $\delta(Z)$, the higher is the randomness of the sequence. As a corollary, if there was a new data instance x_{n+1}, and we were to predict its label based on the available labeled data Z, the confidence in the prediction would be low, if the sequence Z was highly random i.e. $\delta(Z)$ was low.

Evidently, the challenge is the computation of the randomness deficiency, $\delta(Z)$, of a given sequence Z. This is achieved using the Martin-Lof test for randomness, which can be summarized as a function $t : Z^* \to \mathbb{N}$ (the set of natural numbers with 0 and ∞), such that $\forall n \in \mathbb{N}, m \in \mathbb{N}, P \in P_n$:

$$P\{z \in Z^n : t(z) \geq m\} \leq 2^{-m} \tag{8.2}$$

where P_n is the set of computable probability distributions. Equation (8.2) can also be written as:

$$P\{z \in Z^n : t(z) \in [m, \infty)\} \leq 2^{-m} \tag{8.3}$$

Now, if we use the transformation $f(x) = 2^{-x}$, Equation (8.3) can in turn be written in terms of a new function $t'(z)$:

$$P\left\{z \in Z^n : t'(z) \in \left(0, 2^{-m}\right]\right\} \leq 2^{-m} \tag{8.4}$$

Hence, a function $t' : Z^* \to (0, 2^{-m}]$ is a Martin-Lof test for randomness if $\forall m, n \in \mathbb{N}$, the following holds true:

$$P\left\{z \in Z^n : t'(z) \leq 2^{-m}\right\} \leq 2^{-m} \tag{8.5}$$

If 2^{-m} is substituted for a constant, say r, and r is restricted to the interval $[0, 1]$, Equation (8.5) is equiavalent to the definition of a p-value typically used in statistics for hypothesis testing. Given a null hypothesis H_0 and a test statistic, p-value is simply defined as the probability of obtaining a result at least as extreme as the one that was actually observed, assuming that the null hypothesis is true. In other words, the p-value is the smallest significance level of the test for which H_0 is rejected based on the observed data, i.e. the p-value provides a measure of the extent to which the observed data supports or disproves the null hypothesis.

To translate this theory to pattern classification problems, Vovk *et al.* [310] defined a *non-conformity measure* that quantifies the conformity of a data point to a particular class label. This non-conformity measure can be appropriately designed for any classifier under consideration, thereby allowing the concept to be generalized to different kinds of pattern classification problems. To illustrate this idea, the non-conformity measure of a data point x_i for a k-Nearest Neighbor classifier is defined as:

$$\alpha_i^y = \frac{\sum_{j=1}^{k} D_{ij}^y}{\sum_{j=1}^{k} D_{ij}^{-y}} \tag{8.6}$$

where D_i^y denotes the list of sorted distances between a particular data point x_i and other data points with the same class label, say y. D_i^{-y} denotes the list of sorted distances between x_i and data points with any class label other than y. D_{ij}^y is the jth shortest distance in the list of sorted distances, D_i^y. In short, α_i^y measures the distance of the k nearest neighbors belonging to the class label y, against the k nearest neighbors from data points with other class labels (Figure 8.1). Note that the higher the value of α_i^y, the more non-conformal the data point is with respect to the current class label i.e. the probability of it belonging to other classes is high.

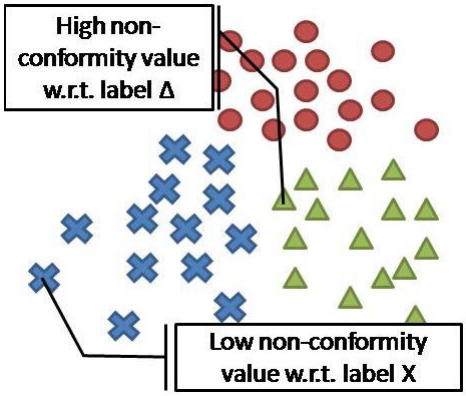


Figure 8.1: An illustration of the non-conformity measure defined for k-NN

Given a new test data point, say x_{n+1}, a null hypothesis is assumed that x_{n+1} belongs to the class label, say, y_p. The non-conformity measures of all the data points in the system so far are re-computed assuming the null hypothesis is true. A p-value function (which satisfies the Martin-Lof test definition in Equation (8.5) is defined as:

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{m} \tag{8.7}$$

where $\alpha_{n+1}^{y_p}$ is the non-conformity measure of x_{n+1}, assuming it is assigned the class label y_p, and m is the total number of data instances. In simple terms, Equation (8.7) states that

the p-value of a data instance belonging to a particular label is the normalized count of the data instances that have a higher non-conformity score than the current data instance, x_{n+1}. It is evident that the p-value is highest when all non-conformity measures of training data belonging to class y_p are higher than that of the new test point, x_{n+1}, which points out that x_{n+1} is *most conformal* to the class y_p. This process is repeated with the null hypothesis supporting each of the class labels, and the highest of the p-values is used to decide the actual class label assigned to x_{n+1}, thus providing a transductive inferential procedure for classification. The largest p-value is called the *credibility* and 1 minus the second largest p-value is refered to as the *confidence*. The general schema of Conformal Predictors in the classification setting is depicted in Algorithm 9.

Algorithm 9 Conformal Predictors for Classification

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, $x_i \in X$, number of classes M, $y_i \in Y = y_1, y_2, \ldots, y_M$, classifier Ξ

1: Get new unlabeled example x_{n+1}.
2: **for** all class labels, y_i, where $i = 1, \ldots, M$ **do**
3: Assign label y_i to x_{n+1}.
4: Update the classifier Ξ, with $T \cup \{x_{n+1}, y_i\}$.
5: Compute non-conformity measure value, $\alpha_{n+1}^{y_i}$ to compute the p-value, P_i, w.r.t. class y_i (Equation 8.7) using the conformal predictions framework.
6: **end for**
7: Output the conformal prediction regions $\Gamma_{1-\varepsilon} = \{y_i : P^{y_i} > \varepsilon, y_i \in Y\}$, where $1 - \varepsilon$ is the confidence level.

One of the key features of this framework is the calibration of the obtained confidence values in an online setting. Probabilities generated by inductive inference approaches in an online setting are often not meaningful since the model needs to be continuously updated with every new example. However, the theory behind the conformal prediction framework guarantees that the probability (or confidence) values obtained using this transductive inference framework manifest as the actual error frequencies in the online setting i.e. they are well-calibrated [288]. This is depicted in Figure 8.2, which plots the cumulative number of errors with increasing number of test samples at various

confidence levels. We note that, at every level of confidence, the number of errors committed by the system is upper bounded by the threshold (which can be set by the user).

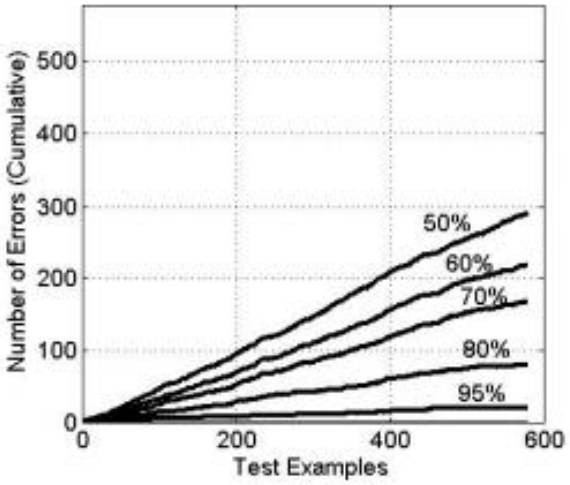


Figure 8.2: Performance of the CP Framework on the Cardiac Patient Dataset. Note that the errors are calibrated at each of the confidence levels. For instance, at 80% confidence level, the number of errors will always be less than 20% of the total number of test examples.

We now present the proposed Generalized Query by Transduction approach for online active learning, which is derived from the conformal predictions framework that was described above.

8.2 Generalized Query by Transduction (GQBT)

The p-values for each of the class labels obtained using the principles of transductive inference, as outlined in the theory of conformal predictions, are used to design the query function in the proposed approach. Ho and Wechsler proposed a similar approach in [104], where the query function was limited to using the top two p-values (amongst the list of p-values obtained for all the class labels). They formally defined the *closeness* between the top two p-values, $I(x_{n+1}) = p_j - p_k$, as the measure of the quality of information in an unlabeled example in the active learning process. The example is queried if $I(x_{n+1}) < \delta$,

for an empirically determined threshold δ. In the proposed approach, we generalize the query function to use all (or as many as required) p-values that are obtained using the conformal predictions framework. We also call this approach *generalized* since it can be integrated into any existing classification algorithm. In addition, we show how this framework can integrate multiple criteria in the proposed query function. We illustrate the proposed approach using suitable examples, and compare the performance of our approach with Ho and Wechsler's QBT [104], along with random sampling, Query by Committee, and a Support Vector Machine (SVM) margin-based active learner.

In the proposed GQBT approach, we define a matrix C which contains the absolute value of the pairwise differences between all the p-values obtained from the conformal predictions framework:

$$C_{ij}(P) = \left|P_i - P_j\right| \tag{8.8}$$

where $i, j = 1, \ldots, M$ and M is the number of classes. Since this C matrix has diagonal elements as zero and is symmetric, its eigendecomposition provides a naturally useful measure with interesting properties. The largest eigen-value of C, say $\eta(C)$, assumes values that are directly proportional to the average pairwise differences between the p-values. Further, it is possible to prove that for any given set of p-values, the matrix C will always have exactly one positive eigenvalue, which we used as a measure of disagreement in this work (please refer Appendix A for a mathematical proof of this proposition). When all the p-values are equal, $\eta(C)$ is trivially zero. As the pairwise differences between the p-values increase, $\eta(C)$ increases proportionately. We now show why $\eta(C)$ provides a natural measure of the extent of disagreement between the p-values, which we intend to use in the proposed approach.

The eigendecomposition of C is given by the characteristic equation:

$$|C - \lambda I| = 0 \tag{8.9}$$

where $|.|$ is the matrix determinant. When the pairwise differences are multiplied by a constant factor, say d, the new C, say C^*, is equal to dC. The characteristic equation for C^* is given by:

$$|C^* - \lambda^* I| = 0 \tag{8.10}$$

where λ^* are the eigenvalues of C^*. Substituting $C^* = dC$,

$$|dC - \lambda^* I| = 0 \Rightarrow \left| d(C - \frac{\lambda^*}{d} I) \right| = 0 \tag{8.11}$$

$$\Rightarrow |dI| \left| C - \frac{\lambda^*}{d} I \right| = 0 \tag{8.12}$$

Since $|dI| \neq 0$,

$$\Rightarrow \left| C - \frac{\lambda^*}{d} I \right| = 0 \tag{8.13}$$

Comparing Equations (8.13) and Equation (8.9),

$$\lambda = \frac{\lambda^*}{d} \tag{8.14}$$

that is, the eigenvalues λ^* are also multiplied by the same constant factor d. For another C matrix, say $\hat{C}$, whose average pairwise difference lies between the original average pairwise difference in C and that in C^*, the corresponding eigenvalues $\hat{\lambda}$ will lie between λ and λ^*. We exploit this ordering of eigenvalues as a natural measure of the extent of disagreement among the p-values obtained.

Since p-values assume values in the interval $[0,1]$, the largest eigenvalue, $\eta(C)$, tends to have low numeric values. For convenience of implementation, we compute the inverse of C, and use the largest eigenvalue of C^{-1} in our work. Since $\eta(C^{-1})$ is *inversely* proportional to the average difference between the p-values, we accordingly factor this in the design of our query condition. The proposed GQBT approach is presented in Algorithm 10.

Algorithm 10 Generalized Query by Transduction for Online Active Learning

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, classifier Ξ, selection threshold δ, stopping threshold γ, number of classes M, number of queried points p, budget constraint β (maximum number of points that can be queried)

1: **initialize** $p \leftarrow 0$
2: **repeat**
3: Get new unlabeled example x_{n+1}.
4: **for** all class labels, y_i, where $i = 1, \ldots, M$ **do**
5: Assign label y_i to x_{n+1}.
6: Update the classifier Ξ, with $T \cup \{x_{n+1}, y_i\}$.
7: Compute non-conformity measure value, $\alpha_{n+1}^{y_i}$ to compute the p-value, P_i, w.r.t. class y_i (Equation 8.7) using the conformal predictions framework.
8: **end for**
9: Construct the matrix C, such that $C_{ij}(P) = \left|P_i - P_j\right|$ (Equation 8.8).
10: Compute $\eta(C^{-1})$ as the largest eigenvalue of C^{-1}.
11: **if** $\eta(C^{-1}) > \delta$ **then**
12: Add x_{n+1} to training set i.e. $T \leftarrow T \cup \{x_{n+1}, y_c\}$, where y_c is the correct label for x_{n+1}.
13: $p \leftarrow p + 1$.
14: **end if**
15: **until** $\eta(C^{-1}) > \gamma$ or $p < \beta$

Almost all online active learning algorithms rely on empirically obtained thresholds to decide if an unlabeled example needs to be queried. In contrast, in this approach, the largest eigenvalue has a straightforward connotation that can be exploited. The selection threshold δ is initialized to the largest eigenvalue of the C^{-1} matrix that is constructed assuming the pairwise differences between the p-values are equal to a unit percentage (i.e. 0.01) each. Similar to what was proved in Equation (8.14), the eigenvalues for C^{-1} are divided by a factor of d, when C is multiplied by d. Hence, when the pairwise differences are equal to 0.02 each, the largest eigenvalue of the corresponding C^{-1} matrix is now equal to $\frac{\delta}{2}$. To apply this in the algorithm, if no examples are selected after, say r, examples are observed, the selection threshold is changed to: $\delta \leftarrow \frac{\delta}{2}$, thus allowing for a more accommodative threshold. Depending on the dataset under consideration, this can progressively be continued at periodic intervals to $\delta \leftarrow \frac{\delta}{3}, \delta \leftarrow \frac{\delta}{4}$, and so on, as may be required in a particular setting. This provides for an automatic methodology to set (and modify) threshold values, where the query condition becomes lenient with time.

We use SVM as the classifier in this work for a few reasons. Firstly, there have been several active learning techniques in the recent past that have used the margin distance in a SVM to query examples in active learning [111, 104], leading to the popularity of SVMs in active learning. Secondly, there have been recent efforts to develop incremental SVMs for an online setting [312] to train newer examples into an existing SVM model. One of the primary limitations of the proposed approach (or any transductive inference approach, for that matter) is the computational overhead in Steps 5-7 in Algorithm 10 for each class label. The use of incremental SVMs substantially offsets this limitation. Thirdly, the Lagrange multipliers obtained while training a SVM are a straightforward choice to consider as non-conformity scores, as pointed out by Vovk *et al.* [310]. The Lagrange multipliers, $\alpha_i, i = 1,\ldots,n$, are computed while maximixing the dual formulation in the soft margin SVM:

$$Q(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i,x_j) + \sum_{i=1}^{n}\alpha_i \tag{8.15}$$

subject to constraints $\sum_{i=1}^{n}\alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C, i = 1,\ldots,n$, and $K(.)$ is the kernel function. The Lagrange multipliers' values are zero for examples outside the margin, and lie between 0 and C for examples on and within the margin, thereby providing a natural monotonic measure of *non-conformity* w.r.t. the corresponding class.

Why Generalized QBT?

Before we present the experimental results, we show how the proposed GQBT is a generalization of the QBT approach proposed by Ho and Wechsler [104]. Ho and Wechsler define the quality of information of a new data example as $I(x_{n+1}) = p_j - p_k$, where p_j and p_k are the highest 2 p-values obtained using the conformal predictions framework. We define the quality of information using the largest eigenvalue of the matrix C containing the pairwise differences between all p-values. In a binary classification problem (or if only

the top 2 p-values are used in a multi-class setting), our approach becomes the same as Ho and Wechsler's. This is because C is now given by:

$$\begin{bmatrix} 0 & |p_1 - p_2| \\ |p_1 - p_2| & 0 \end{bmatrix}$$

whose largest eigenvalue is $|p_1 - p_2|$ itself, which is the measure used by Ho and Wechsler. However, the progressive choice of selection threshold values in our approach (as $\delta, \frac{\delta}{2}$, etc. detailed earlier) performs better than the empirical choice of thresholds in Ho and Wechsler's approach. This is illustrated in Figure 8.3, which shows how the proposed GQBT approach has a lower label complexity i.e. it achieves the highest accuracy by querying much fewer points than Ho and Wechsler's approach.

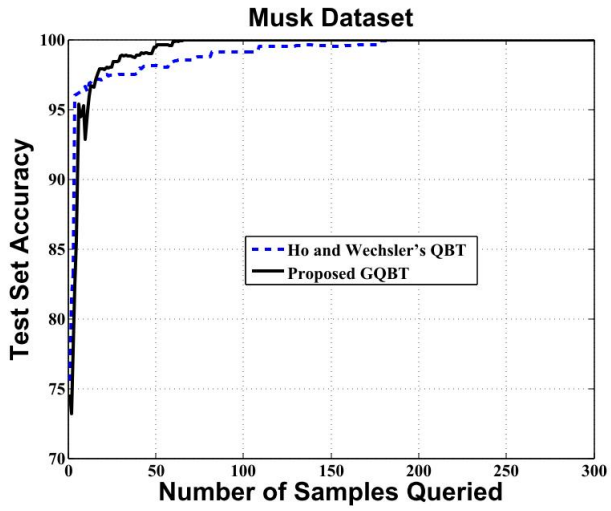


Figure 8.3: Comparison of the proposed GQBT approach with Ho and Wechsler's QBT approach on the Musk dataset from the UCI Machine Learning repository. Note that our approach reaches the peak accuracy by querying ≈ 80 examples, while the latter needs ≈ 160 examples.

Combining multiple criteria for active learning

It may often be essential to combine multiple criteria to decide if a particular unlabeled example needs to be queried for its true label, and included in the training set, and there have been recent efforts in this direction [223]. In our work, for example, in addition to the Lagrange multipliers (whose values are closely related to the distance of the example from the SVM margin), it may be useful to consider another non-conformity measure that estimates the density of examples in the neighborhood of a given unlabeled example. This can be defined using the k-NN classifier (a non-parametric density estimator), as stated earlier in Equation (8.6) in Section 8.1. Evidently, the theory of conformal predictions can also be used with this measure to obtain another set of p-values. We use results from statistical hypothesis testing to combine these p-values. Given that the p-value is a uniformly distributed random variable on the interval $[0,1]$, the combined significance level or p-value of n individual p-values can be given as [313]:

$$k \sum_{i=0}^{n-1} \frac{(-\ln k)^i}{i!} \tag{8.16}$$

where $k = (p_1 \times p_2 \times p_3 \ldots \times p_n)$, the product of the given set of p-values. While we use this approach in our work, there are other methods in hypothesis testing to combine p-values [314], which can be used too. Figure 8.4 shows the improvement in performance obtained (on the same dataset as in Figure 8.3) by combining the p-values obtained using the non-conformity measures computed from the SVM and the k-NN classifier.

8.3 Experiments and Results

We compared the performance of the proposed GQBT approach with three other online active learning algorithms together with random sampling. The methods are briefly outlined below:

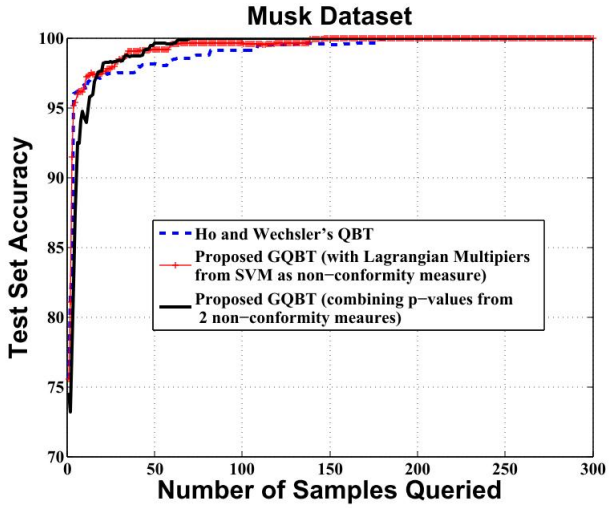


Figure 8.4: Performance comparison on the Musk dataset (as in Figure 8.3). Note the reduction in label complexity obtained by combining the p-values from the two non-conformity measures discussed in Section 8.2. The proposed approach needs only ≈ 50 examples to reach the peak accuracy.

Random Sampling: In this method, when a new example arrives, we randomly decide whether to query this point for its class label or not, i.e. each example is queried with a probability of 0.5.

Margin based SVM: An SVM classifier is constructed from the given set of training instances. For an unlabeled example x_{n+1}, its decision value $f(x) = w.\phi(x) + b$ is computed and if it is below a certain threshold, the point is queried. If a certain number of unlabeled points are not queried in succession, the threshold is updated as the average of the SVM decision values of the unqueried examples.

Query by Committee: A committee consisting of two classifiers, SVM and k-NN (with $k = 10$), was used. For a given unlabeled example, the SVM output values are converted into probabilities using Platt's method [315]. For k-NN, the class probability for the unlabeled example is defined as the fraction of the number of points of a given class occurring in its k nearest neighbors. Once we have the probability values from the two classifiers, we compute the Kullback Leibler divergence between these two sets. A

high divergence implies that the point is informative and should be queried. The threshold for the KL divergence value was updated as described for the margin based SVM.

Query by Transduction: This is the method proposed by Ho and Wechsler [104] as described previously.

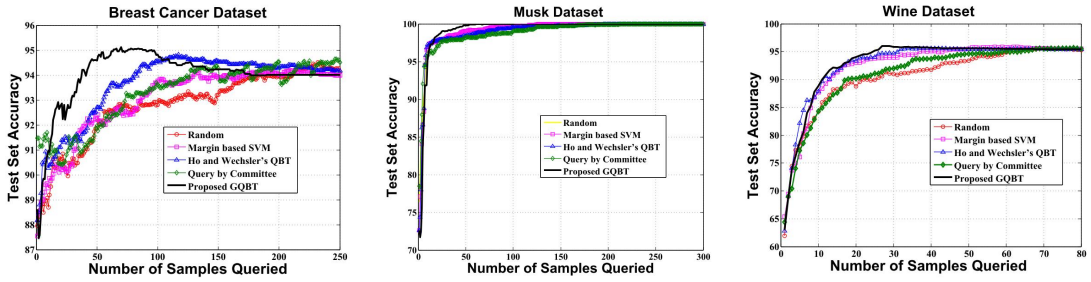
We selected five datasets (with different number of classes, dimensions and instances) from the UCI Machine Learning repository [257] to test the generalizibility of the proposed approach. The datasets and their details are listed in Table 8.1.

Dataset	Classes	Size of dataset	Dim	Initial training set	Size of unlabeled pool	Size of test set
Breast Cancer	2	569	30	10	259	300
Musk	2	1000	166	2	498	500
Wine	3	178	13	3	88	87
Waveform	3	5000	21	15	2485	2500
Image Segmentation	7	2310	19	35	175	2100

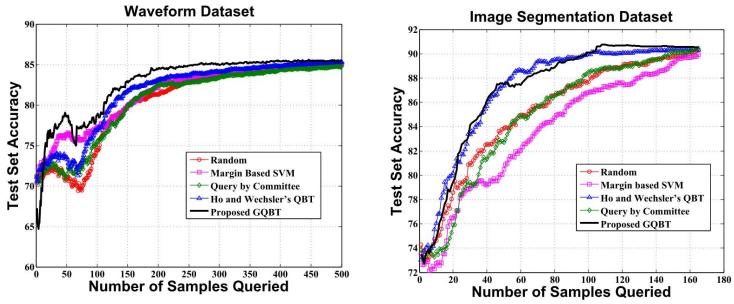
Table 8.1: Datasets from the UCI Machine Learning repository used in our experiments. An equal number of examples from each class was used in the initial training set. For example, for the Breast Cancer dataset, 5 examples from each class were used to form the initial training set of 10 examples.

For each of the datasets, the initial training, testing and unlabeled pools were randomly partitioned three different times and the results were averaged from these 3 runs. Further, in each of the runs, the unlabeled pool was randomly permuted 10 different times to remove any bias on the order in which the points are observed, and the results of these 10 trials were averaged for each run. A polynomial kernel was found to be the most well-suited for all the datasets, as established by the peak accuracies achieved in our results.

The results of our experiments are presented in Figure 8.5 and Table 8.2. In each of these experiments, the formulation of the proposed GQBT approach where the non-conformity measures from the SVM and the k-NN are combined (as in Section 8.2) was used. Table 8.2 shows the label complexity (the percentage of the unlabeled pool that was queried to reach the peak accuracy in the active learning process) of each of the methods.



(a) Breast Cancer dataset (b) Musk dataset (c) Wine dataset



(d) Waveform dataset (e) Image Segmentation dataset

Figure 8.5: Results with datasets from the UCI Machine Learning repository. In the Musk dataset, the results started with an accuracy of $\approx 70\%$, but since all methods had similar initial accuracies, the graph is shown from 85% accuracy onwards, where the differences in performance are clearly seen.

The results are self-explanatory, and demonstrate the improvement in performance gained using the proposed approach.

Application to Face Recognition

To evaluate the performance of the approach on a more challenging real-world problem, we carried out experiments on face recognition from video, where the high redundancy between frames in a video requires an active learning approach. We used the VidTIMIT biometrics dataset [233], of which we used the video recordings of 25 subjects reciting short sentences. Each of the videos are sliced and stored as JPEG images of resolution 512 by 384, on which automated face cropping was performed to crop out the face regions. To extract the facial features, block based discrete cosine transform (DCT) was

Dataset	Random Sampling	Margin-based SVM	Query by Committee	Ho-Wechsler's QBT	Proposed GQBT
Breast Cancer	92.8%	83.6%	80%	46.8%	28%
Musk	77%	55%	72.33%	86.67%	24.33%
Wine	87.5%	78.75%	97.5%	47.5%	35%
Waveform	99.6%	100%	98.2%	98.6%	89.2%
Image Segmentation	100%	100%	100%	98.18%	66.06%

Table 8.2: Label complexities of each of the methods for all the datasets. Label complexity is defined as the percentage of the unlabeled pool that is queried to reach the peak accuracy in the active learning process. Note the low label complexities of the proposed approach in all the cases. Also, note that the label complexities for the other methods on datasets like Waveform and Image Segmentation are very high although the accuracy did increase at a reasonable rate in the active learning process in Figure 8.5. This only implies that these methods reached their peak accuracy when the unlabeled pool was almost exhausted.

used (similar to [235]). Each image was subdivided into 8 by 8 non-overlapping blocks, and the DCT coefficients of each block were then ordered according to the zigzag scan pattern. The DC co-efficient was discarded for illumination normalization, and the first 10 AC co-efficients of each block were selected to form compact local feature vectors. Each local feature vector was normalized to unit norm. Concatenating the features from the individual blocks yielded the global feature vector for the entire image. The cropped face image had a resolution of 128 by 128 and thus the dimensionality of the extracted feature vector was 2560. Principal Component Analysis (PCA) was then applied to reduce the dimension to 100, retaining about 99% of the variance. 50 images of each subject were randomly picked, and divided into the initial training set (10), unlabeled pool (20) and the test set (20). A polynomial kernel was used for the SVM classifier. Similar to the previous set of experiments, the unlabeled pool was randomly permuted 3 different times to remove any bias on the order in which the points are observed, and the results of these 3 trials were averaged. Figure 8.6 shows the results of our experiments. As shown, the proposed GQBT once again demonstrated a significantly improved performance over the other approaches.

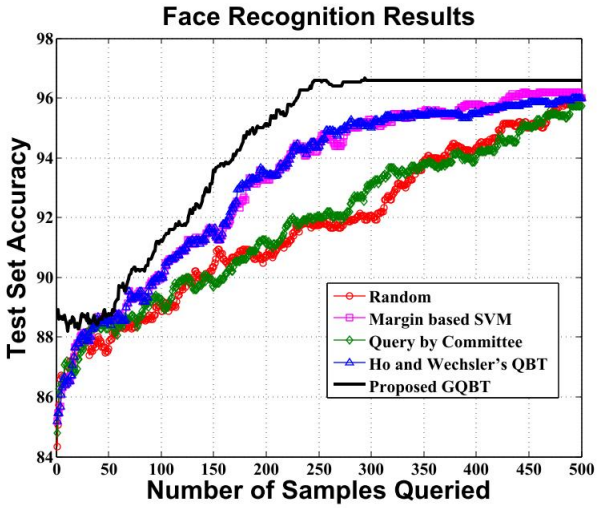


Figure 8.6: Results obtained on the VidTIMIT dataset. Note that the GQBT approach led to a significantly higher peak accuracy, and had a lower label complexity of 58.8% to reach the peak accuracy. Label complexities of the other methods: Ho and Wechsler's QBT - 98.2%; Query by Committee - 100%; Margin-based SVM - 89%; Random sampling - 99.6%

8.4 Discussions

In this chapter, we proposed a Generalized Query by Transduction (GQBT) approach for active learning in the online setting. The results of our experiments with different datasets from the UCI Machine Learning repository and on face recognition from video demonstrated the improvement in performance (and reduction in label complexity) obtained using the proposed research. This approach can be used along with any pattern classification algorithm with the definition of a suitable non-conformity measure. In a binary classifier, the GQBT approach simplifies to the QBT approach proposed by Ho and Wechsler. Further, multiple criteria can be combined using this approach to select appropriate examples from the unlabeled pool, as described in the chapter.

One of the major limitations of this approach, as mentioned earlier, is the computational overhead of transductive inference at each step. With recent advances in incremental classifiers, this limitation can be overcome to a large extent. As future work, we plan to

study other approaches of combining p-values and their influence on the performance of the approach. We also intend to study and identify appropriate stopping criteria for the proposed active learning framework.

Chapter 9

CONCLUSIONS AND FUTURE WORK

The increasing miniaturization of sensing technologies, together with their widespread use, has resulted in the generation of humongous amounts of multimedia data (in the form of images, audio and text among others) in today's world. While this has expanded the possibilities of solving real world problems (such as understanding the behavior of people, objects and activities) using computational learning frameworks, selecting the salient data samples from such huge collections of data has proved to be a significant and practical challenge. Further, to train a reliable classification model, it is indispensable to have a large quantity of labeled training data. Manual annotation of large amounts of data is an expensive process in terms of time, labor and human expertise. This has set the stage for research in the field of batch mode active learning (BMAL) in multimedia pattern recognition applications. Such algorithms automatically identify the salient and representative samples from large quantities of unlabeled data and tremendously reduce human annotation effort in inducing a classification model. BMAL can be applied across all existing classification methods and with any kind of data, thus making it a very generalizable approach. The success of active learning in several multimedia computing applications (such as image retrieval, text/web mining, speech processing and social network analysis) has resulted in the extension of the framework to problem settings beyond regular classification. Batch mode active learning concepts have been extended to newer problem settings like clustering, regression, feature selection and anomaly / rare category detection.

The objective of this dissertation was to develop novel batch mode active learning frameworks for multimedia pattern recognition applications. We specifically focused on computer vision and assistive technology systems (for individuals with visual impair-

ments) as our exemplar application domains. However, the research outcomes of this work are fundamental by their impact and the solutions developed as part of this dissertation are pertinent to a broader audience including information retrieval, document and text mining, spam filtering, healthcare informatics, security systems and in any application where the salient instances need to be automatically identified from large quantities of redundant data. The intellectual merit lies not only in strong contributions in pattern recognition, machine learning and computer vision but also opens up new research directions at the intersection of these disciplines and in assistive technologies and cognitive decision sciences.

9.1 Summary of Contributions

The key contributions made in this dissertation are summarized below:

- **Dynamic Batch Mode Active Learning:** Existing batch mode active learning techniques are all static, that is, they require the batch size as an input parameter for batch selection. However, deciding on a batch size at random and without any knowledge of the data stream in question may lead to poor generalization capacity of the underlying learner. We proposed a dynamic BMAL framework which decides the batch size adaptively based on the complexity of the data stream and the cost of annotation of each unlabeled sample. We developed a stochastic gradient descent based approach to simultaneously solve the batch size and the specific instances to be queried through a single formulation. The method has the same computational complexity as existing static BMAL frameworks where the batch size needs to be pre-specified. We also exploited the properties of sub-modular functions to derive a second adaptive BMAL technique which computes the batch size and selects unlabeled samples for annotation through a single framework. The proposed methodologies were validated on the VidTIMIT and the MOBIO face recognition datasets. Our empirical

evaluations certified the potential of these approaches in appropriately identifying a batch size for a given video stream and consistently delivering high accuracies on unseen samples.

- **Batch Mode Active Learning for Fuzzy Label Problems:** Existing approaches of batch mode active learning schemes are all developed for crisp labels, where there is a clear distinction between the different class labels (like face recognition). However, some problems like facial expression recognition involve a fuzzy label space where there is an inherent imprecision and vagueness in the class label definitions and one class smoothly transitions into the other. A BMAL framework for fuzzy label problems was also proposed as a part of this dissertation. The batch selection problem was represented as the maximization of a sub-modular and monotonically non-decreasing function. A greedy algorithm was used to derive an efficient solution with provable performance guarantees. Our experiments on the MindReading and MMI expression recognition datasets corroborated the potential of this approach over passive sampling and also over crisp BMAL in problems like facial expression recognition.

- **Batch Mode Active Learning with Guaranteed Solution Bounds:** State-of-the-art batch mode active learning frameworks define the batch selection as an NP-hard integer programming problem. Convex relaxations are performed to solve the problem. Even though they depict impressive empirical performance, no formal guarantees have been proved on the qualities of the relaxations. As one of the contributions, the BMAL problem was posed as an NP-hard integer quadratic programming (IQP) problem. Two convex relaxations, one based on linear programming and the other on semi-definite programming were then performed to solve the NP-hard problem. More importantly, a strong deterministic bound was derived on the quality of the first relaxation and a probabilistic bound on the second. The proposed frameworks were

validated on the VidTIMIT and MOBIO face recognition datasets, the MindReading and MMI expression recognition datasets and the Scene and Yeast multi-label datasets. The results depicted that the BatchRank and BatchRand algorithms perform at-par with the state-of-the-art and also deliver high quality solutions. We also demonstrated the robustness of the frameworks to real-world issues like noisy labels and class imbalance.

- **Active Matrix Completion:** Low-rank matrix completion has been extensively used in applications like computer vision and graphics and machine learning among others. However, the problem of integrating human intelligence for matrix completion has not been explored. A framework for active matrix completion has been proposed as one of the contributions in this work. Three different frameworks were developed (based on Conditional Gaussians, committee uncertainty and committee stability) to compute the uncertainty of prediction of every missing entry in the matrix completion process, which was then used to decide the entries to be queried for annotation. Efficient algorithms were used for sparse inverse covariance estimation and singular value decomposition (SVD) of large scale matrices. The algorithms were validated on image and recommendation datasets. The results depicted tremendous promise in using active matrix completion over random or passive sampling in accurately completing a partially observed matrix. The flexibility of the framework was then demonstrated on related active learning problems like transductive active learning, multi-label active learning, active learning in regression and active feature acquisition. The results corroborated the merit of the approach.

Apart from these four basic contributions, the generalizibility of the proposed batch mode active learning frameworks was corroborated by the promising results obtained after varying the criteria for batch selection. The perplexity based objective function (which emphasizes the entropy of the updated learner) depicted the best performance

on challenging real-world face recognition datasets. The framework was also extended to integrate multiple sources of information (like face and speech modalities for person recognition or multiple image features for image classification) for active batch selection. The results depicted the improvement of active batch selection from multiple sources of information over a single source. The flexibility of the framework was further emphasized by extending it to incorporate prior contextual information, which are often available in real-world applications (like the location of a person, the subjects expected to attend a particular meeting). The empirical results depicted improved performance of the context-aware learner over the context-ignorant learner.

A framework for online active learning was also developed using concepts from the Conformal Predictions (CP) theory. This framework offers strong theoretical guarantees on the error frequency of a learner in the online setting. The uncertainty of an unlabeled example was quantified as the largest eigenvalue of the p-value difference matrix, which was used to select samples for manual annotation. We established that the proposed approach is a generalization of a previously proposed query by transduction framework, where only the top two p-values of an unlabeled sample were used to compute the information content. Our algorithm depicted promising results on the UCI Machine Learning Repository and on the VidTIMIT biometric recognition dataset.

These contributions were validated on a number of challenging real-world datasets like the datasets from the UCI Machine Learning Repository, face recognition (VidTIMIT and MOBIO), facial expression recognition (MindReading and MMI), multi-label datasets (Scene and Yeast) and regression problems (FacePix pose estimation, AVEC continuous emotion recognition). All these problems are fundamental challenges in the design and development of a Social Interaction Assistant system, as outlined in Chapter 1. During the course of this work, other contributions related to addressing the problems in these application domains (such as a framework for online active learning using the Conformal

Predictions (CP) theory, an algorithm to maximize efficiency in the CP framework and an algorithm to fuse information from multiple sources using p-values obtained from the CP framework) were also proposed.

Dissemination: The various aspects of the contributions in this work have resulted in a total of 16 peer-reviewed conference and workshop publications, 1 journal publication, 2 book chapters, 1 US patent and 2 US provisional patents. The dissemination venues include the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), ACM Multimedia Conference, the Neural Information Processing Systems (NIPS) conference, Springer Lecture Notes on Computer Science and the Pattern Recognition journal among others. This work was also presented at the Doctoral Symposium of the Association for the Advancement of Artificial Intelligence (AAAI), ACM Multimedia and SIAM Data Mining (SDM) conferences and also at the Multimedia and Vision Meeting organized by IBM Research, the Seventh Annual Machine Learning Symposium organized by the New York Academy of Sciences and at the Machine Learning departments of Microsoft Research, Redmond and Carnegie Mellon University.

9.2 Future Work

The contributions of this dissertation have shown tremendous promise in using batch mode active learning techniques in real-world multimedia recognition applications. The results depict the usefulness of the algorithms in reducing human annotation effort in inducing an appropriate classification / regression model. The possibilities of future work are numerous and a few sample directions are presented in this section.

Dynamic Batch Mode Active Learning

The dynamic batch mode active learning algorithms, detailed in Chapter 3, were derived by appending an L_1 penalty to the objective function for batch selection. As part of our future work, we will explore other mechanisms of dynamic batch size computation (e.g. L_2 regularization as the penalty term) and study the effects on the results.

The problem of adaptive batch selection is closely related to finding the correct number of clusters in a clustering algorithm. Typically, such algorithms define an objective function to quantify the quality of clustering and a penalty is appended on the number of clusters to discourage a high quality clustering with a huge number of clusters. Recent work in this domain has addressed the problem using Dirichlet processes [316, 317] - we plan to investigate these in our ongoing work.

Our future work will also focus on studying the theoretical properties of the frameworks to mathematically establish concrete performance guarantees on the solution qualities for both the dynamic BMAL schemes. Further, in case of the stochastic gradient descent based approach, the quadratic programming problem that needs to be solved as part of the optimization process can be the main bottleneck, which increases the computation time (especially for large scale data). However, there have been recent efforts [318] to efficiently solve QP problems by using a pivoting algorithm and the KKT conditions to significantly reduce computations. This can be used in our approach, making it feasible and meritorious even for large-scale data. We will explore this in our future work.

Batch Mode Active Learning for Fuzzy Classification

The BMAL framework for fuzzy label classification problems like facial expression recognition, based on sub-modular optimization (Chapter 4) assumed every face image as a single unlabeled sample. However, while face recognition can work on single images, it is

more appropriate to consider video streams as data samples for facial expression recognition; an expression can be recognized with certainty over a number of frames rather than a single frame. A promising direction of future research is to extend the active batch selection framework to video streams and study the effect on the results. Future work will also involve rigorous validation of the proposed methodology in large scale learning problems and in other multimedia applications involving fuzzy labels (e.g. recognizing the stages of development of gene expression pattern images).

Batch Mode Active Learning with Guaranteed Solution Bounds

The BatchRank and BatchRand frameworks, discussed in Chapter 5, were based on the assumption that the matrix D which encoded the uncertainty and divergence information had only non-negative entries. As part of our future work, we intend to relax this assumption and prove performance bounds for the case where D can have negative entries. We will also explore efficient methods to compute the Schur complement in order to speed up the SDP solution process in BatchRand. We further plan to extend this work to several other related problems including binary matrix factorization and transfer-active learning.

Active Matrix Completion

The Conditional Gaussian method based active matrix completion framework (presented in Chapter 6) is based on the computation of variance of every unobserved entry of each data sample. This is derived from the sub-matrix of the covariance matrix with rows and columns corresponding to the missing entries of a particular sample. However, apart from the variance, the covariance matrix also provides the correlation among the different entries (given by the off-diagonal elements). It will be interesting to factor the redundancy (correlation) among the different entries in the batch selection criterion for this framework and observe the impact on the results.

Another important direction of future research is to prove performance bounds of the query-by-committee based active matrix completion algorithms (QBC and Stability). There is a rich body of literature on the theoretical guarantees of the QBC algorithm in active learning [63, 319, 140], which quantify the expected number of queries required to attain a certain level of generalization error. It will be interesting to derive analogous bounds for the QBC-based algorithms.

Further, in certain applications, the data is in the form of an array of matrices (known as tensors). For instance, in a video stream, each image can be represented as a matrix of pixel values and the entire video is an array of matrices, one at every time point. Recently, a tensor completion approach has been proposed [273], to estimate missing values in visual data. We intend to extend our active matrix completion algorithms to tensor completion for active sample selection in higher order matrices.

Other Possible Directions of Future Work

Limitations and avenues of future work specific of each of the contributions in this dissertation were stated above. However, there are other possible directions of future work pertaining to active learning in general. Pointers to these directions are briefly described below:

Hierarchical Batch Mode Active Learning: Classification problems usually assume a flat label space. However, in some applications, the label space can be organized in the form of a tree hierarchy. A classical example of such a problem is facial expression recognition which, as described in the context of the Social Interaction Assistant, is one of the main motivating applications of this work. The MindReading dataset (used in the context of fuzzy expression recognition) categorizes human expression into 24 basic emotions and also presents a hierarchy of each emotion depending on its degree/strength. For example, *happy* is a base emotion with 6 subclasses *merry, delighted, amused, tri-*

umphant, jubilant and *exonerated*. This is helpful in studying the emotional state of a person in more intricate details. As another example, consider a malware classification system, which classifies files as clean or malicious. A malware is often represented as a hierarchical structure consisting of a type (e.g. Worm, Backdoor), a platform (e.g. Win 32, Win NT), a family (e.g. Rbot, Puce) and a generation (e.g. Gen!A, Gen!B) [320]. The problem of hierarchical classification has been addressed in the literature [321, 322]. However, to the best of our knowledge, no active learning scheme has been proposed for hierarchical label spaces. An interesting direction of future research is to develop a batch mode active learning framework specially tailored for hierarchical label spaces.

Temporal Redundancy in Video Streams: Most of the active learning algorithms in this dissertation were developed for video based multimedia applications. The video stream was decomposed into images and query functions were designed to select a batch of informative image samples for manual annotation. However, the images forming a video stream share a temporal correlation and by treating each image sample independently, the temporal redundancy has been ignored. An interesting direction of future research is to integrate the temporal component of the video streams in the objective function for batch selection and study the impact on the results.

Alternative Feedback Types: The user feedback in the proposed active learning algorithms was in the form of class labels for classification / regression problems; that is, the active learning algorithms identified a set of promising instances whose class labels were acquired from human oracles. However, it is possible (and sometimes even beneficial) to conceive of other means of user feedback to the learning algorithm. Joshi *et al.* [158] developed a multi-class active learning framework where the user input was binary - in each iteration, the algorithm selected a pair of unlabeled samples and the user merely had to specify whether or not the two samples belonged to the same class. The authors established that the binary feedback system minimized the user interaction time with the

system and succeeded in obtaining more accurate user inputs. A possible direction of future research is to explore applications which require other means of user supervision to the learning systems. For example, in a crowd-sourced setting, the objective is to estimate the underlying scores of objects based on absolute and preference judgments provided by a set of human users. The user input in such an application is in the form of full or partial rankings, relative item comparisons or a combination [323, 324]. An active learner designed for such an application will therefore need to identify a promising subset of objects which need to be ranked manually to better infer the gold-standard scores of all the samples. It will also be interesting to develop active learning algorithms where the user input is in the form of the distribution of class labels in the overall data corpus instead of the label of a single unlabeled sample. Development of new forms of supervision in machine learning applications also opens the door to alternative forms of active learning.

Stopping Criteria: An important aspect of any active learning framework is to know when to stop learning. As noted in some of our results, the generalization accuracy may reach a peak value and then start falling down as more samples are queried. Overfitting may be one possible reason to explain this observation. There have been some efforts to address this problem [325, 326, 327]. These methods are fairly similar, based on the notion of an intrinsic measure of stability or self-confidence of the learner and that active learning ceases to be useful once this measure begins to degrade. However, all these algorithms are based on an intrinsic learner-decided threshold and are heuristic in nature. Identifying an appropriate stopping criterion for active learning is still an open problem and is a promising direction of future research.

REFERENCES

[1] S. Krishna, D. Colbry, J. Black, V. Balasubramanian, and S. Panchanathan, "A systematic requirements analysis and development of an assistive device to enhance the social interaction of people who are blind or visually impaired," in *ECCV Workshop on Computer Vision Applications for the Visually Impaired (CVAVI)*, 2008.

[2] M. Betke, "Intelligent interfaces to empower people with disabilities," in *Springer Verlag*, 2010.

[3] S. Krishna, G. Little, J. Black, and S. Panchanathan, "A wearable face recognition system for individuals with visual impairments," in *Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, 2005.

[4] T. McDaniel, S. Krishna, V. Balasubramanian, D. Colbry, and S. Panchanathan, "Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind," in *IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE)*, 2008.

[5] X. Zhu, "Semi-supervised learning with graphs," in *PhD Thesis, Carnegie Mellon University*, 2005.

[6] B. Settles, M. Craven, and L. Friedland, "Active learning with real annotation costs," in *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Cost-Sensitive Learning*, 2008.

[7] S. Tong, "Active learning : Theory and applications," in *PhD Thesis, Stanford University*, 2001.

[8] E. Horvitz and G. Rutledge, "Time dependent utility and action under uncertainty," in *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 1991.

[9] J. Latombe, "Robot motion planning," in *Kluwer Academic Publishers*, 1991.

[10] D. Heckerman, J. Breese, and K. Rommelse, "Troubleshooting under uncertainty," in *Technical Report, Microsoft Research*, 1994.

[11] A. Wald, "Statistical decision functions," in *Wiley, New York*, 1950.

[12] E. Dale, "Audio-visual methods in teaching," in *New York : The Dryden Press*, 1946.

[13] A. Chickering and Z. Gamson, "Seven principles for good practice," in *AAHE Bulletin*, 1987.

[14] C. Bonwell and J. Eison, "Active learning: Creating excitement in the classroom," in *ASHE-ERIC Higher Education Report*, 1991.

[15] D. Johnson, R. Johnson, and K. Smith, "Active learning: Cooperation in the college classroom," in *Edina, MN: Interaction Book Company*, 1991.

[16] M. Prince, "Does active learning work? a review of the research," in *Journal of Engineering Education*, 2004.

[17] B. Settles and M. Craven, "An analysis of active learning strategies for sequence labeling tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[18] I. Dagan and S. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1995.

[19] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proceedings of the International Conference on Advances in Intelligent Data Analysis (CAIDA)*, 2001.

[20] J. Baldridge and M. Osborne, "Active learning and the total cost of annotation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2004.

[21] R. Hwa, "Sample selection for statistical parsing," in *Computational Linguistics*, 2004.

[22] C. Thompson, M. Califf, and R. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1999.

[23] Z. Zheng and B. Padmanabhan, "On active learning for data acquisition," in *IEEE International Conference on Data Mining (ICDM)*, 2002.

[24] P. Melville, M. Saar-Tsechansky, F. Provost, and R. Mooney, "Active feature-value acquisition for classifier induction," in *IEEE International Conference on Data Mining (ICDM)*, 2004.

[25] M. Saar-Tsechansky, P. Melville, and F. Provost, "Active feature-value acquisition," in *Management Science*, 2009.

[26] R. Greiner, A. Grove, and D. Roth, "Learning cost-sensitive active classifiers," in *Artificial Intelligence*, 2002.

[27] X. Chai, L. Deng, Q. Yang, and C. Ling, "Test-cost sentistive naive bayes classification," in *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2004.

[28] C. Ling, Q. Yang, J. Wang, and S. Zhang, "Decision trees with minimal costs," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[29] V. Sheng and C. Ling, "Feature value acquisition in testing: A sequential batch test algorithm," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

[30] S. Ji and L. Carin, "Cost-sensitive feature acquisition and classification," in *Pattern Recognition*, 2007.

[31] A. Kapoor and E. Horvitz, "Breaking boundaries: Active information acquisition across learning and diagnosis," in *Advances of Neural Information Processing Systems (NIPS)*, 2009.

[32] V. Sindhwani, P. Melville, and R. Lawrence, "Uncertainty sampling and transductive experimental design for active dual supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[33] X. Kong, W. Fan, and P. Yu, "Dual active feature and sample selection for graph classification," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[34] R. Lomasky, C. Brodley, M. Aernecke, D. Walt, and M. Friedl, "Active class selection," in *Proceedings of the European Conference on Machine Learning (ECML)*, 2007.

[35] T. Hofmann and J. Buhmann, "Active data clustering," in *Advances in Neural Information Processing Systems (NIPS)*, 1998.

[36] N. Grira, M. Crucianu, and N. Boujemaa, "Active semi-supervised fuzzy clustering for image database categorization," in *Proceedings the ACM Workshop on Multimedia Information Retrieval (MIR)*, 2005.

[37] Y. Huang and T. Mitchell, "Text clustering with extended user feedback," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006.

[38] F. Wauthier, N. Jojic, and M. Jordan, "Active spectral clustering via iterative uncertainty reduction," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2012.

[39] A. Biswas and D. Jacobs, "Active image clustering: Seeking constraints from humans to complement algorithms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[40] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Technical Report, University of Wisconsin-Madison*, 2005.

[41] M. Sugiyama and N. Rubens, "Active learning with model selection in linear regression," in *SIAM Data Mining Conference (SDM)*, 2008.

[42] M. Sugiyama, "Active learning in approximately linear regression based on conditional expectation of generalization error," in *Journal of Machine Learning Research (JMLR)*, 2006.

[43] K. Yu, J. Bi, and V. Tresp, "Active learning via transductive experimental design," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

[44] R. Burbidge, J. Rowland, and R. King, "Active learning for regression based on query by committee," in *Lecture Notes in Computer Science (LNCS)*, 2007.

[45] D. Hsu, S. Kakade, J. Langford, and T. Zhang, "Multi-label prediction via compressed sensing," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[46] G.Tsoumakas and I. Katakis, "Multi-label classification: An overview," in *International Journal of Data Warehousing and Mining*, 2007.

[47] M. Singh, E. Curran, and P. Cunningham, "Active learning for multi-label image annotation," in *Technical Report, University College Dublin*, 2009.

[48] K. Brinker, "On active learning in multi-label classification," in *SpringerLink*, 2006.

[49] X. Zhang, J. Cheng, C. Xu, H. Lu, and S. Ma, "Multi-view multi-label active learning for image classification," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2009.

[50] X. Li, L. Wang, and E. Sung, "Multi-label SVM active learning for image classification," in *IEEE International Conference on Image Processing (ICIP)*, 2004.

[51] B. Yang, J. Sun, T. Wang, and Z. Chen, "Effective multi-label active learning for text classification," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2009.

[52] C. Hung and H. Lin, "Multi-label active learning with auxiliary learner," in *Journal of Machine Learning Research (JMLR) - Proceedings Track*, 2011.

[53] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang, "Two-dimensional multi-label active learning with an efficient online adaptation model for image classification," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[54] T. Dietterich, R. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," in *Artificial Intelligence*, 1997.

[55] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[56] B. Settles, "Curious machines : Active learning with structured instances," in *PhD Thesis, University of Wisconsin Madison*, 2008.

[57] D. Liu, X. Hua, L. Yang, and H. Zhang, "Multiple-instance active learning for image categorization," in *Springer-Verlag*, 2009.

[58] A.Schein and L.Ungar, "Active learning for logistic regression: An evaluation," in *Machine Learning*, 2007.

[59] Y. Guo and D. Schuurmans, "Discriminative batch mode active learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2007.

[60] C. Gasperin, "Active learning for anaphora resolution," in *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, 2009.

[61] J. Baldridge and A. Palmer, "How well does active learning actually work? time-based evaluation of cost-reduction strategies for language documentation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2009.

[62] K. Tomanek and F. Olsson, "A web survey on the use of active learning to support annotation of text data," in *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing*, 2009.

[63] Y. Freund, S. Seung, E. Shamir, and N. Tishby, "Selective sampling using the query by committee algorithm," in *Machine Learning*, 1997.

[64] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Theory of Probability and Its Applications*, 1971.

[65] R. Gilad-Bachrach, A. Navot, and N. Tishby, "Query by committee made real," in *Advances in Neural Information Processing Systems (NIPS)*, 2006.

[66] W. Wang and Z. Zhou, "Multi-view active learning in the non-realizable case," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[67] ——, "On multi-view active learning and the combination with semi-supervised learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[68] M. Raginsky and A. Rakhlin, "Lower bounds for passive and active learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[69] D. Golovin, A. Krause, and D. Ray, "Near-optimal bayesian active learning with noisy observations," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[70] S. Dasgupta, A. Kalai, and C. Monteleoni, "Analysis of perceptron-based active learning," in *Proceedings of the Conference on Learning Theory (COLT)*, 2005.

[71] S. Dasgupta, "Analysis of a greedy active learning strategy," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[72] ——, "Coarse sample complexity bounds for active learning," in *Adavnces in Neural Information Processing Systems (NIPS)*, 2005.

[73] A. Beygelzimer, S. Dasgupta, and J. Langford, "Importance weighted active learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[74] M. Balcan, S. Hanneke, and J. Wortman, "The true sample complexity of active learning," in *Proceedings of the Conference on Learning Theory (COLT)*, 2008.

[75] S. Dasgupta, "Two faces of active learning," in *Theoretical Computer Science*, 2011.

[76] L. Valiant, "A theory of the learnable," in *Communications of the ACM*, 1984.

[77] M. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2006.

[78] S. Hanneke, "A bound on the label complexity of agnostic active learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

[79] S. Dasgupta, D. Hsu, and C. Monteleoni, "A general agnostic active learning algorithm," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[80] A. Beygelzimer, D. Hsu, J. Langford, and T. Zhang, "Agnostic active learning without constraints," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[81] L. Wang, "Sufficient conditions for agnostic active learnable," in *Advances of Neural Information Processing Systems (NIPS)*, 2009.

[82] S. Dasgupta and D. Hsu, "Hierarchical sampling for active learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[83] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the Conference on Learning Theory (COLT)*, 1998.

[84] A. McCallum and K. Nigam, "Employing EM and Pool-Based active learning for text classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1998.

[85] I. Muslea, S. Minton, and C. Knoblock, "Selective sampling with redundant views," in *Proceedings of the National Conference on Artificial Intelligence*, 2000.

[86] X. Zhu, J. Lafferty, and Z. Ghahramani, "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions," in *Proceedings of the ICML Workshop on the Continuum from Labeled to Unlabeled Data*, 2003.

[87] Z. Zhou, K. Chen, and Y. Jiang, "Exploiting unlabeled data in content-based image retrieval," in *Proceedings of the European Conference on Machine Learning (ECML)*, 2004.

[88] A. Guillory and J. Bilmes, "Online submodular set cover, ranking and repeated active learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[89] X. He, M. Ji, and H. Bao, "A unified active and semi-supervised learning framework for image compression," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[90] R. Sutton and A. Barto, "Reinforcement learning: An introduction," in *MIT Press*, 1998.

[91] L. Mihalkova and R. Mooney, "Using active relocation to aid reinforcement learning," in *Proccedings of the Florida Artificial Intelligence Research Society (FLAIRS)*, 2006.

[92] A. Epshteyn, A. Vogel, and G. DeJong, "Active reinforcement learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[93] S. Hoi and R. Jin, "Active kernel learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[94] D. Angluin, "Queries and concept learning," in *Machine Learning*, 1988.

[95] S. Hoi, R. Jin, and M. Lyu, "Batch mode active learning with applications to text categorization and image retrieval," in *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2009.

[96] A. da Silva, A. falcao, and L. Magalhaes, "Active learning paradigms for cbir systems based on optimum-path forest classification," in *Pattern Recognition*, 2011.

[97] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Batch mode active learning and its application to medical image classification," in *International Conference on Machine Learning (ICML)*, 2006.

[98] S. Hoi, R. Jin, and M. Lyu, "Large-scale text categorization by batch mode active learning," in *International Conference on World Wide Web (WWW)*, 2006.

[99] B. Settles, "Active learning literature survey," in *Technical Report, University of Wisconsin Madison*, 2010.

[100] D. Sculley, "Online active learning methods for fast label-efficient spam filtering," in *Fourth Conference on Email and AntiSpam*, 2007.

[101] N. Bianchi, A. Conconi, and C. Gentile, "Learning probabilistic linear-threshold classifiers via selective sampling," in *Lecture Notes in Artificial Intelligence*, 2003.

[102] C. Monteleoni and M. Kaariainen, "Practical online active learning for classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[103] M. Dredze and K. Crammer, "Active learning with confidence," in *Proceedings of the Association on Computational Linguistics (ACL)*, 2008.

[104] S. Ho and H. Wechsler, "Query by transduction," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2008.

[105] V. Balasubramanian, S. Chakraborty, and S. Panchanathan, "Generalized query by transduction for online active learning," in *IEEE International Conference on Computer Vision (ICCV), Workshop on Online Learning for Computer Vision*, 2009.

[106] P. Melville, S. Yang, M. Saar-Tsechansky, and R. Mooney, "Active learning for probability estimation using jensen-shannon divergence," in *European Comference on Machine Learning (ECML)*, 2005.

[107] J. Attenberg and F. Provost, "Online active inference and learning," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[108] C. Mesterharm and M. Pazzani, "Active learning using on-line algorithms," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[109] W. Chu, M. Zinkevich, L. Li, A. Thomas, and B. Tseng, "Unbiased online active learning in data streams," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[110] A. Guillory and J. Bilmes, "Interactive submodular set cover," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[111] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," in *Journal of Machine Learning Research (JMLR)*, 2000.

[112] S. Tong, "Active learning : Theory and applications," in *Ph.D. Thesis, Stanford University*, 2001.

[113] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, 2001.

[114] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2000.

[115] P. Mitra, C. Murthy, and S. Pal, "A probabilistic active support vector learning algorithm," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2004.

[116] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2000.

[117] J. Cheng and K. Wang, "Active learning for image retrieval with co-SVM," in *Pattern Recognition*, 2007.

[118] T. Osugi, D. Kun, and S. Scott, "Balancing exploration and exploitation: A new algorithm for active machine learning," in *IEEE International Conference on Data Mining (ICDM)*, 2005.

[119] S. Ebert, M. Fritz, and B. Schiele, "Ralf: A reinforced active learning formulation for object class recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[120] C. Loy, T. Hospedales, T. Xiang, and S. Gong, "Stream-based joint exploration-exploitation active learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[121] C. Shannon, "A mathematical theory of communication," in *Bell System Technical Journal*, 1948.

[122] A. Holub, P. Perona, and M. Burl, "Entropy-based active learning for object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2008.

[123] D. MacKay, "Information-based objective functions for active data selection," in *Neural Computation*, 1992.

[124] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," in *Machine Learning*, 1994.

[125] S. Ho and H. Wechsler, "Transductive confidence machine for active learning," in *International Joint Conference on Neural Networks (IJCNN)*, 2003.

[126] M. Li and I. Sethi, "Confidence-Based active learning," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2006.

[127] M. Tang, X. Luo, and S. Roukos, "Active learning for statistical natural language parsing," in *Proceedings Of ACL*, 2002.

[128] D. Lewis and W. Gale, "A sequential algorithm for training text classifiers," in *Proceedings of the ACM SIGIR conference*, 1994.

[129] M. Park, G. Horwitz, and J. Pillow, "Active learning of neural response functions with gaussians processes," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[130] Y. Yan, R. Rosales, G. Fung, and J. Dy, "Active learning from crowds," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[131] A. Kowdle, Y. Chang, A. Gallagher, and T. Chen, "Active learning for piecewise planar 3D reconstruction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[132] J. Roder, B. Nadler, K. Kunzmann, and F. Hamprecht, "Active learning with distributional estimates," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[133] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[134] N. Roy and A. McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2001.

[135] Y. Guo and R. Greiner, "Optimistic active learning using mutual information," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[136] X. He and D. Cai, "Active subspace learning," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[137] D. Cohn, "Neural network exploration using optimal experiment design," in *Advances in Neural Information Processing Systems (NIPS)*, 1994.

[138] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," in *Journal of Artificial Intelligence Research (JAIR)*, 1996.

[139] K. Bellare, S. Iyengar, A. Parameswaran, and V. Rastogi, "Active sampling for entity matching," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2012.

[140] H. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the ACM Workshop on Computational Learning Theory*, 1992.

[141] R. Liere and P. Tadepalli, "Active learning with committees for text categorization," in *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, 1997.

[142] N. Abe and H. Mamitsuka, "Query learning strategies using boosting and bagging," in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, 1998.

[143] I. Dagan, "Committee-based sample selection for probabilistic classifiers," in *Journal of Artificial Intelligence Research (JAIR)*, 1999.

[144] Q. Zhang and S. Sun, "Multiple-view multiple-learner active learning," in *Pattern Recognition*, 2010.

[145] A. Fujii, T. Tokunaga, K. Inui, and H. Tanaka, "Selective sampling for example-based word sense disambiguation," in *Computational Linguistics*, 1998.

[146] M. Lindenbaum, S. Markovitch, and D. Rusakov, "Selective sampling for nearest neighbor classifiers," in *Machine Learning*, 2004.

[147] T. Gao and D. Koller, "Active classification based on value of classifier," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[148] L. Yang, "Active learning with a drifting distribution," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[149] Y. Baram, R. El-Yaniv, and K. Luz, "Online choice of active learning algorithms," *JMLR*, vol. 5, 2004.

[150] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proceedings of International Conference on Machine Learning (ICML)*, 2001.

[151] K. Nigam, A. Mccallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," 1999.

[152] D. Pelleg and A. Moore, "Active learning for anomaly and rare-category detection," in *Advances of Neural Information Processing Systems (NIPS)*, 2004.

[153] A. Schein and L. Ungar, "A-optimality for active learning of logistic regression classifiers," in *The University of Pennsylvania Department of Computer and Information Science Technical Report*, 2004.

[154] H. Tat, H. Nguyen, and A.Smeulders, "Active learning using pre-clustering," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

[155] A. Kapoor, G. Hua, A. Akbarzadeh, and S. Baker, "Which faces to tag : Adding prior constraints into active learning," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2009.

[156] G. Qi, X. Hua, Y. Rui, J. Tang, and H. Zhang, "Two-dimensional active learning for image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[157] R. Kothari and V. Jain, "Learning from labeled and unlabeled data using a minimal number of queries," in *IEEE Transactions on Neural Networks*, 2003.

[158] A. Joshi, F. Porikli, and N. Papanikolopoulos, "Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[159] S. Tong and D. Koller, "Active learning for parameter estimation in bayesian networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2000.

[160] M. Park and J. Pillow, "Bayesian active learning with localized priors for fast receptive field characterization," in *Advances of Neural Information Processing Systems (NIPS)*, 2012.

[161] M. Osborne, D. Duvenaud, R. Garnett, C. Rasmussen, S. Roberts, and Z. Ghahramani, "Active learning of model evidence using bayesian quadrature," in *Advances of Neural Information Processing Systems (NIPS)*, 2012.

[162] H. Tyagi and V. Cevher, "Active learning of multi-index function models," in *Advances of Neural Information Processing Systems (NIPS)*, 2012.

[163] C. Sawade, N. Landwehr, and T. Scheffer, "Active comparison of prediction models," in *Advances of Neural Information Processing Systems (NIPS)*, 2012.

[164] ——, "Active estimation of f-measures," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[165] C. Sawade, N. Landwehr, S. Bickel, and T. Scheffer, "Active risk estimation," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[166] N. Ailon, "Active learning ranking from pairwise preferences with almost optimal query complexity," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[167] K. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[168] L. Charlin, R. Zemel, and C. Boutilier, "Active learning for matching problems," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[169] P. Jain, S. Vijayanarasimhan, and K. Grauman, "Hashing hyperplane queries to near points with applications to large-scale active learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[170] R. Garnett, Y. Krishnamurthy, X. Xiong, J. Schneider, and R. Mann, "Bayesian optimal active search and surveying," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[171] P. Rashidi and D. Cook, "Ask me better questions:active learning queries based on rule induction," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[172] K. Judah, A. Fern, and T. Dietterich, "Active imitation learning via reduction to i.i.d. active learning," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2012.

[173] K. Deng, J. Pineau, and S. Murphy, "Active learning for developing personalized treatment," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2011.

[174] C. Vondrick and D. Ramanan, "Video annotation and trcaking with active learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2011.

[175] S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," in *Europen Conference on Computer Vision (ECCV)*, 2012.

[176] P. Jain and A. Kapoor, "Active learning for large multi-class problems," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[177] S. Vijayanarasimhan and K. Grauman, "Large-scale live active learning:training object detectors with crawled data and crowds," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[178] A. Kovashka, S. Vijayanarasimhan, and K. Grauman, "Actively selecting annotations among objects and attributes," in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[179] X. Wei and Z. Yang, "Coached active learning for interactive video search," in *ACM Multimedia Conference (ACM MM)*, 2011.

[180] A. Vezhnevets, J. Buhmann, and V. Ferrari, "Active learning for semantic segmentation with expected change," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[181] B. Siddiquie and A. Gupta, "Beyond active noun tagging: Modeling contextual interactions for multi-class active learning," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[182] S. Huang, R. Jin, and Z. Zhou, "Active learning by querying informative and representative examples," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[183] L. Li, X. Jin, S. Pan, and J. Sun, "Multi-domain active learning for text classification," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2012.

[184] N. Bianchi, C. Gentile, F. Vitale, and G. Zappella, "A linear time active learning algorithm for link classification," in *Advances of Neural Information Processing Systems (NIPS)*, 2012.

[185] M. Bilgic, L. Mihalkova, and L. Getoor, "Active learning for networked data," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[186] C. Moore, X. Yan, Y. Zhu, J. Rouquier, and T. Lane, "Active learning for node classification in assortative and disassortative networks," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2011.

[187] A. Kuwadekar and J. Neville, "Relational active learning for joint collective classification models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.

[188] A. Harpale and Y. Yang, "Active learning for multi-task adaptive filtering," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[189] A. Borisov, E. Tuv, and G. Runger, "Active batch learning with stochastic query-by-forest(sqbf)," in *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[190] Y. Chen and S. Mani, "Active learning for unbalanced data in the challenge with multiple models and biasing," in *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[191] G. Cawley, "Baseline methods for active learning," in *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[192] L. Lan, H. Shi, Z. Wang, and S. Vucetic, "An active learning algorithm based on parzen window classification," in *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[193] Z. Bodo, Z. Minier, and L. Csato, "Active learning with clustering," in *JMLR Workshop on Active Learning and Experimental Design*, 2011.

[194] M. Salganicoff, L. Ungar, and R. Bajcsy, "Active learning for vision-based robot grasping," in *Machine Learning*, 1996.

[195] A. Morales and E. Chinellato, "Active learning for robot manipulation," in *ECAI*, 2004.

[196] C. Dima, "Active learning for outdoor perception," in *Ph.D thesis, Robotics Institute, Carnegie Mellon University*, 2006.

[197] B. Zhang and S. Kim, "An evolutionary method for active learning of mobile robot path planning," in *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 1997.

[198] R. Cantin, O. D. Freitas, A. Doucet, and J. Castellanos, "Active policy learning for robot planning and exploration under uncertainty," in *Proceedings of Robotics: Science and Systems*, 2007.

[199] C. Dima, M. Hebert, and A. Stentz, "Enabling learning from large datasets: Applying active learning to mobile robotics," in *International Conference on Robotics and Automation (ICRA)*, 2004.

[200] A. Tapus and M. Mataric, "Towards active learning for socially assistive robots," in *Intelligent Service Robotics*, 2008.

[201] J. Wiens and J. Guttag, "Active learning applied to patient-adaptive heartbeat classification," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[202] M. Burl and E. Wang, "Active learning for directed exploration of complex systems," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[203] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.

[204] X. Ding, Y. Zhao, and Y. Li, "A global optimization of SVM batch active learning," in *IEEE International Conference on Intelligent Computing and Intelligent Systems*, 2009.

[205] X. Zhang, D. Zhao, L. Chen, and W. Min, "Batch mode active learning based multi-view text classification," in *International Conference on Fuzzy Systems and Knowledge Discovery*, 2009.

[206] Z. Xu, K. Yu, V. Tresp, X. Xu, and J. Wang, "Representative sampling for text classfication using support vector machines," in *European Conference on Information Retrieval*, 2003.

[207] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," in *IEEE Transactions on Geoscience and Remote Sensing*, 2011.

[208] S. Ananthakrishnan, R. Prasad, D. Stallard, and P. Natarajan, "A semi-supervised batch-mode active learning strategy for improved statistical machine translation," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 2010.

[209] L. Shi, Y. Zhao, and J. Tang, "Batch mode active learning for networked data," in *ACM Transactions on Embedded Computing Systems*, 2011.

[210] S. Hoi, R. Jin, J. Zhu, and M. Lyu, "Semi-supervised SVM batch mode active learning for image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[211] A. Joshi, F. Porikli, and N. Papanikolopoulos, "Multi-class batch-mode active learning for image classification," in *International Conference on Robotics and Automation (ICRA)*, 2010.

[212] L. Shi and Y. Zhao, "Batch mode sparse active learning," in *IEEE International Conference on Data Mining (ICDM) Workshops*, 2010.

[213] J. Azimi, A. Fern, X. Fern, G. Borradaile, and B. Heeringa, "Batch active learning via co-ordinated matching," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[214] W. Zhao, J. Long, E. Zhu, and Y. Liu, "A scalable algorithm for graph-based active learning," in *Springer Link*, 2008.

[215] J. Long, J. Yin, W. Zhao, and E. Zhu, "Graph-based active learning based on label propagation," in *Springer Link*, 2008.

[216] S. Vijayanarasimhan, P. Jain, and K. Grauman, "Far-sighted active learning on a budget for image and video recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[217] Y. Guo, "Active instance sampling via matrix partition," in *Advances of Neural Information Processing Systems (NIPS)*, 2010.

[218] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances of Neural Information Processing Systems (NIPS)*, 2007.

[219] N. Komodakis and G. Tziritas, "Image completion using global optimization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[220] M. Kurucz, A. Benczur, and K. Csalogany, "Methods for large scale svd with missing values," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2007.

[221] P. Tan, M. Steinbach, and V. Kumar, "Introduction to data mining," 2006.

[222] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Advances of Neural Information Processing Systems (NIPS)*, 2005.

[223] D. Shen, J. Zhang, J. Su, G. Zhou, and C. Tan, "Multi-criteria-based active learning for named entity recognition," in *Proceedings of the Association for Computational Linguistics (ACL)*, 2004.

[224] A. Gretton, K. Borgwardt, M. Rasch, B. Scholkopf, and A. Smola, "A kernel method for the two sample problem," in *Advances of Neural Information Processing Systems (NIPS)*, 2007.

[225] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, B. Scholkopf, and A. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," in *Bioinformatics*, 2006.

[226] J. Weston, A. Elisseeff, B. Scholkopf, and M. Tipping, "Use of the zero norm with linear models and kernel methods," in *Journal of Machine Learning Research (JMLR)*, 2003.

[227] J. Nocedal and S. J. Wright, *Numerical optimization*. Springer, 1999.

[228] J. Wang, T. Jebara, and S. Chang, "Graph transduction via alternating minimization," in *Proceedings of International Conference on Machine Learning (ICML)*, 2008.

[229] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of the approximations for maximizing submodular set functions," in *Mathematical Programming*, 1978.

[230] A. Krause and C. Guestrin, "Near-optimal nonmyopic value of information in graphical models," in *UAI*, 2005.

[231] J. Y. Kim, D. Y. Ko, and S. Y. Na, "Implementation and enhancement of GMM face recognition systems using flatness measure," in *Robot and Human Interactive Communication*, 2004.

[232] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, Oct. 2007.

[233] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM Verlag, Jun. 2008.

[234] S. Marcel, C. McCool, and P. Matejka, "Mobile biometry (mobio) face and speaker verification evaluation," *Idiap Research Institute, Technical Report*, 2010.

[235] H. Ekenel, M. Fischer, Q. Jin, and R. Stiefelhagen, "Multi-modal person identification in a smart environment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[236] S. Chakraborty, V. Balasubramanian, and S. Panchanathan, "Dynamic batch size selection for batch mode active learning in biometrics," in *IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2010.

[237] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," in *Journal of Machine Learning Research (JMLR)*, 2008.

[238] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshop on HCI*, 2003.

[239] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2006.

[240] P. Ekman, "Facial expression and emotion," in *American Psychologist*, 1993.

[241] ——, "Strong evidence for universals in facial expressions: A reply to russell's mistaken critique," in *Psychological Bulletin*, 1994.

[242] R. Zhi, Q. Ruan, and Z. Miao, "Fuzzy discriminant projections for facial expression recognition," in *IEEE International Conference on Pattern Recognition (ICPR)*, 2008.

[243] A. Khanam, M. Shafiq, and M. Akram, "Fuzzy based facial expression recognition," in *IEEE Congress on Image and Signal Processing*, 2008.

[244] S. Tsai, J. Jiang, C. Wu, and S. Lee, "A fuzzy similarity-based approach for multilabel document classification," in *IEEE Workshop on Computer Science and Engineering*, 2009.

[245] I. Lizarazo and J. Barros, "Fuzzy image segmentation for urban land-cover classification," in *Photogrammetry Remote Sensing and Spatial Information Sciences*, 2010.

[246] L. Zadeh, "Fuzzy sets," in *Information and Control*, 1965.

[247] N. Singpurwalla and J. Booker, "Membership functions and probability measures of fuzzy sets," in *Journal of the American Statistical Association*, 2004.

[248] S. Al-sharhan, F. Karray, W. Gueaieb, and O. Basir, "Fuzzy entropy : A brief survey," in *IEEE International Fuzzy Systems Conference*, 2001.

[249] M.Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2005.

[250] R. El-Kaliouby and P. Robinson, "Mind reading machines: Automated inference of cognitive mental states from video," in *IEEE International Conference on System, man and Cybernetics*, 2004.

[251] P. Viola and M. Jones, "Robust real-time face detection," in *International Journal of Computer Vision (IJCV)*, 2004.

[252] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," in *Image and Vision Computing (IVC)*, 2006.

[253] N. Pizzi and W. Pedrycz, "Fuzzy set theoretic adjustment to training set class labels using robust location measures," in *International Joint Conference on Neural Networks (IJCNN)*, 2000.

[254] M. Balcan, S. Hanneke, and J. Vaughan, "The true sample complexity of active learning," in *Machine Learning*, 2010.

[255] M. Kukar, "Transductive reliability estimation for medical diagnosis," in *Artifical Intelligence in medicine*, 2003.

[256] M. Goemans and D. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," in *Journal of the Association for Computing Machinery*, 1995.

[257] A. Frank and A. Asuncion, "Uci machine learning repository," in *University of California, Irvine, School of Information and Computer Sciences*, 2010.

[258] M. Zhang and K. Zhang, "Multi-label learning by exploiting label dependency," in *ACM Conference on Knowledge Discovery from Data (KDD)*, 2010.

[259] M. Boutell, J. Luo, X. Shen, and C. Brown, "Learning multi-label scene classification," in *Pattern Recognition*, 2004.

[260] A. Elisseeff and J. Weston, "A kernel method for multi-labeled classification," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

[261] L. Yuan, Y. Wang, P. Thompson, V. Narayan, and J. Ye, "Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data," in *NeuroImage Journal*, 2012.

[262] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. Altman, "Missing value estimation methods for dna microarrays," in *Bioinformatics*, 2001.

[263] T. Schneider, "Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values," in *Journal of Climate*, 2001.

[264] E. Candes and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," in *IEEE Transactions on Information Theory*, 2010.

[265] B. Recht, "A simpler approach to matrix completion," in *Journal of Machine Learning Research (JMLR)*, 2011.

[266] J. Cai, E. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," in *SIAM Journal on Optimization*, 2010.

[267] S. Ma, D. Goldfarb, and L. Chen, "Fixed point and bregman iterative methods for matrix rank minimization," in *Mathematical Programming*, 2009.

[268] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, 2001.

[269] E. Candes and B. Recht, "Exact matrix completion via convex optimization," in *Foundations of Computational Mathematics*, 2009.

[270] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," in *SIAM Journal*, 2010.

[271] N. Srebro, J. Rennie, and T. Jaakkola, "Maximum-margin matrix factorization," in *Advances of Neural Information Processing Systems (NIPS)*, 2005.

[272] T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, and D. Botstein, "Imputing missing data for gene expression arrays," in *Technical Report, Stanford University*, 1999.

[273] J. Liu, P. Musialski, P. Wonka, and J. Ye, "Tensor completion for estimating missing values in visual data," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[274] J. Yi, R. Jin, A. Jain, and S. Jain, "Crowdclustering with sparse pairwise labels : A matrix completion approach," in *AAAI Technical Report*, 2012.

[275] M. Yuan and Y. Lin, "Model selection and estimation in the gaussian graphical model," in *Biometrika Journal*, 2007.

[276] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," in *Biostatistics Journal*, 2007.

[277] N. Stadler and P. Buhlmann, "Missing values: sparse inverse covariance estimation and an extension to sparse regression," in *Journal of Statistics and Computing*, 2012.

[278] R. Mazumder and T. Hastie, "Exact covariance thresholding into connected components for large-scale graphical lasso," in *Journal of Machine Learning Research (JMLR)*, 2012.

[279] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm," in *Mathematical Programming Computation*, 2010.

[280] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," in *Information Retrieval*, 2001.

[281] L. Brozovsky and V. Petricek, "Recommender system for online dating service," in *Proceedings of Conference Znalosti 2007*, 2007.

[282] V. Vapnik, "Statistical learning theory," in *Wiley-Interscience*, 1998.

[283] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," in *MIT Press*, 2006.

[284] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proceedings of the International Conference on Machine Learning (ICML)*, 1999.

[285] R. Collobert, F. Sinz, J. Weston, and L. Bottou, "Large scale transductive SVMs," in *Journal of Machine Learning Research (JMLR)*, 2006.

[286] Z. Xu, R. Jin, J. Zhu, I. King, and M. Lyu, "Efficient convex relaxation for transductive support vector machine," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[287] G. Shafer and V. Vovk, "A tutorial on conformal prediction," in *Journal of Machine Learning Research (JMLR)*, 2008.

[288] V. Vovk, "Online confidence machines are well-calibrated," in *Symposium on Foundations of Computer Science*, 2002.

[289] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.

[290] Y. Zhang, K. Huang, and C. Liu, "Fast and robust graph-based transductive learning via minimum tree cut," in *IEEE International Conference on Data Mining (ICDM)*, 2011.

[291] A. Goldberg, X. Zhu, B. Recht, J. Xu, and R. Nowak, "Transduction with matrix completion: Three birds with one stone," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[292] G. Little, S. Krishna, J. Black, and S. Panchanathan, "A methodology for evaluating robustness of face recognition algorithms with respect to changes in pose and illumination angle," in *International Conference on ACoustics, Speech and Signal Processing (ICASSP)*, 2005.

[293] B. Schuller, M. Valstar, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012 the continuous audio/visual emotion challenge," in *International Conference on Multimodal Interaction*, 2012.

[294] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002.

[295] R. Benmokhtar and B. Huet, "Perplexity-based evidential neural network classifier fusion using mpeg-7 low-level visual features," in *Proceeding of the ACM International Conference on Multimedia Information Retrieval*, 2008.

[296] A. Ross and A. Jain, "Multimodal biometrics : An overview," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2004.

[297] S. Prabhakar and A. Jain, "Decision-level fusion for fingerprint verification," in *Pattern Recognition*, 2001.

[298] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," in *Pattern Recognition*, 2005.

[299] B. Dasarathy, "Decision fusion," in *IEEE Computer Society Press*, 1994.

[300] A. Jain, R. Bolle, , and S. Pankanti, "Biometrics : Personal identification in networked society," in *Springer*, 1999.

[301] A. Jain, P. Flynn, , and A. Ross, "Handbook of biometrics," in *Springer*, 2007.

[302] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," in *IEEE Transactions on Circuits and Systems for Video Technology*, 2004.

[303] D. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision (ICCV)*, 1999.

[304] A. Torralba, "Contextual priming for object detection," in *International Journal of Computer Vision (IJCV)*, 2003.

[305] T. Strat and M. Fischler, "Context-based vision: Recognizing objects using information from both 2D and 3D imagery," in *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1991.

[306] I. Olson and M. Chun, "Perceptual constraints on implicit learning of spatial context," in *Visual Cognition*, 2002.

[307] Y. Song and T. Leung, "Context-aided human recognition - clustering," in *Europen Conference on Computer Vision (ECCV)*, 2006.

[308] A. Dey, G. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," in *Human Computer Interaction*, 2001.

[309] Y. Yang and J. Cao, "The optimization technique for solving a class of non-differentiable programming based on neural network method," in *Nonlinear Analysis: Real World Applications*, 2009.

[310] V. Vovk, A. Gammerman, and G. Shafer, "Algorithmic learning in a random world," in *Springer-Verlag*, 2005.

[311] M. Li and P. Vitanyi, "An introduction to kolmogorov complexity and its applications," in *Springer-Verlag*, 1997.

[312] C. Diehl and G. Cauwenberghs, "SVM incremental learning, adaptation and optimization," in *International Joint Conference on Neural Networks (IJCNN)*, 2003.

[313] L. Jost, "Combining significance levels from multiple experiments or analyses," in *http://www.loujost.com/statistics and physics/statsarticlesindex.html*, 2009.

[314] T. Loughin, "A systematic comparison of methods for combining p-values from independent tests," in *Computational Statistics and Data Analysis*, 2004.

[315] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, 1999.

[316] H. Wallach, S. Jensen, L. Dicker, and K. Heller, "An alternative prior process for nonparametric bayesian clustering," in *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[317] G. Yu, R. Huang, and Z. Wang, "Document clustering via dirichlet process mixture model with feature selection," in *Proceedings of ACM Conference on Knowledge Discovery from Data (KDD)*, 2010.

[318] Y. Liu and Z. Zhang, "A fast algorithm for linearly constrained quadratic programming problems with lower and upper bounds," in *International Conference on Multimedia and Information Technology*, 2008.

[319] P. Sollich and D. Saad, "Learning from queries for maximum information gain in imperfectly learnable problems," in *Advances of Neural Information Processing Systems (NIPS)*, 1995.

[320] "Microsoft malware protection center naming standards," in *http://www.microsoft.com/security/portal/Shared/MalwareNaming.aspx*.

[321] M. Seeger, "Cross-validation optimization for large scale hierarchical classification kernel methods," in *Advances of Neural Information Processing Systems (NIPS)*, 2006.

[322] J. Fürnkranz and J. Sima, "On exploiting hierarchical label structure with pairwise classifiers," in *ACM SIGKDD Explorations Newsletter*, 2010.

[323] B. Carterette, P. Bennett, D. Chickering, and S. Dumais, "Here or there: preference judgments for relevance," in *European Conference on Advances in Information Retrieval (ECIR)*, 2008.

[324] K. Chen, C. Wu, C. Chang, and C. Lei, "A crowdsourceable qoe evaluation framework for multimedia content," in *ACM Multimedia (ACM MM)*, 2009.

[325] A. Vlachos, "A stopping criterion for active learning," in *Computer, Speech and Language*, 2008.

[326] M. Bloodgood and V. Shanker, "A method for stopping active learning based on stabilizing predictions and the need for user-adjustable stopping," in *Proceedings of the Conference on Natural Language Learning*, 2009.

[327] F. Olsson and K. Tomanek, "An intrinsic stopping criterion for committee based active learning," in *Proceedings of the Conference on Natural Language Learning*, 2009.

APPENDIX A

PROOF RELATED TO DISCREPANCY MEASURE IN GENERALIZED QUERY BY TRANSDUCTION

In Chapter 8, as part of the Generalized Query by Transduction approach, we defined a matrix C which contains the absolute value of the pairwise differences between all the p-values obtained from the Conformal Predictions framework:

$$C_{ij}(P) = \left|P_i - P_j\right| \tag{9.1}$$

We claimed that this matrix C will always have exactly one positive eigenvalue (which was used as a measure of disagreement). We prove this claim in this appendix.

Lemma 7. *An N by N square matrix which has* -2 *in all its superdiagonal entries, positive constants in all entries of the last row and* 0 *in all the other positions, always has a positive determinant.*

Proof. Consider the case when $N = 2$. The matrix M_2 can be written as:

$$M_2 = \begin{bmatrix} 0 & -2 \\ d_1 & d_2 \end{bmatrix}$$

where d_1 and d_2 are positive constants. It is trivial to verify that this matrix has a positive determinant. Let us also consider the case when $N = 3$. The matrix M_3 is now given as:

$$M_3 = \begin{bmatrix} 0 & -2 & 0 \\ 0 & 0 & -2 \\ d_1 & d_2 & d_3 \end{bmatrix}$$

Again, it is easy to verify that this matrix has a positive determinant. Let us now assume that the proposition holds for some $N = n$, that is, let us assume that the following matrix M_n has a positive determinant $\det(M_n)$:

$$M_n = \begin{bmatrix} 0 & -2 & 0 & 0 & \dots & 0 \\ 0 & 0 & -2 & 0 & \dots & 0 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & d_4 & \dots & d_n \end{bmatrix}$$

Now, consider the case when $N = n+1$. The matrix M_{n+1} is given by:

$$M_{n+1} = \begin{bmatrix} 0 & -2 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & -2 & 0 & \dots & 0 & 0 \\ \vdots & & & & & & \\ d_1 & d_2 & d_3 & d_4 & \dots & d_n & d_{n+1} \end{bmatrix}$$

The determinant of M_{n+1} is computed as:

$$det(M_{n+1}) = -(-2)\begin{vmatrix} 0 & -2 & 0 & \dots & 0 & 0 \\ 0 & 0 & -2 & \dots & 0 & 0 \\ \vdots & & & & & \\ d_1 & d_3 & d_4 & \dots & d_n & d_{n+1} \end{vmatrix}$$

The $n \times n$ matrix on the right is of a similar form as M_n, and hence its determinant, say $det(\hat{M}_n)$ is greater than zero. Therefore:

$$det(M_{n+1}) = 2 \times det(\hat{M}_n) > 0$$

since the d_is are arbitrary constants. Thus, we see that if the proposition holds for $N = n$, then it also holds for $N = n+1$. Therefore, by the principle of mathematical induction, we conclude that the proposition holds for all N. This proves the lemma.

□

Lemma 8. *An N by N square matrix M where*

- $M_{NN} = 0$
- $M_{ij} = -2$ *for all* i, j *with* $i = j$ *except* $i = N$ *and* $j = N$
- $M_{iN} = 1$, *except when* $i = N$
- M_{Nj} = *a positive constant, except when* $j = N$
- 0*s in all other positions*

has a positive determinant if N is odd and a negative determinant if N is even.

Proof. Let $N = 2$. The matrix M_2 is given by:

$$M_2 = \begin{bmatrix} -2 & 1 \\ d_1 & 0 \end{bmatrix}$$

Trivially, the determinant of M_2 is negative for positive d_1. Now, consider the case when $N = 3$. The matrix M_3 is given by:

$$M_3 = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -2 & 1 \\ d_1 & d_2 & 0 \end{bmatrix}$$

It is easy to verify that the determinant of this matrix is positive for positive values of d_1 and d_2.

Let us assume that the proposition holds for $N = 2n-1$ and $N = 2n$, where n is a positive integer. Let us consider the matrix M_{2n+1}

$$M_{2n+1} = \begin{bmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n} & 0 \end{bmatrix}$$

The determinant is given by

$$det(M_{2n+1}) = (-2) \begin{vmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_2 & d_3 & d_4 & \dots & d_{2n} & 0 \end{vmatrix}$$

$$+1 \times \begin{vmatrix} 0 & -2 & 0 & \dots & 0 \\ 0 & 0 & -2 & \dots & 0 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n} \end{vmatrix}$$

The positive sign appears in front of 1 as it is in an odd position $2n+1$. The first determinant evaluates to a negative value as, by our assumption, the proposition holds for $N = 2n$ and the second determinant is positive by Lemma 7. Thus, $det(M_{2n+1})$ is positive.

Now, consider the matrix M_{2n+2}:

$$M_{2n+2} = \begin{bmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n+1} & 0 \end{bmatrix}$$

Its determinant is given as:

$$det(M_{2n+2}) = (-2)\begin{vmatrix} -2 & 0 & 0 & \dots & 0 & 1 \\ 0 & -2 & 0 & \dots & 0 & 1 \\ \vdots & & & & & \\ d_2 & d_3 & d_4 & \dots & d_{2n+1} & 0 \end{vmatrix}$$

$$-1 \times \begin{vmatrix} 0 & -2 & 0 & \dots & 0 \\ 0 & 0 & -2 & \dots & 0 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \dots & d_{2n+1} \end{vmatrix}$$

The negative sign appears in front of 1 as it is in an even position $2n+2$. The first determinant is positive since it is proved that the proposition holds for $N = 2n+1$ and the second determinant is positive by Lemma 1. Hence, $det(M_{2n+2})$ is negative. Thus, it is proved that if the proposition holds for $N = 2n-1$ and $N = 2n$, then it also holds for $N = 2n+1$ and $N = 2n+2$ and therefore, by the principle of mathematical induction, Lemma 8 holds for all N.

□

Lemma 9. *For any given set of N p-values, the matrix C has a positive determinant if N is odd and a negative determinant if N is even.*

Proof. Consider the case when $N = 3$ and let the three p-values be a, b and c. Let d_1 be the absolute difference between a and b and d_2 be the absolute difference between b and c. The matrix C_3 is given by:

$$C_3 = \begin{bmatrix} 0 & d_1 & d_1+d_2 \\ d_1 & 0 & d_2 \\ d_1+d_2 & d_2 & 0 \end{bmatrix}$$

Its determinant is given by:

$$det(C_3) = \begin{vmatrix} 0 & d_1 & d_1+d_2 \\ d_1 & 0 & d_2 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

Using the transformations Row1 = Row1 - Row2 and Row2 = Row2 - Row3, we have:

$$det(C_3) = \begin{vmatrix} -d_1 & d_1 & d_1 \\ -d_2 & -d_2 & d_2 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

$$= d_1 d_2 \begin{vmatrix} -1 & 1 & 1 \\ -1 & -1 & 1 \\ d_1+d_2 & d_2 & 0 \end{vmatrix}$$

Using the transformations Column1 = Column1 - Column2 and Column2 = Column2 - Column3, we have:

$$det(C_3) = d_1 d_2 \begin{vmatrix} -2 & 0 & 1 \\ 0 & -2 & 1 \\ d_1 & d_2 & 0 \end{vmatrix}$$

$$\Rightarrow det(C_3) > 0$$

by Lemma 8.

In general, let the N p-values be $a_1, a_2, a_3 \ldots a_N$. Let d_1 be the absolute difference between a_1 and a_2, d_2 be the absolute difference between a_2 and a_3 and so on. The determinant of the matrix C is then given by:

$$det(C) = \begin{vmatrix} 0 & d_1 & d_1+d_2 & \ldots & \sum d_i \\ d_1 & 0 & d_2 & \ldots & \sum d_i - d_1 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1+d_2) & \ldots & 0 \end{vmatrix}$$

Using the transformations Row1 = Row1-Row2, Row2 = Row2-Row3 … Row(N-1) = Row(N-1)-RowN, we get:

$$det(C) = \begin{vmatrix} -d_1 & d_1 & d_1 & \ldots & d_1 \\ -d_2 & -d_2 & d_2 & \ldots & d_2 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1+d_2) & \ldots & 0 \end{vmatrix}$$

$$= d_1 d_2 \ldots d_{N-1} \begin{vmatrix} -1 & 1 & 1 & \ldots & 1 \\ -1 & -1 & 1 & \ldots & 1 \\ \vdots & & & & \\ \sum d_i & \sum d_i - d_1 & \sum d_i - (d_1+d_2) & \ldots & 0 \end{vmatrix}$$

Using the transformations Column1 = Column1-Column2, Column2 = Column2-Column3 … Column(N-1) = Column(N-1)-ColumnN, we get:

$$det(C) = d_1 d_2 \ldots d_{N-1} \begin{vmatrix} -2 & 0 & 0 & \ldots & 1 \\ 0 & -2 & 0 & \ldots & 1 \\ \vdots & & & & \\ d_1 & d_2 & d_3 & \ldots & 0 \end{vmatrix}$$

Hence, $det(C) > 0$ if N is odd and $det(C) < 0$ if N is even, by Lemma 8. This proves Lemma 9.

□

Theorem 3. *The matrix C, which contains the absolute values of the pairwise differences between all the p-values obtained from the Conformal Predictions framework, i.e.* $C_{ij}(P) = \left|P_i - P_j\right|$*, will always have exactly one positive eigenvalue.*

Proof. Given an $n \times n$ matrix M, the characteristic polynomial of M is written as:

$$x^n - g_1 x^{n-1} + g_2 x^{n-2} - \ldots + (-1)^n g_n = 0 \tag{9.2}$$

where the coefficient g_j is the sum of the determinants of all the sub-matrices of M taken j rows and columns at a time (symmetrically). Thus, g_1 is the trace of M (i.e., the sum of the diagonal elements), g_2 is the sum of the determinants of the $\frac{n(n-1)}{2}$ sub-matrices that can be formed from M by deleting all but two rows and columns (symmetrically), and so on. Continuing in this way, we can find g_3, g_4,... up to g_n, which of course is the determinant of the entire $n \times n$ matrix. Note that the n roots of the characteristic polynomial are the eigenvalues of the matrix M.

Now, let us assume that we have a similar characteristic polynomial for the given matrix C. From Descartes' rule of signs, if the terms of a single-variable polynomial with real coefficients are ordered by descending variable exponent, then the number of

positive roots of the polynomial is either equal to the number of sign differences between consecutive nonzero coefficients, or less than it by a multiple of 2.

From Lemma 9, we know that $det(C) > 0$ if n is odd and $det(C) < 0$ if n is even. Hence, in the equation for the characteristic polynomial (Equation 9.2), it is evident that g_1 is always positive (since it is the sum of sub-matrices of C, taking 1 row and column at a time, each of whose determinant is positive). Similarly, g_2 is always negative, g_3 is always positive, and so on. Substituting these signs in Equation (9.2), we see that the characteristic polynomial for C has only one sign change between consecutive non-zero co-efficients (between the first and second terms). Thus, from Descartes' rule of signs, the matrix C always has only one positive eigenvalue (root of the characteristic polynomial). This proves the theorem.

□