Building Adaptive Computational Systems for Physiological and Biomedical Data via Transfer and Active Learning

by

Rita Chattopadhyay

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, Co-Chair
Jieping Ye, Co-Chair
Marco Santello
Baoxin Li

ARIZONA STATE UNIVERSITY

May 2013

ABSTRACT

In recent years, machine learning and data mining technologies have received growing attention in several areas such as recommendation systems, natural language processing, speech and handwriting recognition, image processing and biomedical domain. Many of these applications which deal with physiological and biomedical data require person specific or person adaptive systems. The greatest challenge in developing such systems is the subject-dependent data variations or subject-based variability in physiological and biomedical data, which leads to difference in data distributions making the task of modeling these data, using traditional machine learning algorithms, complex and challenging. As a result, despite the wide application of machine learning, efficient deployment of its principles to model real-world data is still a challenge.

This dissertation addresses the problem of subject based variability in physiological and biomedical data and proposes person adaptive prediction models based on novel transfer and active learning algorithms, an emerging field in machine learning. One of the significant contributions of this dissertation is a person adaptive method, for early detection of muscle fatigue using Surface Electromyogram signals, based on a new multi-source transfer learning algorithm. This dissertation also proposes a subject-independent algorithm for grading the progression of muscle fatigue from 0 to 1 level in a test subject, during isometric or dynamic contractions, at real-time.

Besides subject based variability, biomedical image data also varies due to variations in their imaging techniques, leading to distribution differences between the image databases. Hence a classifier learned on one database may perform poorly on the other database. Another significant contribution of this dissertation has been the design and development of an efficient biomedical image data annotation framework, based on a novel combination of transfer learning and a new batch-mode active learning method, capable of addressing the distribution differences across databases.

The methodologies developed in this dissertation are relevant and applicable to a large set of computing problems where there is a high variation of data between subjects or sources, such as face detection, pose detection and speech recognition. From a broader perspective, these frameworks can be viewed as a first step towards design of automated adaptive systems for real world data.

DEDICATION

To my daughters

Bipasa and Akanksha

and to my husband Sandip

for all their love and support

in the completion of this striving effort.

ACKNOWLEDGEMENTS

I would like to begin by expressing my deepest gratitude to Dr. Sethuraman Panchanathan and Dr. Jieping Ye, my advisors, whose experience and expertise was the 'founding stone' for the successful completion of the study presented in this dissertation. Their precious help, inspiration, encouragement and patience made the study a worthwhile and memorable learning experience.

I am thankful to Dr. Marco Santello and Dr. Baoxin Li for serving on my dissertation committee and providing guidance on this research.

I also thank my colleagues in the Center for Cognitive Ubiquitous Computing (CUbiC): John A. Black, Vineeth N Balasubramanian, Troy McDaniel, Sreekar Krishna, Gaurav Pradhan, Shayok Chakraborty, Michael J. Astrauskas, Hemanth Venkateswara, Hiranmayi Ranganathan, Ashok Venkatesan and Narayan Chatapuram Krishnan and in the Center for Evolutionary Medicine and Informatics (CEMI): Jie Wang, Zheng Wang, Cheng Pan, Qian Sun, Shuo Xiang, Rashmi Dubey, Yashu Liu, Jiayu Zhou, Lei Yuan and Sen Yang for the invigorating discussions and insights that helped me shape this work and also for the fun times that made this journey so enjoyable.

My heartfelt thanks to Dr. Ian Davidson, Dr. Wei Fan, Dr. Mark Jesunathadas and Dr. Brach Poston for their active collaboration and interest in this research study.

I wouldn't be doing justice to this opportunity of providing my acknowledgements, if I did not thank Kathleen Fretwell. I will always be grateful to Kathy and would like to thank her for all her help and support.

Lastly, I would like to acknowledge my family for their help and understanding and for continuously being a source of strength and joy for me.

This research was supported by the Science Foundation Arizona's Graduate Research Fellowship (Fall 2008- Spring 2012) and partly by the University Graduate Research Fellowship for Summer 2012. It was also sponsored by the following grants: NSF IIS-0953662, CCF-1025177, NIH LM010730, ONR N00014-11-1-0108 and 2R01 AR47301 from the National Institute of Arthritis, Musculoskeletal and Skin Diseases (NIAMSD) at the National Institutes of Health (NIH).

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

INTRODUCTION

Machine learning is a branch of artificial intelligence, which deals with identification of patterns in data and making predictions on new data based on the similarities with the identified patterns. It has wide applications in several areas such as computer vision, natural language processing, speech and handwriting recognition, image processing, recommendation systems, robotics etc. In recent years machine learning and data mining technologies have received growing attention in the biomedical domain for monitoring, detection and diagnosis of diseases. The availability of a large amount of genomic, proteomic and metabolic data in recent years, has greatly expanded the opportunities of automated data classification in biomedical domain, using data mining and machine learning techniques. In addition, the exponential growth in the design of new sensing devices has generated newer possibilities of tapping into the physiological signals of users. Physiological signals, such as skin conductance, heart rate and brain signals, provide a pathway into understanding users from novel perspectives, and have been used for applications ranging from rehabilitation technologies [21] to natural human-computer interfaces [14]. However modeling the growing amounts of physiological and biomedical data gathered across different users have posed a new challenge because such data can vary greatly between users.

This dissertation presents several novel methods based on transfer and active learning techniques, for addressing the variation of data among multiple users, enabling adaptive computational systems for physiological and biomedical data.

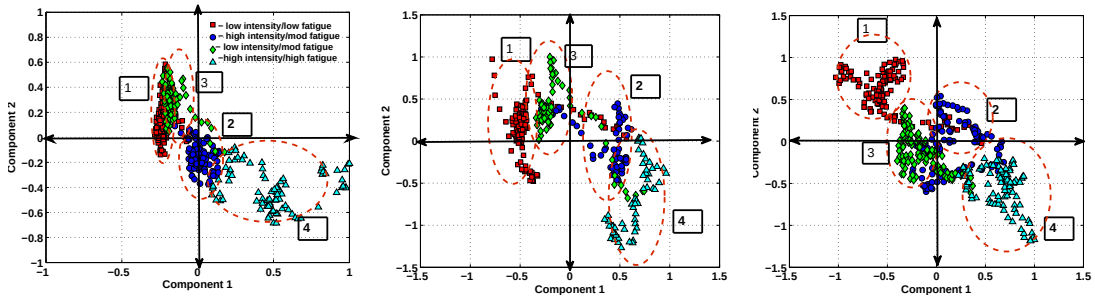
1.1 Motivation

Traditional machine learning algorithms in human-centered computing develop generalizable models from data gathered across a set of users or subjects. However high subject based variability in the physiological and biomedical data makes the task of modeling this data using traditional machine learning algorithms, complex and challenging. Consecutively, most of the works related to physiological and biomedical data are highly subject-specific in nature and a few of the generalized frameworks that are reported in literature have moderate to poor generalization across subjects [54], [47]. Subject-specific methods are based on labeled data from the specific user or subject, for training a reliable classifier. Absence of sufficient labeled data from the specific subject often limits the feasibility of these methods. Hence such methods, though ‘effective’ many not be ‘feasible’. On the other hand, generalized frameworks based on annotated

data from multiple subjects are ‘feasible’ but may not be ‘effective’, due to variations in data across subjects. One effective approach to address this challenge is to develop data models based on annotated data from multiple subjects which can adapt to a new subject, of whom very little knowledge is available; this dissertation calls such systems *person-adaptive* systems. Furthermore, the emergence of miniaturized and wearable sensing devices, enabling capture of subject-specific data in real-time, and the recent advancements in digital computing, storage and communication technologies (such as smart phones), providing fast access to data from multiple subjects, have created a great motivation for developing person adaptive systems, which can adapt themselves in real time, to form personalized or subject specific systems, providing an ‘effective’ and ‘feasible’ solution.

1.2 Problem Statement

Variation in physiological, anatomical, genomic, proteomic, metabolic or behavioral parameters across subjects causes distribution differences between the data in the training and test sets, especially when the system encounters new subjects. Under these situations, the direct application of traditional machine learning and data mining methods which assume that both these data are drawn i.i.d. from the same population, do not provide satisfactory performance [88].



(a) Subject 1 (b) Subject 2 (b) Subject 4

Figure 1.1: Three sample subjects (subjects 1, 2, 4) with four physiological stages forming four classes (shown with dotted circles) in SEMG data set: SEMG data differs predominantly in conditional probability distributions across subjects.

Activities of daily living often involve repetitive or sustained contractions. The result of such contractions is an increase in muscle fatigue, which is commonly defined as a reduction in the maximal force generating capability of the muscle [13, 20, 26]. Electromyography (EMG) is a method for biosignal recording of skeletal muscle activity. Surface Electromyography (SEMG) allows for noninvasive recording of these bio-signals. Localized muscle fatigue has been correlated with a shift in the power spectral density of SEMG signals [22, 6, 72, 51, 13]. However,

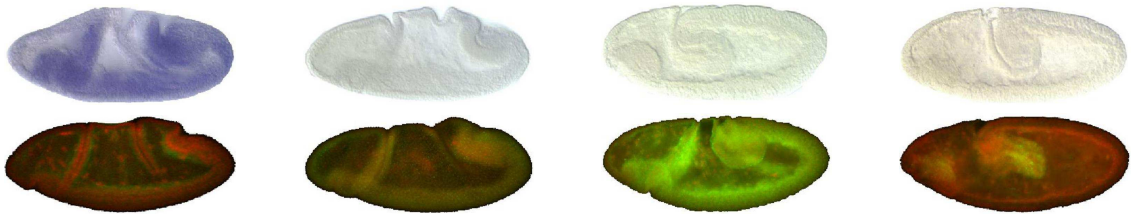


Figure 1.2: Sample *Drosophila* embryo images at different developmental stages (6-9) in BDGP (top row) and Fly-FISH (bottom row) databases.

there is often a large amount of variability in SEMG power spectrum and their shifts, across individuals [13], [26], which creates differences in data distribution. Figure 1.1 shows the distribution of the SEMG data[1] over four stages of muscle fatigue for three different subjects. It was observed that the data distribution during each stage or class varies from subject to subject. This variation leads to predominantly conditional probability differences across subjects. It was also observed that the difference in data distribution varies between subjects.

This dissertation explores machine learning and data mining methodologies to address the distribution differences between data sources, enabling person adaptive computational frameworks. Specifically, it seeks to address the following questions:

- How does subject based variability affect the probability distribution in physiological and biomedical data?
- Can the distribution differences be mitigated and knowledge be transferred across data from multiple subjects or sources?
- How to measure the distribution differences or similarities empirically between the data sets?
- Is it possible to quantify the similarity measures taking into account the mutual interaction between the data from multiple sources?
- How can the training data be used to develop an effective classifier for a test data, drawn from a different distribution?
- Can a theoretical bound be derived for classifiers learned on data with different distributions?
- Can changes in multiple parameters of myoelectric signals be tracked on a subject-independent basis?

[1] 2-dimensional embedding of 12-dimensional SEMG features obtained using factor analysis.

Besides variations due to subject based variabilities, biomedical image data may also vary due to differences in imaging techniques and resolution leading to distribution difference between the image databases. Hence a classifier learned on one database may perform poorly on the other database. Figure 1.2 shows the images of *Drosophila* embryos from two different databases, Fly-FISH [53] and BDGP [85], for the same developmental stages (6-9). This dissertation explores the development of an efficient automated biomedical image annotation framework, which is otherwise performed manually by experts, hence is time consuming and expensive. Specifically, it seeks to address the following questions:

- Can a labeled database be utilized to optimize the time and cost of annotation of a related database?
- Transfer learning addresses the problem of insufficient labeled data by transferring knowledge from an already available dataset; as a different solution, active learning methods [76] focus on selecting a small set of most informative samples from the test data, for which they acquire labels from the domain experts. Can these two methodologies be combined to address the problem of insufficient labels and to develop an efficient annotation framework?

Solutions to these questions would provide a systematic method for addressing distribution differences in real-world data, enabling efficient and wider deployment of machine learning principles for many real-world applications.

1.3 Technical Approaches

To answer the questions posed in the previous section, this dissertation explores transfer learning or domain adaptation methodologies [66], which enable transfer of knowledge under distribution differences. Typical applications of transfer learning include text classification [17], video concept detection [18], sentiment analysis [7] and WiFi Localization [64]. However, most of the domain adaptation methodologies published in literature address marginal probability differences between the data domains [80, 84, 5, 39, 65, 18]. These methodologies are not suitable to address distribution differences due to subject based variability in physiological signals, as these differences are predominantly in the conditional probability distribution. The existing methods generally combine the hypothesis generated by each of the available annotated data sources on the basis of similarities in marginal distributions, computed independently, with the new subject data. This procedure has two potential limitations, Firstly, it minimizes the loss with respect

to the source data and secondly it assumes all data sources are independent, thus it does not exploit the interaction among the multiple subject data or sources.

For addressing the conditional probability based distribution difference between the data sources, this dissertation proposes a new multi-source domain adaptation algorithm. The algorithm uses the data from multiple subjects (sources) and learns a classifier to distinguish the four classes as shown in Figure 1.1 on the basis of some labeled and unlabeled data from the test subject (or target). The key element of this algorithm is a weighting scheme that measures the similarities in conditional probabilities between the source and target domain data based on smoothness assumption on the probability distribution of the target domain data, in a joint optimization framework. Thus it takes care of the mutual interaction between the data from multiple sources while computing the similarity factors.

This dissertation also develops domain adaptation algorithms addressing both marginal and conditional probability differences based on a two stage and a single stage optimization formulation, a nonlinear feature mapping method and a multi-level similarity measurement method. Experiments for detecting stages of fatigue using SEMG data show superior performance of the proposed algorithms compared to the existing well-known domain adaptation methods. This is further confirmed on applications such as document classification and sentiment analysis using publicly available data sets.

Besides, person adaptive classification models, this dissertation proposes a subject-independent grading algorithm for myoelectic signals and an algorithm for feature selection based on robustness to subject based variability.

This dissertation also presents a theoretical analysis of the performance bound of a domain adapted classifier learned on instances from multiple sources and empirical results, supporting the theoretical bound.

To answer the question of development of an efficient annotation algorithm capable of learning on the labeled samples from one database, and developing a classifier for the other, this dissertation proposes a new batch-mode active learning algorithm and an optimization framework combining both batch-mode active learning and transfer learning in a single framework. The proposed algorithm performs transfer and active learning simultaneously, based on a single objective function, i.e., matching marginal probabilities between the training and test data. The

empirical results on biomedical image databases and on publicly available data sets, demonstrate superior performance of the proposed algorithm over the existing state-of-the-art batch-mode active learning methods.

1.4 Contributions

This dissertation proposes several new and efficient methods for developing adaptive systems for physiological and biomedical data via transfer and active learning methodologies. Specifically, it provides the following contribution to machine learning research:

First systematic approach to address subject based variability in data. This dissertation proposes a systematic approach to address subject based variability by considering the sample distribution differences, making this work innovative and significant.

A similarity measure based on conditional probability differences between the data obtained from multiple subjects or sources and the target or test subject data considering mutual interactions among the data from multiple sources.

Multi-source domain adaptation methods to address both conditional and marginal probability differences in data. This dissertation proposes multiple single and multi-source domain adaptation methods to address distribution differences between data sources, including a joint optimization framework for addressing conditional probability differences, a single optimization framework for addressing both marginal and conditional differences, a nonlinear feature map based method and a method based on similarity measure at multiple levels of granularity, i.e., at subject data, at individual class and at instance level.

Performance bound of a domain adapted classifier, learned on multiple sources. The suggested bound depends upon the difference in data distribution between the sources and test data and the degree of knowledge transferred from the multiple sources. This dissertation also presents empirical results supporting the theoretical bound.

A subject-independent method based on latent factor mapping for grading myoelectric features related to muscle fatigue, on a continuous scale from 0 to 1, during a sustained submaximal fatiguing contraction, at real time. This algorithm provides the first step to developing systems that could potentially identify the time when the muscles are approaching a level of fatigue that may cause injury, on a subject-independent basis.

An efficient batch-mode active learning method based on distribution matching principles. Reducing distribution differences between the training and test data has been previously used in the context of *transfer learning* applications, however, performing active learning on the basis of the same criterion, is a novel contribution of this work. The proposed batch-mode active learning is formulated as an efficient quadratic and an equivalent linear programming formulation, providing an alternate characterization and access to more methods for analyzing the problem.

A novel convex optimization problem for performing transfer and active learning simultaneously, based on marginal probability matching principles. To the best of my knowledge, this is the first joint framework for transfer and active learning. The efficient quadratic programming problem can be easily extended to include additional selection criterion and can be easily configured for only transfer or batch-mode active learning, with corresponding parameter changes.

1.5 Statement of Broader Impact

The difference in data across subjects exists in many other real-world applications such as face detection, pose detection, speech and handwriting recognition to name a few. The distribution difference may also arise in the case of medical diagnosis data and market predictions due to differences in data collection protocols, as in the case of biomedical images. The methodologies developed in this thesis are relevant and applicable to all these and many other computing problems where there is a high variation of data between different subjects or sources. From a broader perspective, these frameworks can be viewed as a first step towards design of automated personalized systems over traditional methods that require human intervention for customization, thus broadening the scope of applying machine learning algorithms to real-life home and work settings.

My future work will center around developing new fast and scalable data mining and machine learning methods capable of addressing the challenges of new and evolving applications in the areas of biomedical sciences, health care and business intelligence. Besides, I would also focus on presenting the theoretical analysis of frameworks based on transfer and active learning methodologies.

1.6 Structure of the Thesis

The rest of the thesis is organized as follows:

Chapter 2 starts with background information on transfer learning and domain adaptation. It also presents a short description on Electromyogram (EMG) signals and provides an insight into subject based variability in the EMG power spectrum.

Chapter 3 discusses the proposed single and multi-source domain adaptation methods based on conditional and marginal probability differences. It also presents a theoretical analysis of such frameworks besides the comparative performance of these algorithms for detecting stages of muscle fatigue using SEMG signal.

Chapter 4 discusses the proposed subject-independent grading framework, the related experiments with EMG data and their results.

Chapter 5 presents the PC based real-time system for grading fatigue and intensity level during fatiguing contractions.

Chapters 6 and 7 discuss the proposed marginal probability based batch-mode active learning method and the joint optimization framework for transfer and batch-mode active learning and present their comparative performances.

Chapter 8 reviews the related work in literature on Electromyogram signals as fatigue indicators, domain adaptation, batch-mode active learning and on existing work towards combining transfer and active learning.

Chapter 9 concludes this dissertation with a discussion on the proposed methods and directions for future research addressing the challenges of new and evolving applications.

Chapter 10 presents the journal and conference publications made during this dissertation.

Chapter 2

BACK GROUND

This chapter provides a brief history and an introduction to transfer learning, besides describing the key concepts related to domain adaptation. It also provides some insights into the Electromyogram signals, their application in detecting muscle fatigue and the challenges in modeling such signals.

2.1 A Brief History of Transfer Learning

Traditional data mining and machine learning algorithms make predictions on the future data using statistical models that are trained on previously collected labeled or unlabeled training data. Semi supervised classification addresses the problem that the labeled data may be too few to build a good classifier, by making use of a large amount of unlabeled data and a small amount of labeled data. Both these assume that the distributions of the labeled and unlabeled data are the same. Transfer learning, in contrast, allows the domains, tasks, and distributions used in training and testing to be different. In the real world, we observe many examples of transfer learning. For example, we may find that learning to recognize apples might help to recognize pears. Similarly, learning to play the electronic organ may help facilitate learning the piano. The study of Transfer learning is motivated by the fact that people can intelligently apply knowledge learned previously to solve new problems faster or with better solutions. The fundamental motivation for Transfer learning in the field of machine learning was discussed first in a NIPS-95 workshop on 'Learning to Learn', which focused on the need for lifelong machine learning methods that retain and reuse previously learned knowledge. Research on transfer learning has attracted more and more attention since 1995 in different names: learning to learn, life-long learning, knowledge transfer, inductive transfer, multi task learning, knowledge consolidation, context-sensitive learning, knowledge-based inductive bias, meta learning, and incremental/cumulative learning. Among these, a closely related learning technique to transfer learning is the multi task learning framework, which tries to learn multiple tasks simultaneously even when they are different. A typical approach for multi task learning is to uncover the common (latent) features that can benefit each individual task. In 2005, the Broad Agency Announcement (BAA) 05-29 of Defense Advanced Research Projects Agency's (DARPA) Information Processing Technology Office (IPTO)2 gave a new mission of transfer learning: the ability of a system to recognize and apply knowledge and skills learned in previous tasks to novel tasks. The first workshop on transfer learning was held in NIPS

2005 known as Inductive Transfer : 10 Years Later. In this definition, transfer learning aims to extract the knowledge from one or more source tasks and applies the knowledge to a target task. In contrast to multi task learning, rather than learning all of the source and target tasks simultaneously, transfer learning cares most about the target task. The roles of the source and target tasks are no longer symmetric in transfer learning. Traditional machine learning techniques try to learn each task from scratch, while transfer learning techniques try to transfer the knowledge from some previous tasks to a target task when the latter has fewer high-quality training data. Today, transfer learning methods appear in several top venues, most notably in data mining (ACM KDD, IEEE ICDM, and PKDD, for example), machine learning (ICML, NIPS, ECML, AAAI, and IJCAI, for example) and applications of machine learning and data mining (ACM SIGIR, WWW, and ACL, for example).

2.2 Introduction to Domain Adaptation

Domain adaptation of statistical classifiers, also known as knowledge transfer or transfer learning, is the problem that arises when the data distribution in the test domain is different from that in training domain. The need for domain adaptation is prevalent in many real-world classification problems. For example, spam filters can be trained on some public collection of spam and ham emails. But when applied to an individual person's inbox, we may want to 'personalize' the spam filter, i.e. to adapt the spam filter to fit the person's own distribution of emails in order to achieve better performance.

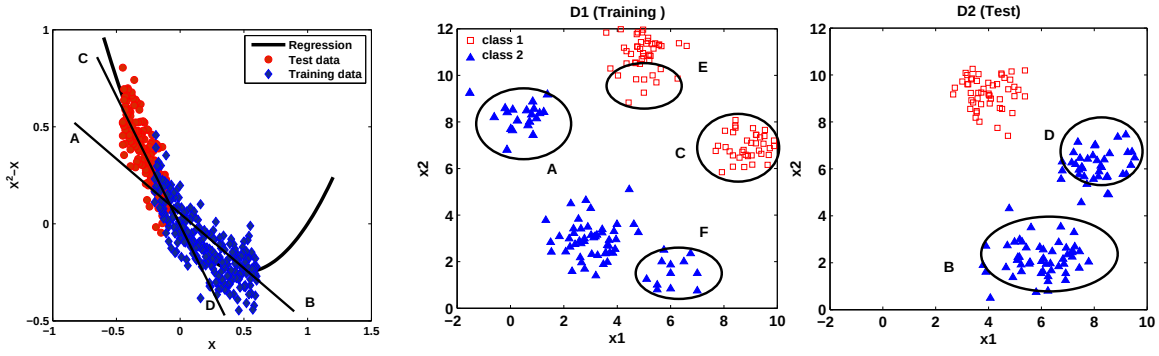
Although the domain adaptation problem is a fundamental problem in machine learning, it only started gaining much attention very recently. However, some special kinds of domain adaptation problems have been studied before under different names including class imbalance, covariate shift and sample selection bias.

A domain is defined by a set of two variables X and Y, where X denote the input variable (i.e. an observation) and Y the output variable (i.e. a class label). $P(X;Y)$ is used to denote the true underlying joint distribution of X and Y, which is unknown. The training domain where labeled data is abundant, is referred to as the source domain, and the test domain where labeled data is not available or very little, as the target domain. In domain adaptation, the joint distribution in the target domain differs from that in the source domain. $P_t(X;Y)$ is therefore used to denote the true underlying joint distribution in the target domain, and $P_s(X;Y)$ is used to denote that in the source domain. $P_t(Y)$, $P_s(Y)$, $P_t(X)$ and $P_s(X)$, denote the true marginal distributions of Y and X in the

target and the source domains, respectively. Similarly, $P_t(X|Y)$, $P_s(X|Y)$, $P_t(Y|X)$ and $P_s(Y|X)$ denote the true conditional distributions in the two domains. Furthermore, lowercase x is used to denote a specific value of X, and lowercase y to denote a specific class label. A specific x is also referred to as an observation, an unlabeled instance or simply an instance. A pair (x; y) is referred to as a labeled instance. Here, $x \in X$, where X is the input space, i.e. the set of all possible observations. Similarly, $y \in Y$, where Y is the class label set. Without any ambiguity, P(X = x; Y = y) or simply P(x; y) should refer to the joint probability of X = x and Y = y. Similarly, P(X = x) (or P(x)), P(Y = y) (or P(y)), P(X = x|Y = y) (or P(x|y)) and P(Y = y|X = x) (or P(y|x)) also refer to probabilities rather than distributions.

It is assumed that there is always a relatively large amount of labeled data available in the source domain: $D_s = \{(x_i^s; y_i^s)\}_{i=1}^{N_s}$ is used to denote this set of labeled instances in the source domain. In the target domain, it is assumed that one always has access to a large amount of unlabeled data: $D_{t,u} = \{x_i^{t,u}\}_{i=1}^{N_{t,u}}$ is used to denote this set of unlabeled instances. Sometimes, there may also be a small amount of labeled data from the target domain, which is denoted as $D_{t,l} = \{(x_i^{t,l}, y_i^{t,l})\}_{i=1}^{N_{t,l}}$. In the case when $D_{t,l}$ is not available, the problem is called as unsupervised domain adaptation, while when $D_{t,l}$ is available, the problem is called as supervised domain adaptation.

Marginal and Conditional Probability Differences



(a) (b) (c)

Figure 2.1: (a) Difference in marginal probability distribution between test and training data. ((b) and (c)) Differences in marginal and conditional probability distributions between training (D1) and test data (D2) having binary classes.

In general, distribution differences between training and test data, referred as source (s) and target domain (t) can be classified into two types based on the marginal and conditional

probabilities of the distributions. If only the marginal probability distribution of X differs between the domains i.e. for all $x \in X\, P_s(X) \neq P_t(X)$ but the conditional probability distributions remain the same i.e. $P_s(Y|X = x) = P_t(Y|X = x)$, this difference is called *covariate shift* in statistical literature [80]. However if the conditional probabilities also differ between the domains, the difference is called *functional relation change* [42]. A typical data distribution with covariate shift (difference in marginal probability distribution alone) is shown in Figure 2.1 (a). Here, the dots with different shapes, represent the target and source domain data respectively. The data points from both distributions in this example have been generated using the equation $Y = X^2 - X$, i.e. their conditional probabilities $P(Y|X)$ are the same, but their marginal probabilities are different as obvious from Figure 2.1 (a). The degree one regression line AB drawn on the basis of training data does not approximate the test data (line CD approximates the test data). In contrast, Figure 2.1(b and c) presents two data distributions D1 and D2 with functional relational change (both marginal and conditional probability distribution differences). The points in circle A differ in marginal probability from points in B and D, whereas points in C and D have conditional probability differences. Note that this data represents a binary classification problem.

2.3 Electromyogram (EMG) Signal

EMG stands for electromyography. It is the study of muscle electrical signals. EMG is sometimes referred to as myoelectric activity.EMG is measured using similar techniques to that used for measuring EKG, EEG or other electrophysiological signals. Electrodes are placed on the skin overlying the muscle, signals so collected are called Surface electromyogram (SEMG). Alternatively, wire or needle electrodes are used and these can be placed directly in the muscle, and the signal collected in this manner are called needle electromyogram signals (EMG). When EMG is acquired from electrodes mounted directly on the skin, the signal is a composite of all the muscle fiber action potentials occurring in the muscle(s) underlying the skin. These action potentials occur at somewhat random intervals so at any one moment, the EMG signal may be either positive or negative voltage. Individual muscle fiber action potentials are sometimes acquired using wire or needle electrodes placed directly in the muscle.

There are many, many applications for the use of EMG. EMG is used clinically for the diagnosis of neurological and neuromuscular problems. It is used diagnostically by gait laboratories and by clinicians trained in the use of biofeedback or ergonomic assessment. EMG is also used in many types of research laboratories, including those involved in biomechanics, motor

control, neuromuscular physiology, movement disorders, postural control, physical therapy, and many others.

The EMG signal is typically described using a variable related to the amplitude of the signal. Rectified, averaged EMG, integrated EMG, and linear envelope displays are all ways to display the amplitude of the EMG signal. Frequency analysis include analysis of zero crossings, spectral analysis, numerous time-frequency algorithms, and many other techniques.

Changes in EMG with Muscle Fatigue

MUSCLE FATIGUE is a condition when the ability of the muscle to contract and produce force is reduced. Localized muscle fatigue is the situation when a muscle or a group of muscles has a reduced ability to contract and produce force, generally as a result of prolonged, relatively strong muscle activity. Muscle fatigue threshold cannot be defined as a simple function of muscle load magnitude and timing because muscle characteristics and capabilities vary from person to person. Undetected fatigue can cause injury-often irreversible-to the subject and besides the pain and suffering, is a financial burden to industry and society.

There are number of techniques that can be used to objectively determine the level of fatigue in a subject. The most reliable of these is the direct measurement of chemical properties in the muscle of the subject. Since this is an invasive technique it is inappropriate for routine utilization, away from the clinical environment.

The properties of the EMG signal are related to the biochemical and physiological changes in skeletal musculature during fatiguing contractions. One consequence of the muscle contraction is the increase in the concentration of lactic acid, a metabolic product. Increased concentration of the lactates is responsible for fatigue by changes in intracellular pH. As a result, muscle fiber conduction velocity (CV) decreases, directly changing the shape of the motor unit action potential (MUAP) waveform and, finally, the properties of the surface EMG as an interference signal of all generated MUAPs. Brody et al. [67] show (in vitro) that the decrease of pH in fact determines the decrease of muscle fiber CV, and as a consequence, the decrease of median frequency (MDF). Therefore, the lowering of muscle fiber conduction velocity (CV) is one of the causes of signal power spectrum shift toward lower frequencies, and also of the increase in SEMG signal amplitude because of a spatial low-pass filtering effect of tissue as a volume conductor [13].

Subject based Variability in EMG Power Spectrum

Several parametric measures of SEMG signal have been used as a relative indicator of the muscle fatigue phenomenon for an individual subject. These include the root mean square (rms), instantaneous frequency, zero crossing rate, mean-frequency, and median-frequency. In general, there is a large variance in both the SEMG power spectrum itself and the nature of the power spectrum shift as a result of fatigue onset for different subjects [13, 26]. These generally unpredictable variations mean that the SEMG is difficult to model and the task of automating the process of signal classification as a generalized tool would be complex, as it would need to adapt to the SEMG power spectral density trends for each individual subject. In addition to these difficulties, muscle properties associated with the production of force are time variant. Motor unit recruitment is dependent both on the load as well as the current fatigue status of the muscle itself, thus introducing a time dependency in the SEMG signal as muscle loading progresses. There is also a random variation in the effectiveness of the body tissues in conducting the signal from the muscle to the surface electrodes. Averaging features over consecutive windows is one way to address the random variations in the EMG signal [48].

Features of EMG / SEMG Signal

For automatic classification of the EMG / SEMG data as a function of muscle force and muscle fatigue in a generalized sense it is important first to reduce the complexity of the signal data, to extract only those features that correlate best with the force of contraction of the muscle and muscle fatigue and at the same time adaptively track the time variation of the properties of the muscle as well as the rate of stimulation. Both time and frequency domain approaches (and a combination of the two) have been attempted in the past. Tools including rectification of the signal, "integration" of the signal, zero crossing count, fast Fourier transform (FFT), and short time Fourier transform (STFT) can provide a basis. Rectification of the signal is a convenient tool to measure the "relative strength" of the signal and "relative strength of contraction" of the muscle. RMS correlates well with the strength of total muscle contraction [15]. FFT, STFT, DFT provide the SEMG spectrum and information related to muscle fatigue status, size of motor units, synchronous activity between motor units, and rate of stimulation of the muscle.

Chapter 3

PERSON-ADAPTIVE DOMAIN ADAPTATION FRAMEWORKS

This dissertation develops multi-source as well as single source domain adaptation algorithms to address distribution difference due to subject based variability in electromyogram signal, enabling person-adaptive classification frameworks to classify physiological status of muscle with respect to fatigue and intensity of activity. Furthermore, it also develops a feature selection method based on subject based variabiity.

3.1 Conditional Probability based Multi-source Domain Adaptation (CP-MDA)

In the proposed framework named as *Conditional Probability based Multi-source Domain Adaptation* (CP-MDA) the unlabeled data are labeled using a weighting scheme that measures the similarities in conditional probabilities between the source and target domain data; the key of this proposed weighting scheme is a joint optimization framework based on smoothness assumption on the probability distribution of the target domain data.

3.1.1 Overview

The proposed framework learns a classifier f^T for the target domain data, using a few labeled samples and a large number of unlabeled samples from the target domain. The framework is based on a novel weighting scheme that integrates multiple source domain data using a set of weights, one for each source domain. These weights are used to compute the labels of the unlabeled target domain data, called 'pseudo labels'. The target domain prediction model is then learned from both labeled and pseudo labeled target domain samples in a regularized framework. Specifically, the proposed multi-source domain adaptation framework is given as follows:

$$\min_{f^T \in H_K} \gamma_A \|f^T\|_K^2 + \frac{1}{n_l} \sum_{i=1}^{n_l} V(x_i, y_i, f^T) + \Omega_r(f_u^T) + \Omega_m(f^T) \tag{3.1}$$

The first term controls the complexity of the classifier f^T in the Reproducing Kernel Hilbert Space (RKHS) H_K, γ_A controls the penalty factor, the second term is the empirical error of the target classifier f^T on the few labeled target domain data D_l^T, and n_l is the number of labeled target domain data. The empirical error on the unlabeled target data, labeled using a conditional probability based weighting scheme, forms the third term. This regularizer enforces the target classifier f^T to have similar decision values to the auxiliary source which has similar conditional probability distribution, explained in detail in Subsection 3.1.2. The fourth term is a manifold based regularizer based on the smoothness assumption [3] on target domain data: if two points x_i and x_j

are close to each other in the intrinsic geometry of marginal distribution then they are most likely to have similar conditional probabilities, i.e., $f^T(x_i)$ should be similar to $f^T(x_j)$. The *manifold based regularizer* is defined as in [3]:

$$\Omega_m(f^T) = \frac{\gamma_I}{n_T^2} f^{T'} L f^T. \tag{3.2}$$

where L is the graph Laplacian matrix constructed on D^T, $f^T = [f^T(x_1), \cdots, f^T(x_{n_T})]$, γ_I controls the complexity of the function f^T in the intrinsic geometry of the marginal probability of x and the normalizing coefficient $\frac{1}{n_T^2}$ is the natural scale factor for the empirical estimate of the Laplace operator, and the symbol $'$ is used to represent the matrix or vector *transpose* operation.

3.1.2 Multi-Source Weighting

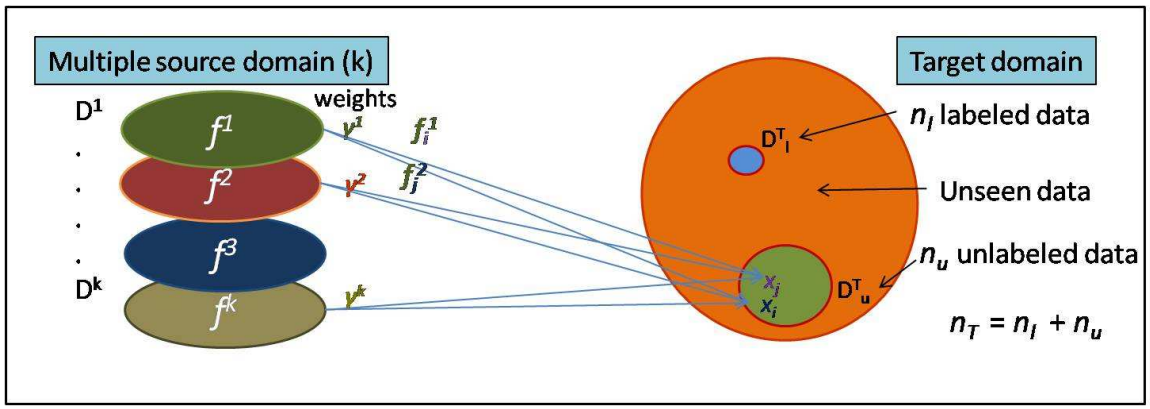


Figure 3.1: Conditional Probability based Multi-Source Weighting

Let $f_u^T = [f_{n_{l+1}}^T \cdots f_{n_T}^T]'$ be the decision values of the target classifier f^T for the unlabeled target domain data and let $f_u^s = [f_{n_{l+1}}^s \cdots f_{n_T}^s]'$ be the decision values of the s-th auxiliary classifier for the same unlabeled target domain data. Let β^s be the measure of relevance or similarity between the distributions of the s-th source and the target data, and let $f_j^T = f^T(x_j)$ be the decision value of target classifier on the target domain data x_j and $f_j^s = f^s(x_j)$ be the decision value of the s-th auxiliary source classifier on x_j. The proposed algorithm uses a weighted combination of the k source domain classifiers f^s to estimate the target classifier. Specifically, the estimated label ($\hat{y}_j$) of the unlabeled target data x_j based on the k source domain classifiers f^s is given by

$$\hat{y}_j = \sum_{s=1}^{k} \beta^s f_j^s, \tag{3.3}$$

where $\beta^s > 0$ is the weight for the s-th source. It is assumed that the weights are normalized, that is, $\sum_s \beta^s = 1$. The auxiliary classifier f^s for the s-th source is pre-computed based on its

respective data. The auxiliary classifiers f^s and the target classifier f^T can be trained using different kernels or even different learning methods. The resulting regularizer, $\Omega_r(f_u^T)$, named as *relevance based regularizer* measures the difference between the target classifier decision value and the estimation based on multiple source data, and is defined as follows:

$$\Omega_r(f_u^T) = \frac{\theta}{2} \sum_{j=n_l+1}^{n_T} \|f_j^T - \sum_{s=1}^{k} \beta^s f_j^s\|^2, \tag{3.4}$$

where $\theta > 0$ is a constant. θ is used to control the relative importance of true labels and pseudo labels. The weight β^s which provides a measure of relevance between the s-th auxiliary source domain and the target domain is computed on the basis of a *Conditional Probability based Weighting Scheme*, which evaluates the similarities in distributions between the source and target domains predominantly based on conditional probability differences.

Next, it is shown, how to estimate the weights β^s's. The proposed weighting scheme evaluates the similarities between auxiliary source data and the target domain data considering the similarities in their conditional probabilities.

Let $F_i^S = [f_i^1 \cdots f_i^k]$ be the $1 \times k$ vector of predicted labels of k auxiliary source models for the i-th sample of target domain data. Let $\beta = [\beta_1 \cdots \beta_k]^T$ be the $k \times 1$ weight vector, where β^s is the weight corresponding to the s-th auxiliary source. Following (3.3), the predicted label for the i-th sample of target domain data is

$$\hat{y}_i = \sum_{s=1}^{k} \beta^s f_i^s = F_i^S \beta. \tag{3.5}$$

This motivates one to estimate the weight vector β based on the smoothness assumption on the conditional probability distribution: the optimal weight vector β is computed by minimizing the difference in predicted labels between two nearby points in the target domain. Specifically, the proposed weighting framework solves the following problem:

$$\min_{\beta:\beta' e=1,\beta\geq 0} \sum_{i,j=n_l+1}^{n_l+n_u} (F_i^S\beta - F_j^S\beta)^2 W_{ij} \tag{3.6}$$

where $F_i^S\beta$ and $F_j^S\beta$ are the predicted labels for i-th and j-th samples of target domain data and W_{ij} is the edge weight between the i-th and j-th samples given by $e^{\frac{-(x_i-x_j)^2}{2\sigma^2}}$. The minimization problem is re-written as follows:

$$\min_{\beta:\beta' e=1,\beta\geq 0} \beta' \left(F^S\right)' L_u F^S \beta \tag{3.7}$$

where F^S is an $n_u \times k$ matrix with each row of F^S being the $1 \times k$ vector of k predicted labels for a sample of target domain data and L_u is normalized graph Laplacian associated with the target domain data D_u^T, given by $L_u = I - D^{-0.5} W D^{-0.5}$, where I is the identity matrix of size n_u, W is the adjacency graph defining edge weights between the n_u unlabled samples in the target domain data, and D is the diagonal matrix given by $D_{ii} = \sum_{j=1}^{n_u} W_{ij}$.

The minimization problem in (3.7) is a standard quadratic problem (QP) and can be solved by applying many existing QP solvers. This dissertation uses the 'quadprog' function in MATLAB. With the computed weights, the labels for the unlabeled target domain data, called *pseudo labels*, are computed using (3.3), and are substituted into the regulariser in (3.4).

Intuitively, by enforcing that nearby points in the marginal distribution of the target data have similar class labels (or conditional probability) via the optimization in (3.7), the proposed weighting scheme is likely to give higher weights to those sources with the conditional probability distribution similar to the target data. This is verified in our empirical study on both SEMG and synthetic data. If a source has a conflicting conditional distribution as the target, it is likely to get a low or even zero weight. In addition, different from many existing weight schemes which compute the weights by considering each source independently, the proposed weighting scheme computes the optimal value of β or the optimal weights of all the k sources simultaneously, thus taking the potential interaction among multiple subjects in the source domain into account.

3.1.3 Multi-source Domain Adaptation

Using the least square error and substituting the regularizers one can rewrite (3.1) as follows:

$$\min_{f^T \in H_K} \gamma_A \|f^T\|_K^2 + \frac{1}{n_l} \sum_{i=1}^{n_l} (f_i^T - y_i^T)^2$$
$$+ \frac{\theta}{2} \sum_{j=n_l+1}^{n_T} \|f_j^T - \sum_{s=1}^{k} \beta^s f_j^s\|^2 + \frac{\gamma_I}{n_T^2} f^{T'} L f^T \tag{3.8}$$

By the Representer theorem [74], one can find an optimal solution of (3.8), which is a linear expansion of the kernel function K, over both the labeled D_l^T and the pseudo labeled target domain data D_u^T given as follows:

$$f^T(x) = \sum_{i=1}^{n_l+n_u} \alpha_i K(x_i, x). \tag{3.9}$$

Substituting this into (3.8), one can obtain the optimal $\alpha = [\alpha_1 \cdots\cdots\cdots \alpha_{n_l+n_u}]^T$ by solving the following optimization problem:

$$\min_{\alpha} \frac{1}{n_l + \theta n_u}(Y - K\alpha)^{'} J(Y - K\alpha) + \gamma_A \alpha^{'} K\alpha + \frac{\gamma_I}{(n_u + n_l)^2}\alpha^{'} KLK\alpha \tag{3.10}$$

where K is the $(n_l + n_u) \times (n_l + n_u)$ kernel Gram matrix over the target domain data, Y is the label vector over labeled and pseudo labeled target domain data points given by:

$$\left[y_1 \cdots y_{n_l} \sum_s \beta^s_{(n_l+1)} f^s_{(n_l+1)} \cdots \sum_s \beta^s_{(n_l+n_u)} f^s_{(n_l+n_u)} \right] \tag{3.11}$$

L is the graph Laplacian defined over labeled and pseudo labeled target domain data, and J is a diagonal matrix of size $(n_l + n_u) \times (n_l + n_u)$ given by $J = diag(1, \cdots 1, \theta, \cdots \theta)$ with the first n_l diagonal entries as 1 and the rest as θ. θ is assigned a number between 0 and 1, thus the pseudo labels of the target domain data get smaller weights compared to the labels of the labeled target domain data. From (3.10), the optimal α^* is given by:

$$\alpha^* = \left(JK + \gamma_A(n_l + \theta n_u)I + \frac{\gamma_I(n_l + \theta n_u)}{(n_u + n_l)^2} LK \right)^{-1} JY.$$

With the computed α^*, the prediction of any unseen test data x is given by:

$$f^T(x) = \sum_{i=1}^{n_l+n_u} \alpha^*_i K(x_i, x). \tag{3.12}$$

Since the proposed domain adaptation framework is based on multiple sources whose similarities to target domain data or weights are computed based on a conditional probability based weighting scheme, this dissertation refers the framework as *Conditional Probability based Multi-Source Domain Adaptation (CP-MDA)*.

3.1.4 Experiments and Results

The proposed algorithms have been evaluated on multi-dimensional feature vectors extracted from SEMG (Surface electromyogram) signals collected from 8 subjects during a fatiguing exercise.

SEMG Data

The SEMG data was collected during a repetitive gripping action performed by the forearm. Figure 3.2 shows the subject with surface EMG differential electrodes on the extensor carpi radialis muscle to record the SEMG signal. The subject performs a cycle of flexion-extension of forearm as shown in Figure 3.2 at two different speeds, i.e., low speed (1 cycles/sec) and high speed (2

cycles/sec) repetitively for about 4 minutes. The cycles of low and high speed are alternated after every minute to form four phases (or classes) as discussed in the introduction.

The raw SEMG activity was recorded by Grass Model 8-16C at 1000Hz and passed through a band pass filter of 20Hz to 500Hz. The data was collected and saved by the LabView software (from National Instruments) running on a PC. Data of the order of 1.92 Million samples ($1000 \times 4 \times 60 \times 8$), was collected from 8 subjects including male and female of the age group of 25 years to 45 years. A set of twelve amplitude and frequency domain features including mean frequency, median frequency, spectral energy, spectral entropy, root mean square, number of zero crossings, to mention a few are derived from running windows of 1000 time samples with 50% overlap [48].

Each subject data consists of around 280 to 400 samples of 12 dimensional feature vectors, belonging to four classes with around 70 to 100 samples per class (some subjects who got fatigued sooner and hence could not maintain the required uniform speed for 1 minute the time period was reduced to 30 to 45 secs per phase, hence the number of samples varies between different subjects).

Experimental Procedure

To evaluate the effectiveness of the proposed methods, this dissertation compares the results with four baseline methods, including SVM-C, SVM-M, SMA, and TSVM (Transductive SVM), and two recently proposed multi-source learning methods, including Locally Weighted Ensemble (LWE) [24] and Domain Adaptation Machine (DAM) [18].

SVM-C refers to *all but one* method where the training data comprises of data from seven subjects and the test data is the data from the remaining subject. SVM-M, refers to the *majority voting* based ensemble framework. The class y assigned to each unlabeled test data x is $\max_y NV(y|x)$ where $NV(y|x)$ is the number of votes given for class y for a particular test sample x by the seven auxiliary sources. SMA refers to *simple model averaging*, which provides equal weight to all the classifiers learned on each auxiliary source domain in an weighted ensemble framework used to generate the label for the target domain data. TSVM refers to Transductive SVM [1] implemented in the svmlight package. It is a semi-supervised method where the training data consists of labeled data from all seven subjects from the source domain and unlabeled data from the target subject.

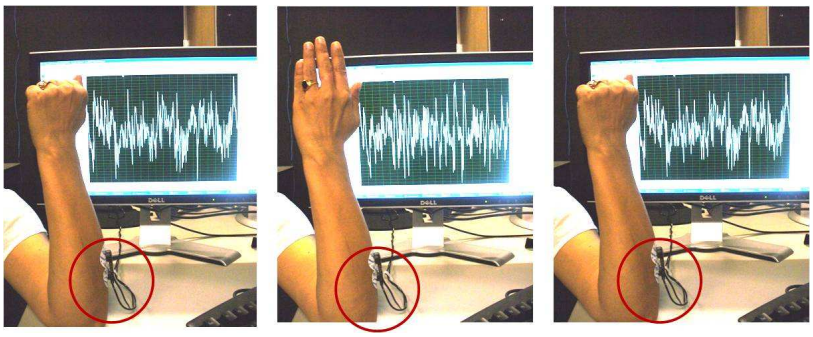


Figure 3.2: SEMG data collection during a repetitive gripping activity

Table 3.1: Comparative performance of CP-MDA on SEMG data - Accuracy (%)

Test Sub	SVM-C	SVM-M	SMA	TSVM	LWE	DAM(1)	DAM(7)	CP-MDA(1)	CP-MDA(7)
1	70.76	33.9	44.96	49.09	67.44	74.83	77.43	81.93	85.25
2	43.69	50.76	44.61	55.68	77.54	81.36	83.35	84.73	87.7
3	50.11	56.85	56.84	65.09	75.55	74.77	78.99	82.45	85.06
4	59.65	47.93	49.67	56.98	81.22	80.63	84.32	81.27	86.4
5	40.37	44.79	50.15	62.5	52.48	76.74	81.14	80.74	86.62
6	59.21	61.45	60.33	71.32	65.77	59.21	74.28	83.12	88.09
7	47.13	46.91	45.76	60.73	60.32	74.27	83.31	81.57	86.4
8	69.85	64.53	74.46	68.55	72.81	84.55	86.6	88.5	90.56
Average	55.09	50.85	53.34	61.24	71.14	75.79	81.18	**83.04**	**87.01**

The number of labeled samples per class in the target domain is varied. DAM(1) and DAM(7) refer to the DAM framework with 1 and 7 target domain labeled samples per class respectively. The proposed CP-MDA method is also implemented using 1 and 7 labeled data from target domain, referred as CP-MDA(1) and CP-MDA(7) respectively. For both cases the unlabeled data from the target domain, is fixed at 10% of the target domain data. The weights of the auxiliary sources computed by the proposed method are also based on this 10% unlabeled target domain data. The rest of the target domain data is treated as unseen target domain data. All the methods are tested on the same pool of *unseen* unlabeled target domain data. The accuracies are computed in a subject independent manner.

Some of the parameters used in implementing the existing and the proposed methods is mentioned here briefly . The values of γ_A and γ_I were kept as 0.014 and 0.01 respectively, as suggested in [3]. The Laplacian graph matrix used in calculating the weights was set as 'binary' type based on the N nearest neighbors with $N = 10$. The value of N was chosen based on 5-fold cross validation over a range of values: $[4, 6, 10, 12, 15,]$ on the labeled test data. The value of Θ was estimated via 5-fold cross validation on the set $\{i10^{-2} | i = 0, 1, \cdots, 100\}$.

Table 3.2: Comparison of SVM-T, DAM, and CP-MDA on Subject 6 (top) and Subject 7 (bottom) in terms of accuracy (%) when the number of labeled target domain data per class varies.

	Number of labeled data per class					
Method	1	2	3	4	6	7
SVM-T	4.26	4.26	49.09	73.63	84.69	85.67
DAM	59.12	59.21	59.35	59.59	65.03	74.28
CP-MDA	**83.12**	83.12	85.45	87.58	87.77	88.09
SVM-T	10.59	45.5	77.79	83.25	85.48	84.97
DAM	74.27	75.11	79.10	81.19	82.43	83.3
CP-MDA	**81.57**	83.99	85.81	86.24	86.32	86.4

Comparative Performance of CP-MDA

This dissertation compares different methods including SVM-C, SVM-M, SMA, TSVM, LWE, DAM and the proposed CP-MDA. The results are summarized in Table 3.10. The first column of the table indicates the subject data under test (target domain). The training data (source domain) consists of the data from the remaining seven subjects. Similar to the results obtained in the case of synthetic data, SVM-C, SVM-M, SMA, and TSVM perform very poorly. Significant improvement in classification accuracy is observed when domain adaptation methodologies are employed. The proposed method CP-MDA(1) provides a 20% to 30% improvement over the baseline methods including SVM-C, SVM-M, SMA and TSVM. The classification accuracies of the proposed method are in average 13% higher than LWE. It is also observed as in the case of synthetic data that CP-MDA(1) performs not only better than DAM(1) but also better than DAM(7) in 5 out of 8 cases. These results verify the effectiveness of the proposed method.

Next, the performance of CP-MDA when the number of labeled target domain data varies, is evaluated. CP-MDA is compared with DAM and SVM-T. SVM-T refers to an SVM classifier trained on the labeled target domain data. The results for two subjects are summarized in Table 3.2; similar results are obtained for the other six subjects and the results are omitted. One can observe from the table that when the number of labeled target domain data per class is small, e.g., 1 to 4 samples per class, both domain adaptation methods perform much better than SVM-T. But with an increasing number of labeled data from the target domain the accuracies become comparable. However the proposed method always performs better than the other two methods. This result demonstrates that domain adaption is especially useful when the amount of labeled target domain data is small.

This dissertation also compares the performance of the weighting schemes used in LWE, DAM and CP-MDA. Table 3.3 summarizes the results for different test cases. One can observe that CP-MDA-WE performs better than the other methods in 6 out of 8 cases, and LWE performs better in the remaining 2 cases. Recall that like CP-MDA-WE, LWE computes weights for the auxiliary source domain based on the conditional probability differences between the source and target domains, while MMD-WE computes the weights based on the marginal probability differences only. Since SEMG data has significant conditional probability differences, CP-MDA-WE and LWE are expected to outperform DAM-WE.

Table 3.3: Comparison of different weighting schemes for different test subjects - Accuracy (%).

Test Sub	LWE	MMD-WE	CP-MDA-WE
1	67.44	68.27	75.12
2	77.54	69.48	83.23
3	75.55	71.84	75.68
4	81.22	62.65	81.09
5	52.48	68.32	78.16
6	65.77	58.91	76.11
7	60.32	67.75	75.07
8	72.81	66.11	78.71
Average	69.14%	66.66%	**77.89**%

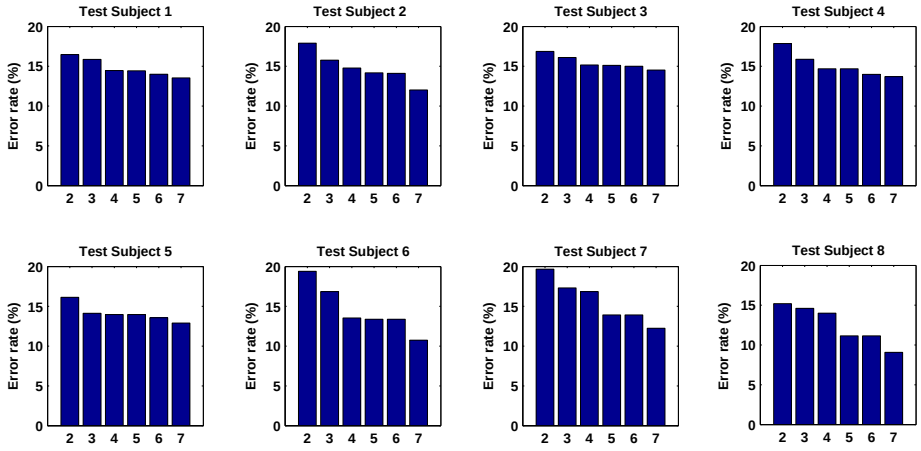


Figure 3.3: The effect of the number of auxiliary source domains (horizontal axis) in the training set on the proposed CP-MDA algorithm in terms of the classification error rates (%) for all eight subjects.

The proposed algorithm CP-MDA computes weights for each class for each of the auxiliary source domain data, thus exploiting the similarities and dissimilarities at the class level. Table 3.4 shows the weights for four different classes assigned to each training subject in the

source domain for test subject 1 in the target domain. One can observe that the proposed weighting scheme assigns different weights to different auxiliary source domain data (subject data) for different classes. Subjects 5 and 8 get higher weights for class 1, subject 7 gets a higher weight for class 2, subject 5 gets a higher weight for class 3, and for class 4 subject 4 gets a higher weight. One observes from Figure 1.1 that the data distribution of class 4 of subject 1 is very similar to that of class 4 of subject 4.

Table 3.4: Weights computed by CP-MDA for four different classes for each of the source domain subjects 2-8 for test target subject 1.

Class	Target subject						
	2	3	4	5	6	7	8
1	0	0	0.02	0.50	0.48	0	0
2	0	0.01	0.03	0	0.11	0.74	0.11
3	0	0.02	0.12	0.75	0	0.01	0.11
4	0.09	0.02	0.66	0.11	0.11	0.01	0

;

Table 3.5: Weights computed by CP-MDA for four different classes for each of the source domain subjects 1-7 for test target subject 8.

Class	Target subject						
	1	2	3	4	5	6	7
1	0	0.22	0.09	0.1	0.31	0.12	.16
2	0	0	0	0	0.01	0.01	0.99
3	0.053	0	0.43	0.01	0.40	0.03	0.08
4	0	0	0	0	0.91	0.04	0.06

One of the key advantage of the proposed algorithm is that it exploits the information from multiple source domains for classifying the target data. It will be interesting to study how the number of sources used in the training set affects the classification. Figure 3.3 presents the error rates obtained for each test subject when the number of subjects in the source domain varies (from 1 to 7); Subjects are added to the training set in increasing order of the subject number. One can observe that for all subjects, the error rate decreases monotonically when the number of subjects increases. These results demonstrate the effectiveness of the proposed algorithm for extracting useful information from multiple sources.

To evaluate the benefit of a multi-source domain adaption framework for addressing subject based variability, the proposed algorithm is compared with three representative single-source domain adaption algorithms Table 3.6 summarizes the classification accuracies obtained by different methods for each of the test subjects. The target data is from the subject shown in column 1 and the source data consists of the combined data from the remaining seven

Table 3.6: Comparison of CP-MDA with three single source domain adaptation algorithms (KMM, TCA, and KE) - Accuracy(%).

Test Sub	CP-MDA(7)	KMM	TCA	KE
1	85.25	65.15	45.15	71.85
2	87.7	46.96	68.93	74.62
3	85.06	59.55	56.78	74.79
4	86.4	73.38	52.68	69.35
5	86.62	45.31	60.15	73.44
6	88.09	70.62	76.92	83.92
7	86.4	51.13	55.64	77.97
8	90.56	42.79	67.24	79.48
Average	**87.01**	56.86	53.84	75.67

subject. Classification results were averaged over 10 runs with different sets of randomly selected 7 labeled samples per class from the target domain data. One can observe from the table that combining all the subject data and forming a single domain degrades the performance. One also observes that among the three single domain adaption algorithms, KMM [39] or TCA [64] which consider the marginal probability differences only perform worse than KE [95]. These results are expected as SEMG data has significant conditional probability differences. These results demonstrate the effectiveness of the proposed multi-domain framework for dealing with subject based variability in SEMG data.

3.2 Two Stage Weighting Framework for Multi-source Domain Adaptation (2SW-MDA)

This section develops a two-stage domain adaptation methodology which combines weighted data from multiple sources based on marginal probability differences (first stage) as well as conditional probability differences (second stage), with the target domain data. The weights for minimizing the marginal probability differences are estimated independently, while the weights for minimizing conditional probability differences are computed simultaneously by exploiting the potential interaction among multiple sources. This section also provides a theoretical analysis on the generalization performance of the proposed multi-source domain adaptation formulation using the weighted Rademacher complexity measure. Empirical comparisons with existing state-of-the-art domain adaptation methods using three real-world datasets demonstrate the effectiveness of the developed approach.

3.2.1 Overview

The multi-source domain adaptation framework, named as *Two Stage Weighting Framework for Multi-source Domain Adaptation* (2SW-MDA) computes weights for the data samples from multiple sources to reduce both marginal and conditional probability differences between the source and target domains. In the first stage, the framework computes weights of the source domain data samples to reduce the marginal probability differences, using Maximum Mean Discrepancy (MMD) [8, 39] as the measure. The second stage computes the weights of multiple sources to reduce the conditional probability differences; the computation is based on the smoothness assumption on the conditional probability distribution of the target domain data. Finally, a target classifier is learned on the re-weighted source domain data. A novel feature of the weighting methodologies is that no labeled data is needed from the target domain, thus widening the scope of their applicability. The framework is easily extendable to the case where a few labeled data may be available from the target domain.

3.2.2 2-Stage Source Instance Weighting

The developed framework consists of two stages. The first stage computes the weights of source domain data based on the marginal probability difference; the second stage computes the weights of source domains based on the conditional probability difference, as described in Section 3.1.2. A target domain classifier is learned on these re-weighted data.

Re-weighting Data Samples based on Marginal Probability Differences

The difference between the means of two distributions after mapping onto a reproducing kernel Hilbert space, called Maximum Mean Discrepancy, has been shown to be an effective measure of the differences in their marginal probability distributions [8]. This measure is used to compute the weights α_i^s's of the s-th source domain data by solving the following optimization problem [39]:

$$\min_{\alpha^s} \quad \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \alpha_i^s \Phi(x_i^s) - \frac{1}{n_T} \sum_{i=1}^{n_T} \Phi(x_i^T) \right\|_H^2 \tag{3.13}$$

$$\text{s.t.} \quad \alpha_i^s \geq 0$$

where $\Phi(x)$ is a feature map onto a reproducing kernel Hilbert space H [82], n_s is the number of samples in the s-th source domain, n_T is the number of samples in the target domain, and α^s is the n_s dimensional weight vector. The minimization problem is a standard quadratic problem and can be solved by applying many existing solvers. This thesis uses the 'quadprog' function in MATLAB.

Re-weighting Sources based on Conditional Probability Differences

In the second stage the proposed framework modulates the α^s weights of a source domain s obtained on the basis of marginal probability differences in the first stage, with the weighting factor β^s computed as described in Section 3.1.2. The weight β^s reflects the similarity of a particular source domain s to the target domain with respect to conditional probability distributions.

To illustrate the two-stage framework, the effect of re-weighting data samples in source domains D1 and D2 of the toy dataset (shown in Figure 3.4), based on the computed weights is demonstrated in Appendix A. Figure 3.4 shows two source distributions, along with their

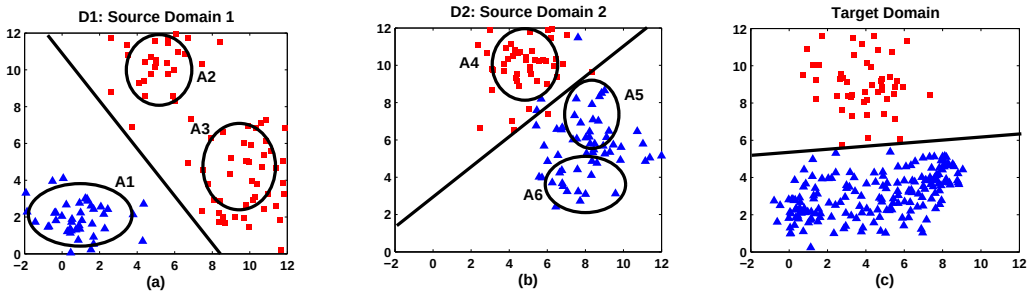


Figure 3.4: Two source domains D1 and D2 and target domain data with different marginal and conditional probability differences, along with conflicting conditional probabilities (the red squares and blue triangles refer to the positive and negative classes).

hypotheses obtained based on traditional machine learning methodologies and a target data

distribution. It is evident that the hypotheses learned by the two source distributions D1 and D2 would perform poorly on the target distribution.

3.2.3 Learning the Target Classifier

The target classifier is learnt based on the re-weighted source data and the few labeled target domain data (if available). An additional weighting factor μ, is incorporated to provide a differential weight to the source domain data with respect to the labeled target domain data. Mathematically, the target classifier $\hat{h}$ is learnt by solving the following optimization problem:

$$\hat{h} = \underset{h}{\operatorname{argmin}} \quad \mu \left(\sum_{s=1}^{k} \frac{\beta^s}{n_s} \sum_{i=1}^{n_s} \alpha_i^s \mathcal{L}(h(x_i^s), y_i^s) \right) + \sum_{j=1}^{n_l} \frac{1}{n_l} \mathcal{L}(h(x_j^T), y_j^T) \tag{3.14}$$

where n_l is the number of labeled data from the target domain.

Algorithm 2 below summarizes the main steps involved in 2SW-MDA.

ALGORITHM 1: 2SW-MDA

1: **Input** μ, k source domain data $\{D^s\}_{s=1}^k$, unlabeled target domain data D_u^T and labeled target domain data D_l^T (if available)
2: **Output** Target classifier h
3: **for** $s = 1, \ldots, k$ **do**
4: Compute α^s by solving (8.3)
5: Learn a hypothesis h^s on the α^s weighted source data
6: **end for**
7: Form the $n_u \times k$ prediction matrix H^S as in Section 3.2.2
8: Compute matrices W, D and L using the unlabeled target data D_u^T
9: Compute β^s by solving (3.7)
10: Learn the target classifier $\hat{h}$ by solving (3.14)

Theoretical Analysis

This dissertation also presents a theoretical analysis of the joint loss function given in Equation 3.14 and present an upper bound on the error of the target classifier h (learned by minimizing 3.14) on target domain data. To do this, an upper bound on the empirical joint error with respect to the true joint error is proved and then an upper bound on the error on the target domain data only, is proved.

Bound on joint error function

For convenience of presentation, the empirical joint error function on (α, β)-weighted source domain and the target domain defined in (3.14) is re-written as follows:

$$\hat{E}_{\alpha,\beta}^S(h) = \mu \hat{\epsilon}_{\alpha,\beta}(h) + \hat{\epsilon}_T(h) = \mu \sum_{s=1}^{k} \frac{\beta^s}{n_s} \sum_{i=1}^{n_s} \alpha_i^s \mathcal{L}(h(x_i^s), f_s(x_i^s)) + \sum_{i=1}^{n_l} \frac{1}{n_l} \mathcal{L}(h(x_i^0), f_0(x_i^0)) \tag{3.15}$$

where $y_i^s = f_s(x_i^s)$ and f_s is the labeling function for source s, $\mu > 0$, (x_i^0) are samples from the target, $y_i^t = f_0(x_i^0)$ and f_0 is the labeling function for the target domain, and $S = (x_i^s)$ include all samples from the target and source domains. The true (α, β)-weighted error $\epsilon_{\alpha,\beta}(h)$ on weighted source domain samples is defined analogously. Similarly, $E^S_{\alpha,\beta}(h)$ is defined as the true joint error function. For notational simplicity, denote $n_0 = n_l$ as the number of labeled samples from the target, $m = \sum_{s=0}^{k} n_s$ as the total number of samples from both source and target, and $\gamma_s^i = \mu\beta^s\alpha_i^s/n_s$ for $s \geq 1$ and $\gamma_s^i = 1/n$ for $s = 0$. Then the empirical joint error function in (3.15) can be re-written as:

$$\hat{E}^S_{\alpha,\beta}(h) = \sum_{s=0}^{k}\sum_{i=1}^{n_s} \gamma_i^s \mathcal{L}(h(x_i^s), f_s(x_i^s)).$$

Next, the difference between the true joint error function $E^S_{\alpha,\beta}(h)$ and its empirical estimate $\hat{E}^S_{\alpha,\beta}(h)$ is bounded using the weighted Rademacher complexity measure [2, 49] defined as follows:

Definition 1. *(Weighted Rademacher Complexity) Let $\mathbb{H}$ be a set of real-valued functions defined over a set X. Given a sample $S \in X^m$, the empirical weighted Rademacher complexity of $\mathbb{H}$ is defined as follows:*

$$\hat{\mathfrak{R}}_S(H) = E_\sigma\left[\sup_{h\in\mathbb{H}} |\sum_{s=0}^{k}\sum_{i=1}^{n_s} \gamma_i^s \sigma_i^s h(x_i^s)| \,\middle|\, S = (x_i^s)\right].$$

The expectation is taken over $\sigma = \{\sigma_i^s\}$ where $\{\sigma_i^s\}$ are independent uniform random variables taking values in $\{-1, +1\}$. The weighted Rademacher complexity of a hypothesis set $\mathbb{H}$ is defined as the expectation of $\hat{\mathfrak{R}}_S(H)$ over all samples of size m:

$$\mathfrak{R}_m(H) = E_S\left[\hat{\mathfrak{R}}_S(H) \,\middle|\, |S| = m\right].$$

Our main result is summarized in the following lemma, which involves the estimation of the Rademacher complexity of the following class of functions:

$$\mathbb{G} = \{x \mapsto \mathcal{L}(h'(x), h(x)) : h, h' \in \mathbb{H}\}.$$

Lemma 1. *Let $\mathbb{H}$ be a family of functions taking values in $\{-1, +1\}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, the following holds for $h \in \mathbb{H}$:*

$$\left|E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right| \leq 2\mathfrak{R}_S(\mathbb{H}) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}.$$

Furthermore, if $\mathbb{H}$ has a VC dimension of d, then the following holds with probability at least $1-\delta$:

$$\left|E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right| \leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\left(\sqrt{2d\log\frac{em}{d}}+1\right),$$

where e is the natural number.

The proof is provided in Appendix A.

Error bound on target domain with re-weighted source domain

The previous section presented an upper bound on the difference between the true joint error function and its empirical estimate and established its relation to the weighting factors γ_i^s. Next, the main theoretical result, i.e., an upper bound of the error function on target domain data, i.e., an upper bound of $\epsilon_T(\hat{h})$ is presented. The following definition of divergence are needed for the main result:

Definition 2. *For a hypothesis space $\mathcal{H}$, the symmetric difference hypothesis space $d_{\mathbb{H}\Delta\mathbb{H}}$ is the set of hypotheses*

$$g \in \mathbb{H}\Delta\mathbb{H} \Leftrightarrow g(x) = h(x) \oplus h^{'}(x)\ for\ some\ h, h^{'} \in \mathcal{H},$$

where $\oplus$ is the XOR function. In other words, every hypothesis $g \in \mathbb{H}\Delta\mathbb{H}$ is the set of disagreements between two hypotheses in $\mathcal{H}$.

The $\mathbb{H}\Delta\mathbb{H}$-divergence between any two distributions D_S and D_T is defined as

$$d_{\mathbb{H}\Delta\mathbb{H}}\left(D_S, D_T\right)) = 2\sup_{h,h'\in\mathbb{H}}\left|Pr_{x\backsim D_S}[h(x)\neq h'(x)] - Pr_{x\backsim D_T}[h(x)\neq h'(x)]\right|.$$

Theorem 1. *Let $\hat{h}\in\mathbb{H}$ be an empirical minimizer of the joint error function on similarity weighted source domain and the target domain:*

$$\hat{h} = \textit{arg}\min_{h\in\mathbb{H}}\hat{E}_{\alpha,\beta}(h) \equiv \mu\hat{\epsilon}_{\alpha,\beta}(h) + \hat{\epsilon}_T(h)$$

for fixed weights μ, α, and β and let $h_T^* = \min_{h\in\mathbb{H}}\epsilon_T(h)$ be a target error minimizer. Then for any $\delta\in(0,1)$, the following holds with probability at least $1-\delta$:

$$\begin{aligned}\epsilon_T(\hat{h}) &\leq \epsilon_T(h_T^*) + \frac{2\Re_S(H)}{1+\mu} + \frac{2}{1+\mu}\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} \\ &+ \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right), \end{aligned} \tag{3.16}$$

if $\mathbb{H}$ has a VC dimension of d, then the following holds with probability at least $1-\delta$:

$$\begin{aligned}\epsilon_T(\hat{h}) &\leq \epsilon_T(h_T^*) + \frac{2}{1+\mu}\left(\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\left(\sqrt{2d\log\frac{em}{d}}+1\right)\right) \\ &+ \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right), \end{aligned} \tag{3.17}$$

where $\lambda_{\alpha,\beta} = \min_{h\in\mathbb{H}}\{\epsilon_T(h)+\epsilon_{\alpha,\beta}(h)\}$, and $d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)$ is the symmetric difference hypothesis space for (α,β)-weighted source and target domain data.

The proof as well as a comparison with the result in [4] is provided in the Appendix A.

It is observed that μ and the divergence between the weighted source and target data play significant roles in the generalization bound. The proposed two-stage weighting scheme aims to reduce the divergence. Next, the effect of μ is analyzed. When $\mu = 0$, the bound reduces to the generalization bound using the n_l training samples in the target domain only. As μ increases, the effect of the source domain data increases. Specifically, when μ is larger than a certain value, for the bound in (3.17), as μ increases, the second term will reduce, while the last term capturing the divergence will increase. In the extreme case when $\mu = \infty$, the second term in (3.17) can be shown to be the generalization bound using the weighted samples in the source domain only (the target data will not be effective in this case), and the last term equals to $2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)$. Thus, effective transfer is possible in this case only if the divergence is small. It is also observed in the experiments that the target domain error of the learned joint hypothesis follows a bell shaped curve; it has a different optimal point for each dataset under certain similarity and divergence measures.

3.2.4 Results and Analysis

The comparative performance of CP-MDA with 2SW-MDA is presented followed by a discussion on the relative performance of the two developed methods.

The proposed algorithms have been evaluated on three real-world datasets 20 Newsgroup [1], Sentiment Analysis[2] and dataset of multi-dimensional feature vectors extracted from SEMG (Surface electromyogram) signals collected from 8 subjects during a fatiguing exercise.

Each document is represented as a binary vector of the 100 most discriminating words determined by Weka's info-gain filter [89]. Out of the 20 categories, 13 categories are used to form

[1] Available at http://www.ai.mit.edu/∽jrennie/20Newsgroups/
[2] Available at http://www.cs.jhu.edu/∽mdredze/

the source and target domains. For each of these categories the negative class was formed by a random mixture of the rest of the seven categories, as suggested in [19]. The details of the 13 categories used can be found in the supplemental material. The Sentiment Analysis dataset contains positive and negative reviews on four categories (or domains) including *kitchen*, *book*, *dvd*, and *electronics*. The Sentiment Analysis dataset was processed to reduce the feature dimension to 200 using a cutoff document frequency of 50.

Table 3.7: Comparison of different methods on three real-world and one toy datasets in terms of classification accuracies (%).

Dataset	SVM-C	LWE	KE	KMM	TCA	DAM	CP-MDA	2SW-MDA
talk.politics.mideast	46.00%	50.66%	49.01%	45.78%	58.66%	52.03%	61.44%	73.49%
	49.33%	49.39%	53.48%	39.75%	56.00%	52.00%	57.10%	65.06%
	49.33%	50.27%	54.67%	43.37%	52.04%	51.81%	57.34%	62.65%
talk.politics.misc	48.83%	53.62%	46.77%	62.32%	55.90%	53.22%	62.31%	63.67%
	48.22%	51.12%	48.39%	59.42%	53.23%	54.12%	53.62%	60.87%
	48.31%	50.72%	55.01%	59.07%	54.83%	54.12%	56.08%	68.12%
comp.sys.ibm.pc.hardware	48.42%	51.25%	49.50%	50.56%	61.25%	52.50%	68.53%	62.92%
	47.44%	51.44%	49.44%	59.55%	57.50%	52.50%	53.93%	60.67%
	45.93%	49.88%	48.00%	58.43%	59.75%	57.80%	56.17%	64.04%
rec.sport.baseball	56.25%	61.51%	47.50%	61.79%	61.75%	61.25%	74.15%	79.78%
	58.75%	50.09%	51.25%	64.04%	57.75%	53.75%	67.41%	60.22%
	56.35%	59.26%	56.25%	58.43%	57.83%	55.05%	52.80%	61.24%
kitchen	35.55%	40.12%	49.38%	64.04%	64.10%	58.61%	55.0%	70.55%
electronics	35.95%	42.66%	48.38%	65.55%	54.20%	52.61%	53.88%	59.44%
book	37.77%	40.12%	49.38%	58.88%	55.01%	54.10%	53.33%	59.47%
dvd	36.01%	49.44%	48.77%	50.00%	50.00%	52.61%	51.11%	51.11%
SEMG- 8 subjects	70.76%	67.44%	63.55%	64.94%	66.35%	74.83%	81.93%	83.03%
	43.69%	77.54%	74.62%	63.63%	59.94%	81.36%	84.73%	87.96%
	50.11%	75.55%	62.50%	64.06%	56.78%	74.77%	82.54%	88.96%
	59.65%	81.22%	69.35%	52.68%	73.38%	80.63%	81.27%	88.49%
	40.37%	52.48%	65.61%	49.77%	57.48%	76.74%	80.74%	86.14%
	59.21%	65.77%	83.92%	70.62%	76.92%	59.21%	83.12%	87.10%
	47.13%	60.32%	77.97%	51.13%	55.64%	74.27%	81.57%	87.08%
	69.85%	72.81%	79.48%	67.24%	42.79%	84.55%	88.50%	93.01%
Toy data	60.05%	75.63%	81.40%	68.01%	64.97%	84.27%	93.21%	98.54%

Comparative Studies.Table 3.7 shows the classification accuracies of different methods on the real-world and the toy datasets. One can observe that SVM-C performs poorly for all cases. This may be attributed to the distribution difference among the multiple source and target domains. It is observed that 20 Newsgroups and Sentiment Analysis datasets have predominantly marginal probability differences. In other words, the frequency of a particular word varies from one category of documents to another. In contrary physiological signals, such as SEMG are predominantly different in conditional probability distributions due to the high subject based variability in the power spectrum of these signals and their variations as fatigue sets in [13, 25, 26]. One can also observe that the proposed CP-MDA and 2SW-MDA methods outperform other domain adaptation methods and achieve higher classification accuracies in most cases, specially for the SEMG dataset.

For some cases of the 20 Newsgroup and Sentiment Analysis datasets which have significant marginal probability differences, CP-MDA which addresses predominantly conditional probability differences, is outperformed by 2SW-MDA and KMM methods, which address marginal probability differences. The accuracies of an SVM classifier, on the toy dataset, when learned only on the source domains D1, D2 individually and on the combined source domains, are 60.67% and 71.84% and 60.05% respectively, while 2SW-MDA achieves an accuracy of 98.54%. More results are provided in Appendix A.

It is observed that 2SW-MDA performs better than CP-MDA, this can be attributed to the fact that 2SW-MDA addresses both marginal and conditional probability differences, where as CP-MDA addresses only conditional probability differences. Also the conditional probability based weights are computed with source hypothesis learned on re-weighted source instances (as per marginal probability differences) , thus increasing the accuracy of computed weights.

It is interesting to note that instance re-weighting method KMM and feature mapping based method TCA, which address marginal probability differences between the source and target domains perform better than LWE and KE for both 20 Newsgroups and Sentiment Analysis data. They also perform better than DAM, a multi-source domain adaptation method, based on marginal probability based weighted hypotheses combination. It is worthwhile to note that LWE is based on conditional probability differences and KE tries to address both the differences. Thus, it is not surprising that LWE and KE perform better than KMM and TCA for the SEMG dataset, which is predominantly different in conditional probability distributions. DAM too performs better for SEMG signals. However the proposed CP-MDA and 2SW-MDA methods outperform all the other methods in most cases. The experiments verify the effectiveness of the proposed frameworks.

The accuracies of an SVM classifier, on the toy dataset, when learned only on the source domains D1, D2 individually and on the combined source domains, are 60.67% and 71.84% and 60.05% respectively, while 2SW-MDA achieves an accuracy of 98.54%. More results are provided in the Appendix A.

Parameter Sensitivity Studies. In this experiment, the effect of μ on the classification performance is studied. Figure 3.5 shows the variation in classification accuracies for some cases presented in Table 3.7, with varying μ over a range $[0\ 0.001\ 0.01\ 0.1\ 0.3\ 0.5\ 1\ 100\ 1000]$. The x-axis of the figures are in logarithmic scale. The results for the toy data are included in Appendix A. The results show

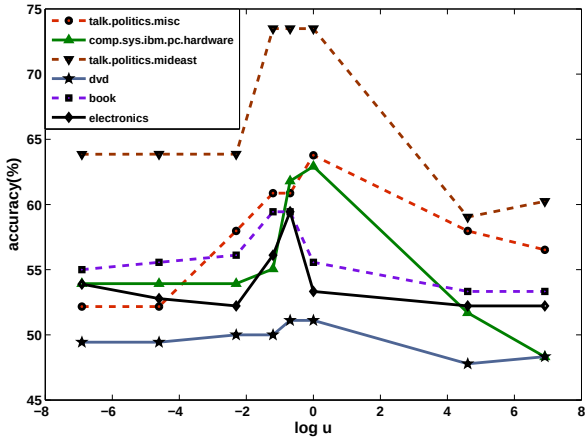


Figure 3.5: Performance of the proposed 2SW-MDA method on 20 Newsgroups dataset and Sentiment Analysis dataset with varying μ.

that in most cases, the accuracy values increase as μ increases from 0 to an optimal value and decreases when μ further increases. When $\mu = 0$ the target classifier is learned only on the few labeled data from the target domain. As μ increases the transfer of knowledge due to the presence of additional weighted source data has a positive impact leading to increase in classification accuracies in the target domain. Results show that after a certain value of μ the classifier accuracies drop, due to the distribution differences between the source and target domains. These experimental results are consistent with the theoretical results established in this thesis.

3.3 Hierarchical Confidence Weighted Multi-Source Domain Daptation (HC-MDA)

Large variations in Surface Electromyogram (SEMG) signal across different subjects make the process of automated signal classification as a generalized tool, challenging. This dissertation develops a domain adaptation method based on a hierarchical sample selection algorithm. The developed method selects samples from multiple training subjects, based on their similarity with the target subject at different levels of granularity.

3.3.1 Overview

It is clear from Figure 1.1 that in SEMG signal different classes vary differently over subjects. For example, subjects 1 and 2 have similar data distribution for classes 1 and 3, and subjects 2 and 4 have similar data distribution for classes 4. Hence computing a single similarity measure for a source domain subject data with respect to target data does not capture the differences at the class level. The proposed approach presented below addresses this challenge by following a hierarchical confidence weighted sample selection strategy that considers similarities at all levels.

3.3.2 Hierarchical Confidence Weighted Sample Selection

In the developed method the similarities between the source domain subject data and the target subject data is measured at three different levels, each with increasing granularity. The framework *Hierarchical Confidence Weighted Multi Source Domain Adaptation* (HC-MDA) is outlined in Algorithm 2.

As a first step, the proposed approach learns a model M^{T_l} from the labeled target subject data D_l^T. Next, it computes the classification accuracy obtained on each subject data $\{D^k\}_{k=1}^K$, by this model. This classification accuracy given by w^k is used to estimate the similarity of a particular source domain subject k with respect to the target subject. The classification accuracy reflects the differences in both marginal and conditional probability distributions. The classification accuracies are normalized across subjects to obtain a relative similarity measure between a source domain and a target domain subject. While this measure encapsulates the overall similarity of a source domain subject k with respect to target subject, it does not address the distribution difference or similarities at individual class level.

The next level involves computing similarity between the individual classes of source domain subject and the target subject. This similarity measure is computed by determining the average true positive rate, w_c^k, for a class c belonging to a source domain subject k. Normalized

w_c^k, class wise across subjects, reflects conditional probability differences between source and target subject classes.

However this measure still overlooks the dissimilarities between the instances of source subject with respect to target subject data e.g. even if a particular class has a true positive rate as high as 80%, there are still 20% of the instances which are not similar to the target domain distribution. In order to avoid selecting these instances for domain adaptation, the proposed framework advocates computation of similarity at a still finer level of granularity. This selection is achieved by concentrating only on the correctly classified instances of a class and selecting instances with higher confidence of prediction, $D_{sel}^{K,C}$. A classifier is learned these and labeled target samples D_l^T and the learned model is used to classify and label the unlabeled target domain data D_u^T.

ALGORITHM 2: The Hierarchical Confidence Weighted-Multi-Domain Adaptation

Input Source domain subject samples $\{D^k\}_{k=1}^K$ and small amount of target domain subject training examples D_l^T
Output $D_{sel}^{K,C}$
Learn a model M^{T_l} using D_l^T
for $k = 1, \ldots, K$ **do**
    Weight for source D^k : w^k = Classification accuracy for D^k using $M^T l$
    for $c = 1 \ldots, C$ **do**
        Weight for class c of source D^k: w_c^k = True positive rate for class c on the samples from D^k using $M^T l$
    end for
end for
Normalize the w^k and w_c^k for each c, over all K subjects
Compute the number of samples N_c^k to be selected from a source subject k for a class c

$$N_c^k = (w_c^k \times w^k) \times |\left\{x_i : M^{T_l}(x_i) = y_i and y_i = c\right\}| \tag{3.18}$$

$D_{sel}^{k,c}$ = first N_c^k samples $\in \left\{x_i : M^{T_l}(x_i) = y_i and y_i = c\right\}$ with the highest confidence of classification using M^{T_l}
Output $D_{sel}^{K,C}$

3.3.3 Experiments and Results

Results are obtained from a leave one subject out cross validation process. A set of 4 labeled samples are randomly selected from the target subject to constitute D_l^T. These labeled test data is added to all the methods for fair comparison. The classification accuracies are averaged over ten folds of execution to remove any bias due to selection of any specific labeled data set from test subject.

Average cross validated accuracy for each of the methods is summarized in Table 3.10. The first column of the table indicates the test subject and the training data consists of the data from the remaining seven subjects. The results show that the proposed method significantly improves the performance over other methods by an average gain of 15% to 20%. LWE, that

Table 3.8: Comparative performance of different methods on SEMG data - Accuracy (%)

Test Sub	SVM-C	SVM-T	TCA	KE	KMM	LWE	HC-MDA
1	72.76	62.12	55.45	65.45	72.42	67.44	82.61
2	53.69	67.50	59.94	60.98	63.63	77.54	80.06
3	55.11	62.58	72.57	63.16	68.69	75.55	81.45
4	59.65	64.42	69.89	59.68	72.38	81.22	87.05
5	60.37	71.87	64.06	61.33	62.5	52.48	87.97
6	59.21	49.09	59.02	54.54	70.62	65.77	78.86
7	57.13	51.09	62.42	60.17	61.13	60.32	80.43
8	64.85	70.79	62.48	83.41	74.79	68.55	85.73
Average	60.34	62.43	63.22	64.41	68.27	69.14	**83.02**

addresses conditional probability distribution difference, performs better than other methods that address only marginal probability differences. However LWE, performs poorer than HC-MDA since it computes the weights of each subject depending upon a overall similarity factor, overlooking the similarities at different levels.

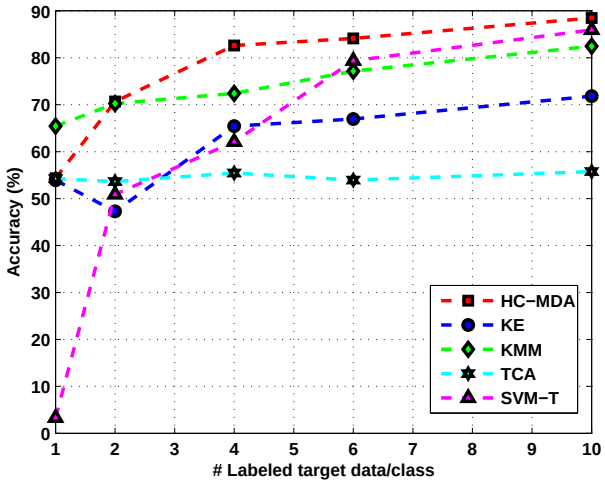


Figure 3.6: Accuracy (%) Vs No of labeled target data

Figure 3.6 presents the variation in classification accuracies for different methods with varying number of labeled data available from the target subject. The performance of HC-MDA is moderate when the number of labeled samples from target subjects is either 1 or 2. This is because the small number of samples is insufficient to learn the target model M_l^T. However beyond 2 labeled samples from target subject, HCMDA performs better than other techniques. Performance of TCA is low because of conditional probability differences in the data, that negatively affects the feature mapping process of TCA [42]. Even though KE too involves a feature mapping step, the sample selection strategy helps it to perform better than TCA.

SVM-T performs very poorly upto 4 samples per class. After 6 samples per class, SVM-T performs better than most of the approaches as the number of labeled samples from target subject is sufficient to learn a reliable model.

3.4 Optimization based Domain Adaptation (ODA)

This section presents a single source domain adaptation method addressing both marginal and conditional probability differences between the source and the target domains. An unique feature of this method is an optimization formulation which addresses both these differences in a single stage, unlike the existing methods which achieve this objective following a two step strategy.

3.4.1 Overview

The developed ODA method selects instances from the source domain (data available from other subjects) based on a novel optimization formulation, to ensure that the selected instances are similar in distribution to the target domain (test subject data) in both marginal and conditional probability distributions. As the proposed domain adaptation framework is based on an optimization formulation it is referred to as Optimization-based Domain Adaptation (ODA). The proposed method (ODA) is different from the KMM, TCA and LWE methods (see Chapter 8) as it addresses both marginal and conditional probability differences between the source and target domains. And unlike KE (Chapter 8), ODA addresses both these distribution differences based on a sample selection strategy, rather than feature mapping, and it does so using an unified optimization framework rather than addressing these differences using a two-step strategy. The section below presents the single stage Optimization-based Domain Adaptation (ODA) framework.

3.4.2 Single Stage Optimization for Marginal and Conditional Differences

Let us consider the following classification scenario. Each of several subjects s in the source domain is characterized by sample set $D^s = (x_i^s, y_i^s)|_{i=1}^{n_s}, s = 1, 2, \cdots k$ where x_i is the feature vector, y_i the corresponding label, n_s the total number of data for the subject s and k is the total number of subjects in the source domain. The target domain consists of sample sets of few labeled data $D_l^T = (x_i^T, y_i^T)|_{i=1}^{n_l}$, and plenty of unlabeled data $D_u^T = x_i^T|_{i=n_l+1}^{n_l+n_u}$, from the test subject. Here n_l and n_u are numbers of labeled and unlabeled data respectively and $D^T = D_l^T \bigcup D_u^T$ and $n_T = n_l + n_u$.

The goal of the proposed person-adaptive framework is to develop a classification model for the test subject or target domain data using the source domain data consisting of data from many other subjects. The greatest challenge in doing so is addressing the distribution difference between the data of the subjects forming the source and target domain. Hence, this domain

adaptation methodology selects data from the source domain such that these data points are similar to target domain both in marginal and conditional probability distribution. In other words, if we were to select samples from the training data shown in Figure 2.1(b) to better classify samples in test data shown in Figure 2.1(c) then we should select points lying in circles E and F, taking care that points shown in the circle C do not get selected, as these points have conflicting conditional probabilities and would degrade the performance of the target classifier learned on these data points.

An objective function which satisfies these requirements is formulated under the assumption that the number of samples to be selected from the source domain is specified a-priori and there are a few labeled samples from the target domain. Two criteria are employed for this purpose. (i) Firstly, in order to ensure that selected samples from the source domain are similar in marginal probability to the target domain samples, the method selects those source domain samples which are closer to the target domain by intrinsic data geometry in the feature space. (ii) Secondly, in order to ensure that these points also have similar conditional probabilities, a classifier is learned on the selected samples from the source domain and validated on the unlabeled target domain data. If the samples selected from source domain have similar conditional probability distributions as well, then the entropy of labels obtained on the target domain data will be minimum.

To formulate the first criterion, let d_i denote the average distance of a data point x_i belonging to source domain data D^S from the available labeled pool of target domain data D_l^T. A greater value of d_i denotes that the point is located away from the high-density region in the intrinsic geometry of target domain data. In order to ensure that the objective function is differentiable, Euclidean distance is used in this work. Any other differentiable distance metric can also be used for this purpose.

The objective of the second criterion is to select a set L with m data samples from the source domain data D^S ($D^S = \sum_{s=1}^{k} D^s$) in such a way that the classifier trained on $D_l^T \bigcup L$ (labeled target domain data and the selected source domain data whose labels are known) has minimum entropy of labels on the unlabeled target domain data D_u^T as illustrated in Figure 3.7. Let w^t denote the current classifier, and w^{t+1} denote the classifier at time $t+1$ which is obtained by training on $D_l^T \bigcup L$, where L is the set of samples selected at time t using w^t (w^{t+1} is also called as the future classifier in this work). Let C denote the possible number of classes, then the entropy

S of the labels or the conditional distribution, at time t+1, which is given by $P(y|x_j, w^{t+1})$, where x_j is the j^{th} sample in the unlabeled pool of target domain data and y the class label, is computed as:

$$S(y|x_j, w^{t+1}) = -\sum_{y \in C} P(y|x_j, w^{t+1}) log P(y|x_j, w^{t+1}) \tag{3.19}$$

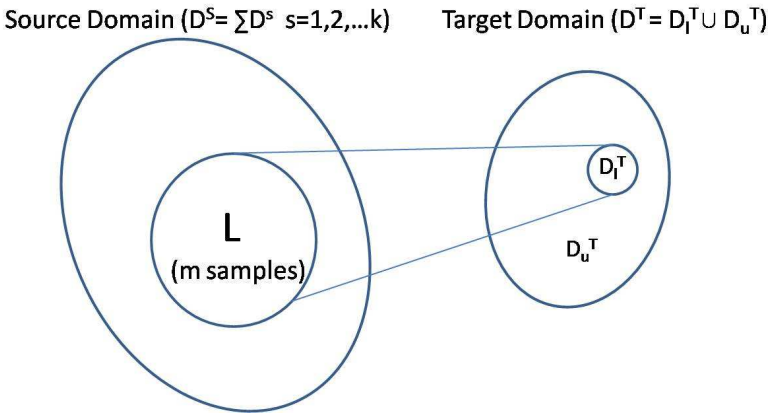


Figure 3.7: Framework based on selecting m samples from the source domain (D^S) which are similar in marginal and conditional probability distribution to the target domain (D^T).

The two conditions described above are then combined to form a cost function $f(L)$ as follows:

$$f(L) = \sum_{i \in L} d_i + \lambda \sum_{j \in D_u^T} S(y|x_j, w^{t+1}) \tag{3.20}$$

The first term denotes the sum of the average distances of m selected points from the source domain to the labeled target domain data D_l^T, and the second term quantifies the sum of label entropies on the unlabeled target domain data D_u^T using w^{t+1}. λ is a trade-off parameter which controls the relative importance of the two terms.

Hence, the problem now reduces to selecting a set of m points belonging to L which produces a minimum score of $f(L)$. Since the search space is exponentially large, exhaustive search methods are infeasible, and numerical optimization techniques need to be used to solve this minimization problem. Let $|D^S|$ be the number of points in the source domain data. A binary vector B of size $|D^S|$ is defined, where each entry, B_i, denotes whether the data point x_i is selected in the current batch of m data points. If a point is selected, the corresponding entry B_i is 1, else 0. B is obtained by minimizing the value of the objective function $f(L)$, modified as below:

$$\min_B \sum_{i \in D^S} B_i d_i + \lambda \sum_{j \in D_u^T} S(y|x_j, w^{t+1}) \tag{3.21}$$

subject to the constraints:

$$B_i \in \{0, 1\} \tag{3.22}$$

$$B'\mathbf{1} = m \tag{3.23}$$

where **1** is a vector of the same dimension as B with all entries equal to 1 and symbol $'$ is used to represent the matrix or vector *transpose* operation. Evidently, the cost function in Equation 3.21 is an alternative representation of the cost function $f(L)$ in Equation 3.31. The first term denotes the sum of the average distances of the selected points in the source domain from the target domain. Note that if a point x_i is not selected in the current set then B_i will be 0 and this term would vanish. The second term is the sum of the entropy values on the unlabeled target domain, as obtained by the future classifier w^{t+1} which is trained on source samples selected using B (note that the second term thus depends on B implicitly). The first constraint ensures that each entry in B is either 0 or 1 and the second constraint ensures that exactly m entries of B are 1, meaning exactly m points are selected from the source domain, where m is specified a priori by the user.

The above problem is an integer programming problem and is NP hard. Hence the constraints are relaxed to make it a continuous optimization problem, as below.

$$\min_{B} \sum_{i \in D^S} B_i d_i + \lambda \sum_{j \in D_u^T} S(y|x_j, w^{t+1}) \tag{3.24}$$

subject to the constraints:

$$0 \leq B_i \leq 1 \tag{3.25}$$

$$B'\mathbf{1} = m \tag{3.26}$$

B is obtained from this formulation and the top largest m entries are set as 1 to approximate the integer solution. Since it is not possible to obtain an analytical global solution for such a formulation, the iterative *Quasi Newton* method is deployed to produce a locally optimum solution. Quasi Newton assumes that the function can be well-approximated as a quadratic in the neighborhood of the optimum point and iteratively updates the variable to guide the functional value towards a local optima. In each iteration a quadratic programming problem is solved to obtain an updated direction for B, and the Hessian matrix is updated according to the BFGS method, suggested by Broyden, Fletcher, Goldfarb and Shanno in 1970 [41]. The step size to update B is obtained by using backtracking line search method based on the Armijo Goldstein equation. As mentioned earlier, the classifier w^{t+1} is obtained after training on $D_l^T \bigcup L$, where L is obtained by solving B. The entropy of labeling the unlabeled target domain data is computed

using Equation 3.19 and the value of the new cost function is computed using Equation 3.31. The iterations are continued till the change in cost function value is negligible or below an epsilon value. As this is a known method please refer [41] for more details on this numerical optimization technique.

3.4.3 Experiments and Results

Effectiveness of proposed framework ODA is evaluated against the baseline methods SVM-C, SVM-T and TSVM; and against state-of-the-art domain adaptation methods namely Kernel Mean Matching (KMM), Transfer Component Analysis (TCA), Kernel Ensemble (KE) and Locally Weighted Ensemble (LWE). In SVM-C, the training data comprises of data from seven subjects and the test data is from the remaining one subject. In SVM-T, the model is learned only on the labeled test data D_l^T. TSVM refers to Transductive SVM [1], which is a semi-supervised method where the training data consists of unlabeled data from the test subject, besides the labeled data from the remaining seven subjects. The description of the relevant domain adaptation methods are presented in Section 8.

Results are obtained using a leave-one-subject-out strategy i.e., at each experiment, data from seven subjects are considered as training data and the data from the remaining subject as test data. A set of 4 labeled samples from each class (around 4% of the target domain data) is randomly selected from the test data to constitute D_l^T, and the number of training samples to be selected for domain adaptation in case of ODA was fixed at 10% of training data (results with varying portions of training data are presented in Section 3.4.3). The same labeled test data is added to all the methods for fair comparison. The classification accuracies are averaged over ten trials to remove any bias due to random selection of initial labeled test data.

Model parameters for different techniques were obtained through a cross validation process on a set aside validation data set consisting of samples from all subjects. SVM-C was trained with a Gaussian kernel with $\sigma = 0.5$ and high C. For KMM, a Gaussian kernel with $\sigma = 10$ gave best results on the validation set. TCA was implemented with linear kernel and feature mapping was performed with dimension value of 10. The same dimension value of 10 was used for feature mapping in KE.

Table 3.9: Comparative performance of different methods on SEMG data - Accuracy (%)(training data for all methods have 4 labeled samples/class from test subject data for fair comparison).

Test Sub	SVM-C	SVM-T	TSVM	TCA	KE	KMM	LWE	ODA
1	67.27	62.12	51.69	55.45	65.45	72.42	67.44	**83.57**
2	53.69	67.50	57.08	59.94	60.98	63.63	77.54	**79.70**
3	55.11	62.58	67.09	72.57	63.16	68.69	75.55	**77.01**
4	59.65	64.42	58.78	69.89	59.68	72.38	81.22	**85.75**
5	60.37	71.87	64.35	64.06	61.33	62.5	52.48	**80.47**
6	59.21	49.09	72.02	59.02	54.54	70.62	65.77	**86.71**
7	57.13	51.09	62.53	62.42	60.17	61.13	60.32	**70.34**
8	64.85	70.79	70.75	62.48	83.41	74.79	68.55	**86.46**
Average	59.66	62.43	63.03	63.22	64.41	68.27	69.14	**81.25**

Results and Discussion

The results of the experimental procedure explained above are described in this section. The average cross-validated accuracy for each of the methods is summarized in Table 3.9. The first column of the table indicates the test (target) subject and the training data consists of the data from the remaining seven subjects. An analysis of the performance with different amounts of labeled test data and training data is presented in later sections. The results show that the proposed method ODA significantly improves the average performance over the baseline methods SVM-C, SVM-T and TSVM by 19% to 21% and over the state-of-the art domain adaptation methods TCA, KE, KMM and LWE by 12% to 18%. Please note although TSVM performs better than other baseline methods, it does not perform better than the domain adaptation methods. This maybe due to the conflicting conditional probabilities in SEMG data, which affects the label propagation process in TSVM adversely.

In comparison with other domain adaptation methods, the proposed method ODA performs better than both Transfer Component Analysis (TCA) and Kernel Mean Matching (KMM) as they address only marginal probability differences, whereas ODA addresses both marginal and conditional probability differences, as mentioned earlier. It is interesting to observe that though both KMM and TCA are based on marginal probability differences, KMM, which is based on instance selection, performs better than TCA in 5 out of 8 cases. This can be attributed to the fact that TCA is based on feature mapping methodologies as explained in Section 8. Since SEMG data has significant conditional probability differences in addition to marginal probability differences, feature mapping followed by dimensionality reduction degrades the performance, by increasing the entropy of labels [42]. Similarly, although KE addresses both marginal and conditional probability distribution differences, this method performs poorer than the proposed

ODA framework since it also involves a feature mapping using Kernel Discriminant Analysis. LWE, which addresses only conditional probability distribution differences, performs better than other methods but is inferior to the performance of ODA. This may be caused due to two reasons: (i) Firstly, the LWE method does not consider marginal probability differences, and (ii) Secondly, the method is inherently based on clustering, thus expecting the data to lie on a certain cluster manifold which may not be true for this data at higher dimension.

Sensitivity to Parameter Settings

The proposed ODA algorithm has two major parameters in the formulation: (i) the number of labeled samples available from the test subject i.e., $|D_l^T|$, and (ii) the number of samples selected from the source domain or training data i.e., m which is defined apriori by the user, as explained in Section 3.4. The section below presents the sensitivity of the ODA method to these two parameters.

Number of Labeled Samples from Test Subject

Figure 3.8 presents the variation in classification accuracies for different methods with varying number of labeled samples from the test subject. One can observe that the performance of all methods is moderate when the number of labeled samples (from each class) from the test subject is either 1 or 2 (1-2% of target data). This is because the small number of samples is insufficient to represent the test data distribution reasonably well. However, when more than 2 labeled samples (from each class) are selected from the test subject, ODA performs better than other domain adaptation methods. It is interesting to note that the improvement in performance is marginal when the number of labeled test samples per class is increased from 4 to 10. This shows that 4 labeled test samples per class (4% of target data) were sufficient to represent the test data distribution reasonably well (thus explaining this choice in our experiments in the previous section).

It can also be seen that the performance of TCA does not vary much with increase in number of labeled test samples (due to the conflicting conditional probabilities in our SEMG data). For the same reason, KE performs better than TCA and its classification accuracy improves with the number of labeled test data. It is interesting to observe that although KMM addresses only the marginal probability difference, it performs better than KE and TCA which involve feature mapping, since it is inherently based on an instance selection strategy. TSVM, which is a traditional semi-supervised method also improves with more number of labeled samples from the test subject, but does not perform as well as most domain adaptation methods. SVM-T, which is a

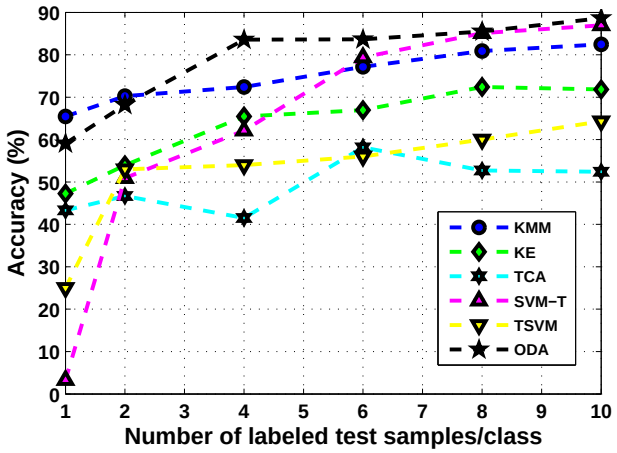


Figure 3.8: Accuracy(%) vs Number of labeled test samples per class (x-axis) for different domain adaptation methods.

classifier learned only on the available labeled test samples, performs poorer than other methods when the number of labeled target data samples is low. However, it performs better when there are more number of labeled target data samples. In summary, ODA outperformed all other method in our studies, thus demonstrating the effectiveness of the proposed framework.

Number of Samples Selected from Training Set for Domain Adaptation

Another significant parameter in the ODA framework is the number of samples to be selected from training data (m), which is specified by the user. Figure 3.9 presents the results of varying m (x-axis) against the accuracy under different conditions of labeled samples from the test subject. This parameter is unique to the formulation of the ODA framework, and hence is not relevant to other methods.

The results that the accuracy values *initially* improve with values of m across all the studied conditions. However, beyond a certain value of m, the accuracy decreases. This suggests that there is an optimum number of training samples to be selected for a given number of labeled test samples, which provides best classification over the test data. This brings to light a very interesting phenomenon associated with transfer learning or domain adaptation - that more is always not better, since there is a fundamental difference in distributions between training and test data. Detecting this number automatically can be a potential future work.

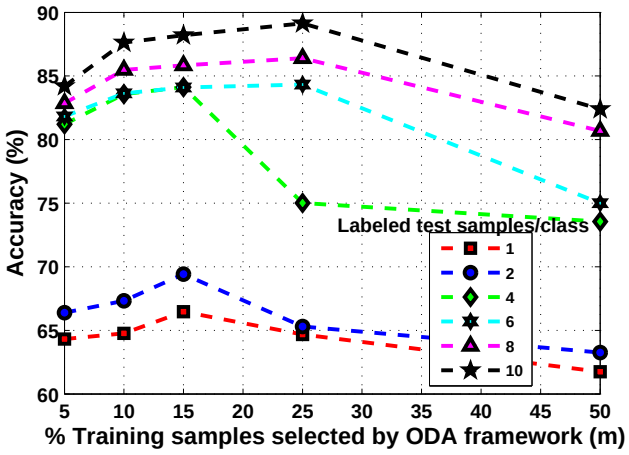


Figure 3.9: Accuracy(%) obtained by ODA vs % Training samples selected (m) with varying number of labeled test samples/class (x-axis represents the percentage training samples selected from a total of 2500 samples).

3.5 Topology Preserving Domain Adaptation (TPDA)

Spectral and amplitude variations in surface myoelectric signals (SEMG) are analyzed to determine the fatigue state of a muscle. But variations in the spectrum and magnitude of myoelectric signals across subjects cause variations in both marginal and conditional probability distributions in the features extracted across subjects, making it difficult to model the signal for any automated signal classification. However one can observe that the manifold of the multidimensional SEMG data have an inherent similarity as the physiological state moves from no fatigue to fatigue state. This method exploits this specific feature of the SEMG data and develops a domain adaptation technique that is based on intrinsic manifold of the data preserved in a low dimensional space, thus reducing the marginal probability differences between the subjects, followed by an instance selection methodology, based on similar conditional probabilities in the mapped domain. The proposed method provides significant improvement in classification accuracies compared to cases without any domain adaptation methods and also compared to other state-of-the-art domain adaptation methodologies.

3.5.1 Overview

It is observed that SEMG data collected over a fatiguing exercise from different individuals shows variations in both conditional and marginal distributions. In spite of these differences, the SEMG

data shows specific topological patterns that is consistent across different individuals. This pattern is illustrated in Figure 3.10.

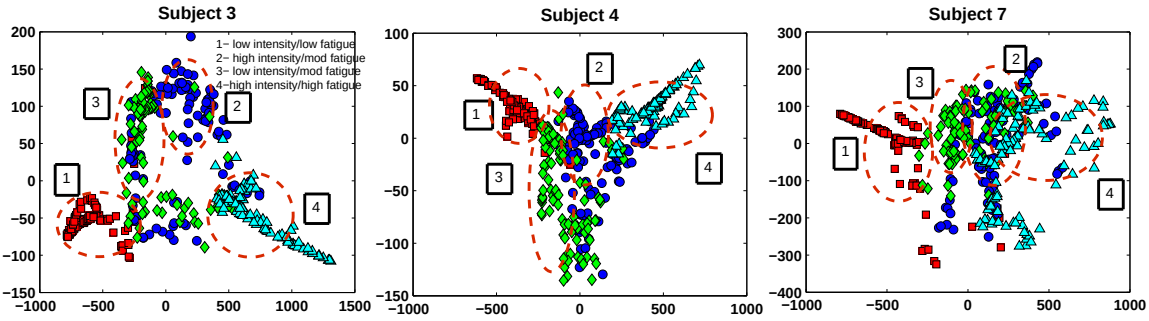


Figure 3.10: Three sample subjects (subjects 3, 4, 7) with four classes (four physiological stages) in our SEMG data set: Projected using Isomap based topology preserving methodology

The developed method exploits this similarity across subjects and uses a domain adaptation methodology that preserves the topology of the input data distribution, to map the data from multiple subjects into a common domain. This process minimizes the differences in the marginal probability distributions across the subjects. It is observed that the SEMG data as shown in Figure 3.10 has significant differences in conditional probabilities across subjects, having conflicting conditional probabilities in several cases. The second part of the framework reduces this difference through an instance selection technique. To summarize the proposed framework is dividd into two parts. The first part learns a new low dimensional feature space using a nonlinear dimensionality reduction technique called ISOMAP [45], which preserves the topology of the input data distribution and the second part performs instance selection in the mapped domain, based on conditional probability similarities between the distributions. Thus the proposed approach tries to reduce both marginal and conditional probability differences between the distributions.

3.5.2 Isomap based Domain Adaptation

Isomap [45] is a manifold learning technique that extends the traditional multi-dimensional scaling by incorporating the geodesic distances imposed by a weighted graph. The Isomap algorithm takes as input the distances $d_(i,j)$ between all pairs x_i, x_j from N data points in the high-dimensional input space X, measured in the standard Euclidean metrics and outputs coordinate vectors in a D -dimensional Euclidean space Y that best represents the intrinsic geometry of the data. The algorithm has three main steps. The first step constructs the neighborhood graph G over all the N data points by connecting points x_i and x_j. The weights

associated with the edges connecting two points x_i and x_j is set to the Euclidean distance between these points $d_x(i,j)$. The second step computes the shortest paths between any two points by initializing $d_G(i,j) = d_x(i,j)$ if x_i, x_j are linked by an edge, $d_G(i, j = \inf$ otherwise. Then for each value of $k = 1, 2, \cdots, N$ in turn, replace all entries $d_G(i,j)$ by $\min\{d_G(i,j), d_G(i,k) + d_G(k,j)\}$. The final matrix $D_G = \{d_G(i,j)\}$ would contain all the shortest path distances between all pairs of points in G. Finally, the third step applies classical MDS (multi dimensional scaling) to the graph matrix D_G and constructs an embedding of the data in a D-dimensional Euclidean space Y that best preserves the manifold's estimated intrinsic geometry. The coordinate vectors y_i for points in Y or the mapped features in the new domain are obtained by minimizing the cost function:

$$\phi(Y) = \sum_{ij} (\|x_i - x_j\| - \|y_i - y_j\|)^2 \tag{3.27}$$

where $\|x_i - x_j\|$ is the Euclidean distance between the high-dimensional data points x_i and x_j and $\|y_i - y_j\|$ is the Euclidean distance between the corresponding low dimensional data points y_i and y_j. Thus the low dimensional representation of the data in isomap preserves the geodesic distances in the input data. Other classical dimensionality reduction algorithms such as PCA or MDS do not detect the inherent manifold in the data. PCA finds a low dimensional representation that best preserves the variance of the data in the high dimensional input space and MDS finds an embedding that preserves the interpoint Euclidean distances. Isomap algorithm uses the MDS framework to preserve the geodesic distances as per the manifold of the high dimensional input data.

The proposed algorithm uses isomap to project the training and test data from multiple subjects to a common domain and then utilizes a K nearest neighbor based methodology to select instances from the training data which have similar conditional probability as the test data. Since the proposed domain adaptation algorithm addresses conditional probability as well, it requires a few labeled samples from the target domain data D^T. The training data D^S consists of data from multiple subjects and the test or target domain data (data from a subject under test) consists of some labeled data D_l^T and lots of unlabeled data D_u^T such that $D^T = D_l^T + D_u^T$. The main steps of the proposed approach are as follows:

- Step 1: Compute the low dimensional projection of the available labeled test data D_l^T using Isomap methodology.

- Step 2: Use the mapping to project the training data D^S and unseen unlabeled test data D_u^T as well, into the same mapped space.
- Step 3: In the mapped domain, compute the Euclidean distance from each labeled test data in D_l^T to each of the data points in the training data D^S belonging to the same class.
- Step 4: In the mapped domain, sort the distances in increasing order of the value and select k nearest points from the training data D^S for each of the data samples in D_l^T to form $D^S_{selected}$.
- Step 5: Learn a classifier on the mapped selected training data $D^S_{selected}$ and mapped labeled target domain data D_l^T and compute the labels of the mapped test data D_u^T.
- Step 6: Compute new low dimensional projection again, using the selected data from training domain i.e.$D^S_{selected}$ and labeled test data D_l^T in the original space, and obtain the new mapping.
- Step 7: Go back to Step 2, for N number of iterations.
- Step 8: Compute the class declared majority number of times in N iterations for each of the D_u^T and assign the same to the data instance.
- Compute classification accuracy of the D_u^T.

3.5.3 Experiments and Results

The effectiveness of the proposed Topology Preserving Domain Adaptation (TPDA) methodology is evaluated by comparing its results with four baseline method SVM-C, SVM-M, SMA, TSVM (Transductive SVM) and with three recently published domain adaptation methods; namely Transfer Component Analysis (TCA), Locally Weighted Ensemble (LWE), and Kernel Ensemble (KE). The details of each of these techniques are provided in Section 8. The results were also obtained with PCA based domain mapping followed by instance selection methodology as performed in TPDA, referred as PCA-DA. The definition of other methodologies are as follows: SVM-C refers to *all but one* method where in the training data comprises of data from all seven subjects and the test data is the data from the eighth subject. SVM-M, refers to *majority voting* based ensemble framework. The class y assigned to each unlabeled test data x is $\max_y NV(y|x)$ where $NV(y|x)$ is the number of votes given for class y for a particular test sample x by the seven auxiliary sources. SMA refers to *Simple Model Averaging*, which provides equal weight to all the

Table 3.10: Comparative Performance of Proposed Method (TPDA) on SEMG data - Accuracy (%)

Test Sub	SVM-C	SVM-M	SMA	TSVM	TCA	LWE	KE	PCA-DA	TPDA
1	70.76	33.9	44.96	49.09	45.15	67.44	71.85	66.81	82.12
2	43.69	50.76	44.61	55.68	68.93	77.54	74.62	59.46	75.76
3	50.11	56.85	56.84	65.09	56.78	75.55	74.79	65.79	83.96
4	59.65	47.93	49.67	56.98	52.68	81.22	69.35	78.02	81.75
5	40.37	44.79	50.15	62.5	60.15	52.48	73.44	67.29	83.20
6	59.21	61.45	60.33	71.32	76.92	65.77	73.92	66.08	82.32
7	47.13	46.91	45.76	60.73	55.64	60.32	77.97	72.46	83.33
8	69.85	64.53	74.46	68.55	67.24	72.81	79.48	68.34	84.13
Average	55.09	50.85	53.34	61.24	60.43	69.14	74.42	68.03	82.07
Std_Dev	11.53	9.88	10.24	7.25	10.14	9.56	3.20	5.39	2.69

classifiers learned on each auxiliary source domain in an weighted ensemble framework used to generate the label for the target domain data. SVM from LibSVM package was used as basic classifier for all these methods. TSVM is semisupervised method called Transductive SVM implemented using svmlight package.

For TCA, KE and the proposed method TPDA, 10% of the test data which is 6 to 7 samples per class, is made available for training purpose. Rest of the 90% of the test data was unseen test data. All the methods are tested on the same pool of unseen unlabeled test data. The accuracies are computed in a subject independent manner. In a particular experiment a subject data is considered as test data and the training data comprises of the data from the rest of the seven subjects. TCA was performed using linear kernel. The neighburhood parameter used in K-Isomap and also in the KNN based instance selection algorithm was set to 12. The reduced dimension was kept as 6 for TCA, KE, TPDA and PCA-DA. These parameter values were selected on the basis of best performance as a result of 5 fold cross validation over the test training data. All the results presented are average over 10 rounds of execution with different sets of random data selected as labeled test data for each test subject.

Results and Discussion

Classification accuracy was used as metrics for performance evaluation. The results shown in Table 3.10 are obtained by implementing the methods SVM-C, SVM-M, SMA, TSVM, TCA, LWE, KE and the proposed TPDA methodologies. The first column of the Table 3.10 indicates the subject data under test. The training data consists of the data from rest of the seven subjects. The results show that the different weighting schemes for SVM, namely SVM-C, SVM-M and SMA result in poor performance. All these schemes pertain to the case where brute force transfer of knowledge is done causing negative transfer [70]. It is also observed that results obtained by

implementing TCA which addresses only marginal probability differences are comparable to TSVM, which does not perform any domain adaptation, instead employs a label propagation technique to determine the labels of the target domain data. Results show that LWE which addresses conditional probability differences only between the domains gave better results than TCA and KE for test subjects 2,3 and 4. LWE gave better results than TPDA for subject 2 and comparable for subject 4. These results show that considering conditional probability difference is paramount for addressing the distribution difference across subjects in SEMG data. It is observed that the overall or average classification accuracies obtained when both marginal and conditional probabilities are addressed by methodologies KE and TPDA are better than any of the other methodologies implemented. However TPDA outperforms KE for all cases with an average performance improvement of around 7.6%. PCA based domain adaptation followed by instance selection, known as PCA-DA, was also implemented, for comparison with the proposed methodology, as PCA is by far the most popular linear technique. The results show that for SEMG data, preserving input data topology, significantly improves the average subject independent classification accuracies.

Thus results show that TPDA provides a gain of 21% to 32% over the methods with out any domain adaptation (SVM-C, SVM-M, SMA and TSVM) and 8% to 22% over the state-of-the-art domain adaptation methods (TCA, LWE and KE). We also observe that standard deviation of the results across the subjects is minimum for TPDA. This also shows that the proposed method is able to address better the variations in distributions across the subjects and present a less variant common representation for the SEMG data.

3.6 Feature Selection based on Robustness to Subject based Variability

This dissertation addresses the subject based variabilities in time series physiological data from SEMG sensors and presents a learning based framework that monitors fatigue across subjects. Feature selction is key to developing an efficient learning model. This section presents a feature selection method based specifically on robustness to subject based variability.

3.6.1 Overview

The performance of any automated classification tool depends on the discriminative power of features extracted from the data. Hence, if we want to have better generalization across subjects, we need to base our classification on features which are not only highly discriminative across different classes but also robust to subject based variability. This dissertation presents here a metric to measure subject based variability and propose a new feature selection method that is scalable to the number of subjects as well as to number of classes.

3.6.2 Cost Function for Class and Subject based Variability

ALGORITHM 3: Proposed Feature Selection Method

Feature Selection based on Robustness to Subject based Variability
Input: D, S_{labels}, P_{labels}
Output: $F_{selected}$
for each feature i **do**
  for each phase p **do**
    calculate ICC_p^i from Equation 4.5
  end for
  for each subject s **do**
    calculate ICC_s^i from Equation 3.29
  end for
end for
Compute $mpICC^i$=Mean (ICC_p^i) across all phases
Compute $msICC^i$=Mean (ICC_s^i) across subjects
% Select features with minimum $mpICC$ ***and maximum*** $msICC$
$arg_f\ max\ [(1 - mpICC^f) \times msICC^f]$ over all features f

In order to develop a subject independent measurement framework, we need to select those features or parameters which are less susceptible to subject based variation. So the proposed feature selection method not only measures the discriminative power of the features across classes or phases of physiological status as in our case with SEMG data, but also measures the subject based variability within each class or physiological phase. The high class based variability is a necessary condition for identifying different phases by the generalized classification framework and low subject based variability is an essential requirement of the

generalized measurement framework so as to classify phases or classes accurately across subjects. The variability metrics used in the proposed algorithm are defined as follows:

Definition: ***Subject based variability*** *is defined as the variation in a feature across subjects under similar conditions such as for the same activity or physiological status.*

Intraclass correlation coefficient (ICC) is a statistical tool that is used to measure fraction of the total variance that is due to variance between groups. This method uses ICC to measure subject based variability for a particular feature for a particular class or phase. When measuring the subject based variability of a particular feature over a specific phase, the groups are defined as feature values of a particular phase for different subjects. And the variance across groups denoted by each subject, for a particular feature f over a specific phase p are defined by the equation4.5

$$ICC_p^f = \frac{\sigma_{sb}^2(f,p)}{\sigma_{sb}^2(f,p) + \sigma_{sw}^2(f,p)} \quad p = 1 \cdots P \tag{3.28}$$

where P is the number of phases or classes, $\sigma_{sb}(f,p)$ is the variance in the feature f between subjects and $\sigma_{sw}^2(f,p)$ is the variance within subjects over a phase p.

Definition: ***Phase based variability*** *is defined as the variation in a feature across phases for the same subject.*

Similar to subject based variability, ICC is also used to measure phase based variability of a particular feature. Hence, the equation 3.29 defines phase based variability for each feature across each subject.

$$ICC_s^f = \frac{\sigma_{pb}^2(f,s)}{\sigma_{pb}^2(f,s) + \sigma_{pw}^2(f,s)} \quad s = 1 \cdots S \tag{3.29}$$

where S is the number of subjects, $\sigma_{pb}(f,s)$ is the variance in the feature f between phases and $\sigma_{pw}^2(f,s)$ is the variance within phases for a subject s. The variance within groups $\sigma_{pw}(f,s)$ and the variance between groups $\sigma_{pb}(f,s)$ were obtained using the within group Mean Square Error ($MS_{pw}(f,s)$) and between group Mean Square Errors ($MS_{pb}(f,s)$) obtained by Kruskal-Wallis test on the data. The respective ICC_s^f is then computed as per equation 3.30, where n is the number of samples.

$$ICC_s^f = \frac{MS_{pb}(f,s) - MS_{pw}(f,s)}{MS_{pb}(f,s) + (n-1)MS_{pw}(f,s)} \quad s = 1 \cdots S \tag{3.30}$$

ICC_p^f is computed similarly using the Mean Square errors obtained by Kruskal-Wallis test on the data across the subjects. Kruskal-Wallis test was performed as the test of normality using Kolmogorov- Smirnov's normality test gave p < 0.05.

The cost function used for most of the state-of-the-art feature selection algorithms such as boosting based, area under receiver operating characteristic curve (AUC), ReliefR etc tries to

maximize the discriminative power of the features with respect to different classes. In the proposed feature selection algorithm, another dimension is added to the cost function, i.e. robustness to subject based variability. Hence the proposed algorithm maximizes a cost function:

$$\max_{f} [(1 - mpICC^{f}) \times msICC^{f}] \tag{3.31}$$

where $mpICC^{f}$=Mean(ICC_{p}^{f}) across all phases and $msICC^{f}$=Mean (ICC_{s}^{f}) across all subjects, also shown in step 11 of Figure 3. The cost function thus not only takes into account high discriminative power of the features across the phases or classes but also tries to bias the selection towards features with low subject based variability. Figure 3 summarises the proposed feature selection method. Input to this method is the complete $n \times m$ feature matrix D where n is number of feature vectors belonging to all the S subjects and m is the dimension of feature vectors, S_{labels} i.e labels pertaining to each subject and labels pertaining to each class or phase i.e. P_{labels}. Output of this method is $F_{selected}$ i.e. features in decreasing order of cost function value.

3.6.3 Experiments and Results

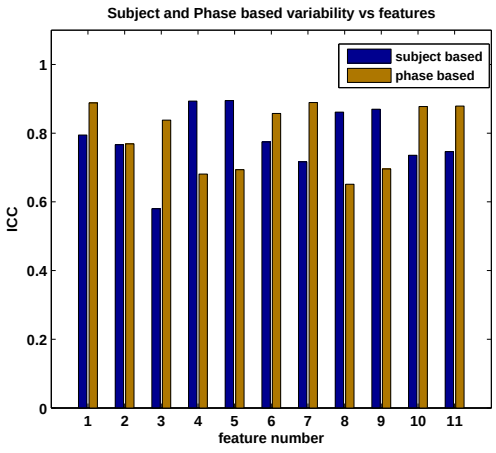


Figure 3.11: Subject and Phase based variability(ref Table 4.2 for feature names)

The proposed and state-of-the-art feature selection methods were applied on a training dataset (25%). To measure the performance of these feature selection methods subject independent classification accuracies were computed on the remaining of 75% of the data. In order to remove bias towards any particular learning algorithm, classification accuracies were computed using different learning algorithms such as AdaBoost, SVM, HMM and KNN. Significant parameters for these learning algorithms were set as 100 iterations for AdaBoost, standard

deviation of radial basis function for SVM was set as 0.5 , nearest neighbor number for KNN was kept at 5 and window size used was 10 for HMM. These parameters were cross validated on a set aside training set. The results were computed across all the subjects following an all but one strategy i.e. training data did not include any data from the test subject. Software package 'weka' was used to obtain the selected features as per the state-of-the-art feature selection methods.

The subject and phase based variability computed for each of the eleven features as per proposed algorithm are shown in Figure 3.11. The variability values along with their cost function computed as per algorithm in Figure 3 are presented in Table 4.2.

Table 3.11: Feature ranking based on subject and class or phase based variability (proposed method)

No.	Features	ICC based on subjects	ICC based on phases	Cost function
3	**Peak of LE**	0.58	0.84	**0.36**
7	**RMS of burst**	0.72	0.89	**0.25**
10	**Energy of burst**	0.74	0.88	**0.23**
11	**Spectral Entropy**	0.75	0.88	**0.22**
6	Max peak value	0.77	0.86	0.20
1	Mean of LE	0.79	0.88	0.18
2	Variance of LE	0.76	0.76	0.18
9	Mean Frequency	0.87	0.90	0.12
8	Median Frequency	0.86	0.65	0.09
10	Zero crossing	0.89	0.68	0.08
11	Ave time bet peaks	0.9	0.69	0.07

Table 3.12: Feature Selection Method Vs Features Selected

Feature Selection Method	Features Selected
ReliefR	mean LE, zero count, RMS, median frequency
AdaBoost	mean LE, max peak, median frequency
AUC	max peak, RMS, zero count, entropy
Infogain	mean LE, RMS, energy, entropy
Factor Loadings	peak LE, zero count, RMS, energy
Proposed Method	**peak LE, RMS, energy, entropy**

The proposed feature selection algorithm ranks the features as per their cost function value (Equation 3.31). Minimal improvement in subject independent classification accuracies were observed after first four features, in the rank list, hence the first four features were selected for the framework. Table 3.12 shows the first four significant features selected by the state-of-the-art feature selection techniques along with the features selected by the proposed method. Eight out of eleven features are selected as most discriminative by at least one feature selection method.

The proposed method selects Peak of LE, RMS, Energy, and Spectral entropy. All these four features have relatively low subject based variability and high phase based variability. The

results show that though Mean of LE has very high subject based variability, it gets selected by 3 out of 5 existing feature selection techniques, as it has high phase based variability. Similarly, though zero crossing count and median frequency have very high subject based variability they got selected in 3 out of 5 and 2 out of 5 existing feature selection techniques respectively. Hence, it can be concluded that the existing feature ranking or selecting techniques do not necessarily consider subject based variability as a key measure.

Table 3.13 presents the subject dependent and subject independent classification accuracies of the different feature selection methods, based on their respective selected features, on the same test dataset. The Table 3.13 also presents the classification accuracies for the case where all features are used, for comparison.

Table 3.13: Subject Independent Classification Accuracy Vs Feature Selection Method

Feature Selection Method	AdaBoost	SVM	HMM	KNN
ReliefR	67.25	59.15	40.68	59.88
AdaBoost	62.54	60.02	45.9	60.61
AUC	66.13	65.76	68.51	60
Infogain	66.27	63.48	68.82	60.52
Factor loadings	65.76	62.1	68.25	58.41
Proposed Method	**71.5**	**73.67**	**80.65**	**76.09**
All features	60.14	61.28	65.52	58.25

From these results one can observe that:

- The proposed feature selection technique based on robustness to subject based variability gave an improvement of 11.45% for AdaBoost, 12.39% with SVM, 15.13% with HMM and 17.84% KNN over the case with all features with 64% reduction in dimensionality. Hence a proper selection of features can help addressing the subject based variability in developing generalized framework for physiological data.
- The proposed feature selection method gave about 10%-18% higher subject independent classification accuracy compared to that obtained when features used based on existing state-of-the-art feature selection techniques.
- The low subject independent accuracies obtained by existing feature selection techniques show that none of them specifically consider robustness to subject based variability.

Our study reveals the existing feature selection techniques do not necessarily or specifically consider variation across subjects. The unsupervised techniques select features based on redundancy and irrelevance and the supervised techniques such as ReliefR method [38] increases weights of those features which are more similar to another data point of same class in the feature domain and decreases the weight of those features which are similar to the data point of dissimilar class. AdaBoost gives more weight to the features which get selected as decision boundaries more often. Area under Receiver Operating Characteristics Curve (AUC)[38] select features based on better classification between the given classes. Infogain methodology selects features based on minimizing the weighted average impurity measure (entropy) and factor loadings select features based on high correlation to major factor components obtained by doing a factor analysis.

Chapter 4

SUBJECT INDEPENDENT GRADING FRAMEWORK

Besides proposing person adaptive classification frameworks based on domain adaptation methodologies, this dissertation also proposes a subject independent grading framework for grading electromyogram features related to fatigue, during a submaximal fatiguing contraction. This chapter presents the experimental setup, experimental procedure, the grading framework and the results obtained.

4.1 Methods

The proposed algorithm was applied to intramuscular EMG signals collected from 12 hand muscles during a sustained isometric fatiguing contraction and compared to the results from standard amplitude and spectral measures. The data are from a published report [73], which investigated the influence of fatigue on EMG-EMG coherence across hand muscles.

Subjects

Eight adults [5 men, 3 women; 27 $\pm$ 6 yrs.] participated in the study. Subjects reported being without any neurological disorders or musculoskeletal injuries of the hand and were right handed. Subjects gave written informed consent before participation in the study, and all experimental procedures were approved by the Institutional Review Board at Arizona State University.

Experimental Setup

Subjects sat in an adjustable chair facing a computer monitor. The right forearm was placed on a flat rigid platform with the wrist and hand in a semi-supinated and neutral position, respectively. Movement of the forearm and wrist was prevented by rigid dowel restraints inserted into the platform around the forearm and wrist. The hand was positioned in a 3-digit (thumb, index, middle finger) grasp posture for the fatiguing contractions as shown in Figure 4.1.

Force Measurement

The isometric force of the distal pads of thumb, index and middle fingers were each measured with six-dimensional force/torque transducers (ATI Nano-17/S1 Apex, NC) mounted on a manipulandum. They measure force in all three spatial dimensions and torque about all three spatial axis. The contact surfaces of the sensors for the index and middle fingers were 3 cm apart vertically (center-to-center), and 8 cm apart horizontally from the surface of the force/torque sensor for the thumb.

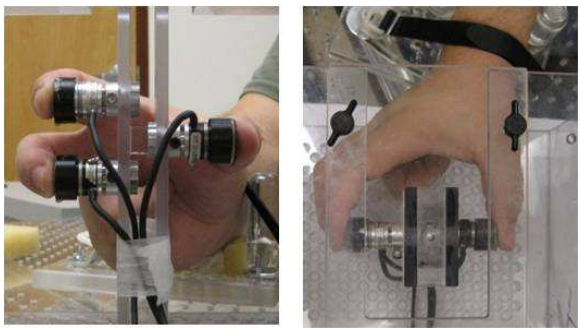


Figure 4.1: Experimental Setup: hand positioned in a 3-digit grasp posture.

EMG Measurement

Intramuscular EMG was recorded with electrodes comprised of one $50\mu m$ stainless steel wire insulated with Formvar (California Fine Wire, Grover Beach, CA) from 12 muscles of the hand. Insulation was removed from the recording tip and the opposite end of wire. One electrode was inserted into the belly of each muscle via a 27-gauge hypodermic needle. Once the electrode was inserted the quality of the signal was optimized by changing the depth and angle of the needle. Electrode placement was further verified with electrical stimulation. The needle was then removed and the wire electrode remained in the muscle for the duration of the experiment. The muscles recorded and analyzed included 6 intrinsic and 6 extrinsic hand muscles (for details, see [73].

The EMG recordings were amplified 2000x and band-pass filtered between 3 Hz and 1 kHZ (Neurodata Acquisition System model 12, Grass Instruments, Warwick, RI). A reference surface electrodes (gold plated silver disc, Grass Instruments; West Warwick, RI) was placed on the skin over the radial styloid to serve as a reference electrode.

Data Acquisition

Force and EMG recordings were acquired and digitized at 2 kHZ with 12-bit A/D converter boards (PCI-6225, National Instruments, Austin, TX) displayed, and stored on computer using LabView v 6.0, National Instruments.

Experimental Procedure

Maximal Voluntary Contractions (MVC)

Subjects increased digit isometric grasp force of all three fingers from rest $(\sim 1N)$ to maximum over a period of 3 s and then maintained the maximum force for 6 s. Subjects performed three MVC trials with a minimum of 3 min of rest between trials. The trial with the greatest total sum of

digit forces was used as the reference value to compute the target force for the submaximal fatiguing contraction.

Submaximal Fatiguing Contraction

Subjects performed a sustained isometric contraction with the thumb, index and middle fingers at a total force of 40 % of MVC. Subjects began the task by increasing digit total force from rest $(\sim 1N)$ to the target force. Data collection began when the subject was able to maintain the target force for $\geq$ 3 s. Visual feedback of the target force was provided as a horizontal line on the computer for the duration of the contraction. The subject's total digit force was also displayed as a red trace in real-time. The contraction was ended when the subject could not maintain the total force within $\pm 10\%$ of the target force or they were unable to maintain the same hand and forearm posture despite strong verbal encouragement for 3 s. However, all subjects failed due to an inability to maintain the target force.

Data Analysis

A set of nine time and frequency domain features were extracted from the interference EMG, after down sampling to 1 kHz (Table 4.1). The linear envelope of the rectified EMG signal was obtained

Table 4.1: Time and Frequency domain features extracted from EMG signal

	Time Domain Features
1	Mean of linear envelope (LE)
2	Peak of linear envelope (LE)
3	Zero crossing count [22]
4	Maximum peak amplitude
5	Root mean square (RMS)[50, 26]

	Frequency Domain Features
6	Median frequency [50, 56]
7	Mean frequency [6, 26, 91]
8	Spectral entropy [72]
9	Spectral energy [37]

with a Butterworth lowpass filter of 3rd order with a cutoff frequency of 2.5 Hz. Features were derived from running windows of 1000 time samples with a 50% overlap i.e., features were derived every 500 ms. The spectral components were obtained using a 1024 point DFT, yielding a spectral resolution of 0.97 Hz. We have used multiple time and frequency domain features to extract more information about the amplitude and spectral characteristics of the EMG signal. Mean and peak of linear envelope, maximum peak amplitude, root mean square [50, 26] and spectral energy [37] provided information about the amplitude of the EMG signal, where as zero crossing count [22], median frequency [50, 56], mean frequency [6, 26, 91] and spectral entropy [72] provided information about the frequency content of the EMG signal. A brief description of the features is as follows.

Mean and peak of linear envelope are mean and peak voltages respectively, of the 2.5 Hz linear envelope obtained as described before. Zero crossing count is obtained by counting the number of times the EMG voltage changes sign in a window of 1000 time samples or 1 secs. The maximum peak amplitude is the maximum voltage obtained in every window. Median frequency (mf), which is a measure of skewness in power spectrum, is defined as the frequency that divides the power spectrum into two equal parts and was obtained as follows:

$$\sum_{i=1}^{mf} I_i = \sum_{i=mf+1}^{n} I_i. \tag{4.1}$$

where n is the number of frequency bins, which is equal to the bandwidth of the signal, i.e., 500. and I_i is the amplitude or intensity of spectrum at i-th bin. Mean Frequency defined as the normalized, first order spectral moment was calculated as the sum of the product of the spectrogram amplitude and the frequency, divided by the total sum of spectrogram amplitude or intensity, as follows:

$$Meanfrequency = \frac{\sum_{i=1}^{n} I_i \times f_i}{\sum_{i=1}^{n} I_i}. \tag{4.2}$$

where f_i is the frequency of spectrum at i-th bin of n frequency bins and the spectral energy was obtained as follows:

$$Spectral\ energy\ = \frac{1}{n}\sum_{i=1}^{n} I_i^2. \tag{4.3}$$

The spectral entropy, defining the entropy in the power spectrum was obtained as below:

$$Spectral\ entropy\ = -\sum_{i=1}^{n} p_i \times log(p_i). \tag{4.4}$$

where $p_i = \frac{I_i}{\sum_{j=1}^{n} I_j}$.

4.2 Subject-independent Model for Grading EMG Features Related to Fatigue

Even though EMG amplitude and spectral features vary differently across subjects during a contraction, the relation among the features tended to be subject-independent. Therefore, the framework was based on the relation between multiple features of the EMG signal. Principal Component Analysis on the nine features listed in Table 4.1 from all twelve muscles over eight subjects revealed that the first two components captured $98.71 \pm 1.1\%$ of the total variance, therefore factor analysis [62] was performed on these nine features using two factors. Factor analysis was used to describe variability among the features to define a potentially lower number of variables i.e. latent factors. These experiments revealed two latent factors, which were used for the analysis of the reference framework.

Because the aim was to develop a generalized subject-independent method for grading the changes in the EMG signal during a fatiguing contraction, the variability of each feature and latent factor, between subjects, was measured with an Intra Class Correlation (ICC) analysis. The ICC coefficient measures the fraction of the total variance that is due to between group variance (groups are values of a particular feature or latent factor over all twelve muscles belonging to different subjects) as follows:

$$ICC^f = \frac{\sigma^2_{sb}(f)}{\sigma^2_{sb}(f) + \sigma^2_{sw}(f)} \tag{4.5}$$

where $\sigma^2_{sb}(f)$ is the variance in the feature or latent factor f between subjects and $\sigma^2_{sw}(f)$ is the variance within subjects. Between-subjects and within-subjects variances of each feature and latent factors were obtained using Kruskal-Wallis one way analysis of variance [57] across multiple subjects.

Since ICC values capture the between subject variance, with respect to the total variance in a feature, higher ICC values indicate greater variation across subjects. The individual EMG features had greater ICC values compared to the latent factors (Table 4.2). Despite the large variation in feature values across subjects, the relation between the features and the two latent factors did not differ significantly across subjects, as indicated by significantly lower ICC values of the latent factors, shown in bold in Table 4.2.

Table 4.2: Subject based variability for features and latent factors

No.	Features/ Latent factors	ICC based on subjects
1	Mean of LE	0.738
2	Peak of LE	0.716
3	Zero crossing count	0.929
4	Max peak value	0.839
5	Root mean square	0.714
6	Median Frequency	0.914
7	Mean Frequency	0.914
8	Spectral Entropy	0.775
9	Spectral Energy	0.737
10	Latent Factor 1	**0.076**
11	Latent Factor 2	**0.098**

The latent factors provide a reference model to monitor changes in the EMG signal of a subject with respect to a reference group of subjects in real time. Therefore, these latent factors were used to derive our automatic algorithm. Analysis of the relation of the EMG features with the two latent factors over all subjects and muscles, revealed that features related to EMG amplitude were strongly correlated with Latent Factor 1 while the features related to frequency content of the

Table 4.3: Correlation coefficients: Features Vs Latent factors (Bold values indicate significant correlation ($p < 0.05$) coefficient.)

Sl.no	Features	Latent Factor 1	Latent Factor 2
1.	Mean of LE	**0.98**	0.02
2.	Peak of LE	**0.87**	-.01
3.	Zero crossing count	-0.01	**0.90**
4.	Max peak value	**0.81**	0.09
5.	Root mean square	**0.98**	0.05
6.	Median Frequency	0.02	**0.95**
7.	Mean Frequency	0.01	**0.96**
8.	Spectral Entropy	**-0.79**	0.31
9.	Spectral Energy	**0.97**	-0.01

EMG signal were strongly correlated with Latent Factor 2 (Table 4.3). Corresponding significant coefficients are in bold for reference.

Framework for grading of EMG features related to Fatigue

The main requirement of the subject-independent framework was to obtain a generalized or reference mapping of features extracted from the EMG signal, which could be used to map and quantify the progression of multiple features continually in real-time, during a fatiguing exercise of a subject, without the need to calibrate the algorithm with subject specific data.

The reference mapping, in the proposed framework, was obtained through the latent factors (Latent Factor 1 and 2). The mapping was then used to monitor the changes in EMG features on a continuous scale from 0-1 of a test subject in real time. The states 0 represent the beginning of the contraction whereas the state 1 represents the estimated point of task failure with respect to the reference framework.

The proposed framework consisted of two major components: (1) developing the reference framework and (2) projecting a test subject's features into the reference framework. The reference framework was developed with the EMG features from seven subjects and cross validated on the remaining eighth subject, known as test subject, following a Leave-one-out strategy. As per this strategy, the reference data consisted of features from all seven subjects and the test data consisted of features from the remaining eighth subject referred to as the test subject. Details of the framework are provided in Figure 4.2. Table 4.4 defines notations used in Figure 4.2.

Development of the reference framework

A set of nine features (see section 4.1), were extracted from each window of 1000 time samples of the pre-processed EMG signal of a subject k, with 50% overlap. Pre-processing of the raw EMG signal involved de-noising using a high pass filter of 20 Hz. In step 2 of the framework, as shown

Table 4.4: Notation I

Notation	Explanation
K	Number of subjects in reference set
$S_k(t)$	Input time series signal for a subject k in reference set
$S_T(t)$	Input time series signal for test subject T
N	Total number of windows in signal
D	Feature dimension
F	Feature matrix
L	Reference factor loadings
FS	Factor scores

Algorithm 1: Subject-Independent Framework for Grading EMG Features.
Input: $S_k(t)\ \forall 1 \leq k \leq K$ and $S_T(t)$
Output: L, EMG feature gradings of test subject.
1. Extract features from each window of 1000 samples of $S_k(t)$ to form $F_k^{N\times D}$.
2. Combine $F_k^{N\times D}$ of K subjects to form $F_{reference}$.
3. Obtain L and $FS_{reference}$ using factor analysis on $F_{reference}$.
4. Extract features from $S_T(t)$ to form F_{test}.
5. Obtain FS_{test} for test subject as follows:
$FS_{test} = (L^TL)^{-1}L^T(F_{test} - \overline{F}_{reference})$.
where $\overline{F}_{reference}$ is the mean feature vector of K subjects.
6. Obtain the projections of FS_{test} on latent factors 1 and 2.
$P1 = FS_{test}(:,1)$, $P2 = FS_{test}(:,2)$
7. Obtain 0 to 1 gradings with respect to the reference framework:
$P1_{grading} = \frac{P1-\min(FS^1_{reference})}{\max(FS^1_{reference})-\min(FS^1_{reference})}$.
$P2_{grading} = 1 - \frac{P2-\min(FS^2_{reference})}{\max(FS^2_{reference})-\min(FS^2_{reference})}$.
where $FS^1_{reference} = FS_{reference}(:,1)$ and
$FS^2_{reference} = FS_{reference}(:,2)$.

Figure 4.2: Subject-Independent Framework for Grading EMG features.

in Figure 4.2, features extracted from all K subjects were appended to form $F_{reference}$. Factor analysis was performed on $F_{reference}$, in step 3, to obtain the factor loadings L and the factor scores $FS_{reference}$. Factor analysis was performed with two principal components, using standard Matlab function 'factoran'. Factor loadings define the relation of each feature to the two latent factors (Latent Factor 1 and Latent Factor 2). Typical factor loadings obtained for a reference group of subjects is shown in Table 4.3. This provides a reference mapping for grading EMG signal from a test subject.

Projecting a test subject's features into the reference framework

The feature vectors extracted from EMG signal of a test subject in step 4 were mapped in the reference framework, using the equation defined in step 5 of Figure 4.2. The factor scores FS_{test} obtained in step 5 map the test subject features in the reference framework using the reference

factor loadings L. The latent factor 1 and latent factor 2 components of FS_{test} i.e., P1 and P2 respectively, obtained in step 6, provided a composite view of the changes in the features of the EMG signal from the test subject. These values were normalized with respect to maximum and minimum values of the respective components of the reference factor scores $FS_{reference}$ to generate a continuous grading in the range [0,1], in step 7. The grading on latent factor 2 were inverted to get an increasing value with a decrease in values of latent factor 2.

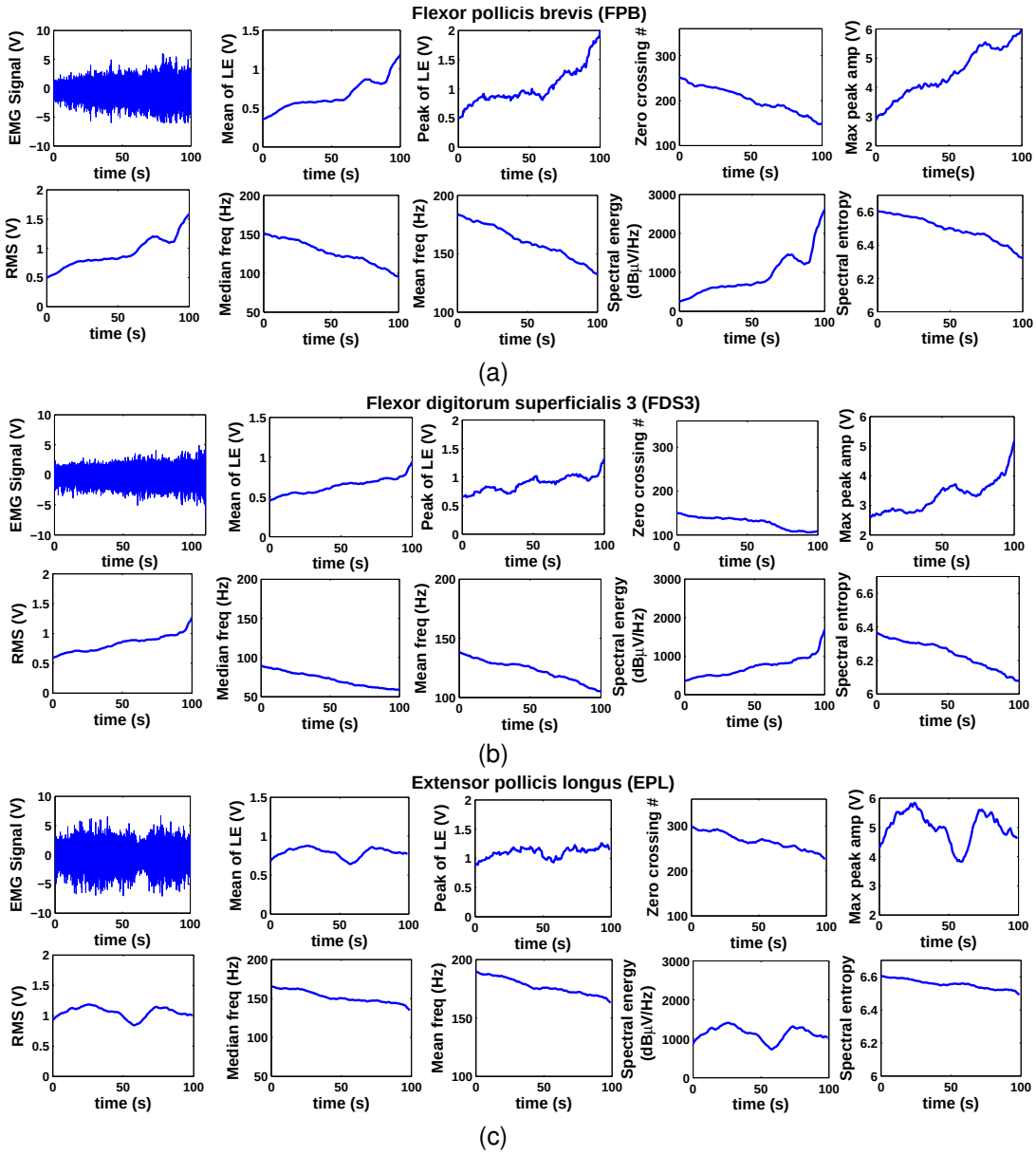


Figure 4.3: Time and frequency domain features during the fatiguing contractions for muscles (a) flexor pollicis brevis (FPB) (b) flexor digitorum superficialis 3 (FDS3) (c) extensor pollicis longus (EPL) for a representative subject (subject 1).

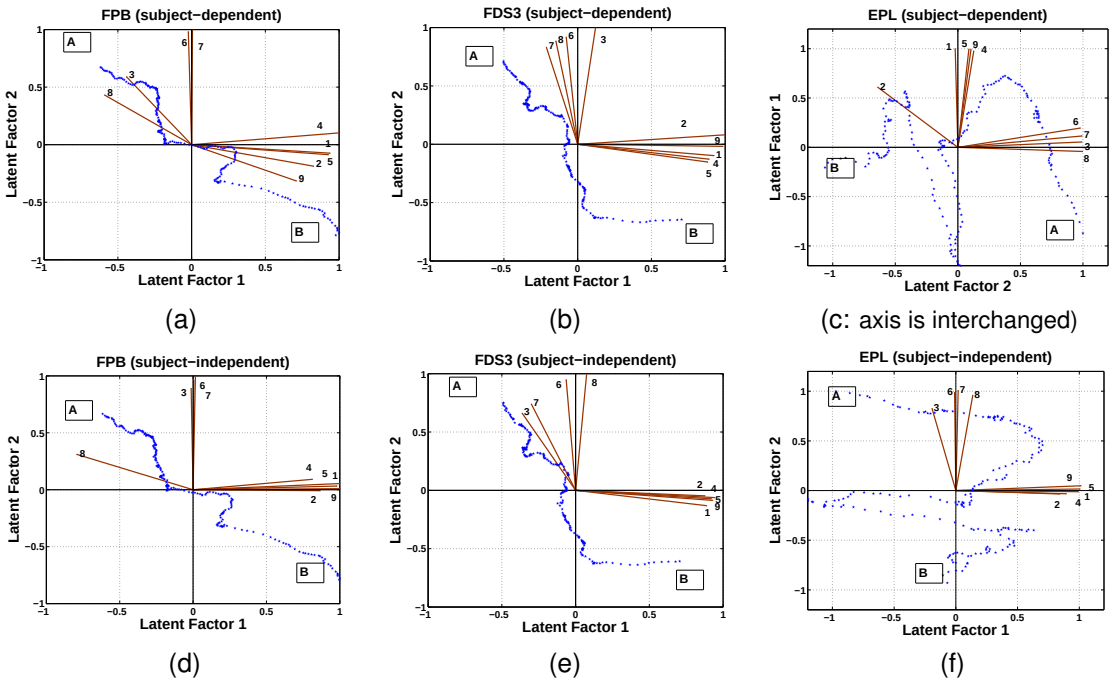


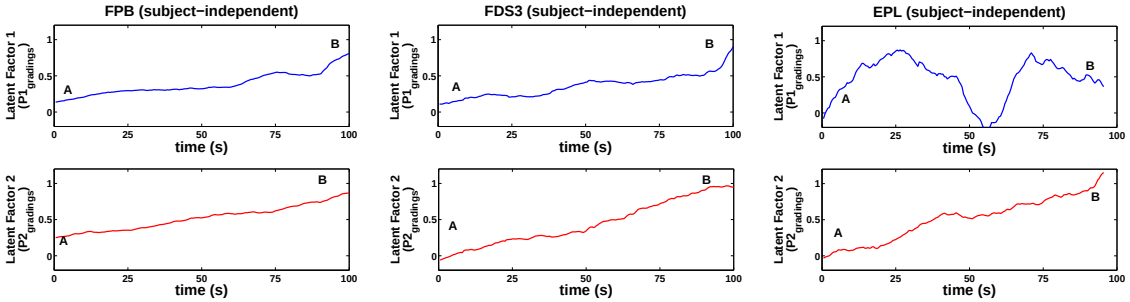
Figure 4.4: Subject 1 factor scores with respect to own i.e. subject-dependent framework and rest seven subjects features i.e. subject-independent framework, from beginning to end of contractions (A to B) for the muscles Flexor pollicis brevis (FPB) ((a) and (d)), Flexor digitorum superficialis 3 (FDS3) ((b)and (e)) and Extensor pollicis longus (EPL) (sub1) ((c) and (f)). Feature axis numbered as per Table 4.1. The angles between each feature axis and the two Latent Factors indicate the respective correlation between each feature and the Latent Factors. Smaller angle signifies higher correlation.

4.3 Results and Discussion

Figure 4.3 shows typical trends of the extracted features from three different muscles namely (a) Flexor pollicis brevis (b) Flexor digitorum superficialis 3 and (c) Extensor pollicis longus for a representative subject (subject 1). For the flexor muscles the values of amplitude related features such as mean of linear envelope (LE), peak of linear envelope (LE), maximum peak amplitude, root mean square (RMS) and spectral energy increased throughout the contraction and the values of frequency related features such as median frequency, mean frequency, spectral entropy and zero crossing count decreased as the power spectrum shifted towards lower frequencies. These trends are consistent with the recruitment of additional motor units and a decrease in the conduction velocity of action potentials along the muscle fibers as discussed in Section 1. The features of the extensor muscles did not always follow the same trend as those of the flexors especially for subject 1 shown in Figure 4.3 (c). However the framework was still able to

Table 4.5: Linear regression results on factor score distributions for subject-dependent and subject-independent frameworks

	subject_dependent		subject_independent	
Muscle	slope	intercept	slope	intercept
First dorsal interosseus (FDI)	-0.68 ± 0.18	-0.00 ± 0.01	-0.67 ± 0.16	-0.00 ± 0.01
First palmar interosseus (FPI)	-0.31 ± 0.12	-0.01 ± 0.02	-0.29 ± 0.14	-0.01 ± 0.02
Second dorsal interosseus (SDI)	-0.64 ± 0.13	-0.00 ± 0.02	-0.67 ± 0.11	-0.00 ± 0.01
Second palmar interosseus (SPI)	-0.55 ± 0.36	-0.00 ± 0.05	-0.51 ± 0.31	-0.00 ± 0.02
Abductor pollicis breves (APB)	-0.86 ± 0.46	-0.18 ± 0.51	-0.85 ± 0.57	-0.19 ± 0.53
Flexor pollicis breves (FPB)	-0.84 ± 0.10	-0.01 ± 0.09	-0.85 ± 0.12	-0.04 ± 0.09
Extensor digitorum 2 (ED2)	-0.49 ± 0.01	-0.00 ± 0.00	-0.51 ± 0.00	-0.00 ± 0.00
Extensor digitorum 3 (ED3)	-0.75 ± 0.12	-0.01 ± 0.12	-0.77 ± 0.31	-0.01 ± 0.11
Extensor pollicis longus (EPL)	-0.62 ± 0.24	-0.00 ± 0.00	-0.59 ± 0.25	-0.00 ± 0.00
Flexor digitorum superficialis 2 (FDS2)	-0.82 ± 0.31	-0.00 ± 0.00	-0.83 ± 0.28	-0.00 ± 0.00
Flexor digitorum superficialis 3 (FDS3)	-0.74 ± 0.14	-0.00 ± 0.01	-0.73 ± 0.11	-0.00 ± 0.02
Flexor pollicis longus (FPL)	-0.72 ± 0.13	-0.02 ± 0.01	-0.74 ± 0.12	-0.02 ± 0.01



(a) (b) (c)

Figure 4.5: Subject 1 EMG feature gradings with respect to reference framework for muscles (a) Flexor pollicis brevis (FPB) (b) Flexor digitorum superficialis 3 (FDS3)(c) Extensor pollicis longus (EPL) obtained by projecting respective subject-independent factor scores (A-B) in Figure 4.4 on both the latent factors.

derive the factor score distributions for this muscle in a subject-independent manner as shown in Figure 4.4(f).

Figure 4.4 shows the distribution of factor scores for the three muscles for a test subject (subject 1), obtained by the subject-dependent and subject-independent frameworks. Factor scores obtained by using factor analysis (standard Matlab function 'factoran') solely on subjects's own features are referred to as subject-dependent and the subject-independent factor scores are obtained based on the features of the other 7 subjects, as per algorithm described in Figure 4.2. The angle between the axis representing a feature and the axis representing the specific Latent Factor indicates the correlation between the feature and the Latent Factor. A greater correlation corresponded to a smaller angle between the two axis.

The start of contraction is denoted as (A) and the end of contraction as (B). As mentioned earlier in section 4.2 the amplitude related features correlated well with Latent Factor 1 and the frequency related features correlated well with Latent Factor 2. This corresponded to a shift of factor scores towards higher values for Latent Factor 1 and lower values for Latent Factor 2, respectively, during the duration of the contractions. This is in accordance with the observation that as fatigue progresses the amplitude of the EMG signal typically increases while the frequency of the signal shifts to lower values. Consequently, the factor scores for most cases, started in the upper left quadrant at (A) and ended in the fourth quadrant of the framework at the end of contractions (B).

To compare the factor scores obtained by the subject-dependent and subject-independent frameworks a linear regression using degree one polynomial was performed on the two factor score distributions. Table 4.5 shows the mean slope and intercept values obtained for both subject-dependent and subject-independent factor scores, for all the twelve muscles, averaged over eight subjects. Leave-one-out strategy, as explained in section 4.2, was used to obtain the subject-independent factor scores for each subject. The distribution of factor scores did not differ significantly whether the reference model was based on subject's own features (subject-dependent framework) or on the features of other 7 subjects (subject-independent framework). A multi-variable analysis of variance revealed that the means of the slope and intercept values obtained using subject-dependent and subject-independent frameworks did not differ significantly $(p > 0.05)$.

Note that for each of the three different muscles, the factor loadings for the subject-dependent frameworks shown in Figures 4.4 (a), (b) and (c) are very similar to the factor loadings of their respective subject-independent frameworks shown in Figures 4.4 (d), (e) and (f) respectively. This is due to the low subject based variability of the factor loadings shown in Table 4.2. Hence the distribution of the factor scores obtained by the subject-independent framework is similar to the distribution in the subject-dependent framework. This is also evident from the similarities in slope and intercept values in the two frameworks presented in Table 4.5. Also note that though the factor score distributions are different across the flexor (FPB and FDS3) and extensor muscle (EPL) but the factor loadings are very similar across both the muscle types.

Figures 4.5 (a), (b), and (c) show the gradings of the EMG features (associated with muscle fatigue) obtained as per the algorithm shown in Figure 4.2, by projecting factor scores

obtained in Figures 4.4 (d), (e), and (f) on both the latent factors for the three muscles, Flexor pollicis brevis (FPB), Flexor digitorum superficialis 3 (FDS3) and Extensor pollicis longus (EPL) respectively. The EMG feature gradings increased on a scale between 0 to 1 from the start of contractions defined by (A) to end of the contractions (B), during the progression of the fatiguing exercise for all three muscles. Please note that for the extensor muscle EPL (Figure 4.5 (c)), the $P1_{gradings}$ which capture the EMG amplitude, did not increase with fatigue up to point B. However, the $P2_{gradings}$ related to EMG spectral characteristics, increased from 0 at the start of contractions (A) to 1 at the end of contractions (B), signifying the importance of capturing changes in both EMG amplitude and spectral characteristics for grading features related to fatigue.

The subject independent framework presented here provides a practical way to track the changes in EMG features associated with muscle fatigue. The framework is based on hidden factors or latent factors derived from multiple EMG amplitude and spectral features. The goal of the proposed algorithm is to associate the beginning of the contraction and the time of task failure, in which there is significant level of muscle fatigue, to the changes in the multiple EMG features, in a subject-independent manner. This was shown in two primary ways; one, the factor scores for most cases, started in the upper left quadrant of the subject-independent framework and ended in the fourth quadrant of the framework (Figure 4.4) and secondly the gradings of the EMG features obtained from the projections of the factor scores on the Latent factor 1 and Latent factor 2 (Figure 4.5) increased in a scale from 0 to 1 with respect to the reference framework in either both or one of the Latent factors. So far this analysis cannot measure the extent of muscle fatigue. However, the framework provides a method to identify in a subject-independent fashion when the muscles are approaching a level of fatigue that would result in task failure. Another advantage of this framework is that it operates in real time.

Although surface EMG is more practical for clinical purposes, its current ability to record from individual hand muscles is limited. Therefore, intramuscular electrodes were used in this study to record the EMG activity of small hand muscles due to its higher selectivity and reduced risk of cross-talk. Nevertheless, studies using surface EMG or intramuscular EMG during sustained sub-maximal isometric contractions have revealed similar changes in the signal features, such as an increase in EMG amplitude and a left shift in frequency content of the power spectra [71, 63]. Although the present study is based on data from eight subjects, but the EMG data have been collected from 12 muscles per subject: thus, increasing the number of muscles

sampled. The developed framework has been applied successfully to several extensor and flexor muscles. Results show that the performance of the algorithm was relatively not sensitive to the type of the muscle, indicating flexibility in the choice of muscles.

Chapter 5

REAL TIME FATIGUE AND INTENSITY GRADING

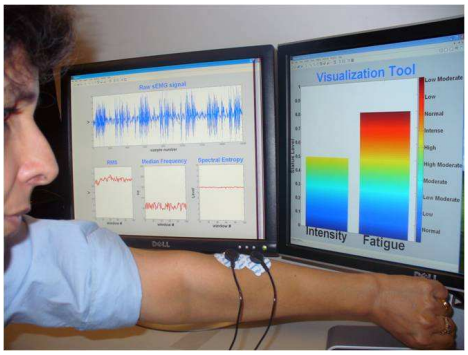


Figure 5.1: Real time Fatigue-Intensity Measurement System

The subject-independent method for grading EMG features was deployed for real time monitoring and measurement of the physiological status with respect to fatigue and intensity of activity while the subject performed repetitive gripping activity at varied cycle time. The subject could chose to do the activity at any cycle time, not necessarily the ones followed during developing the training framework. The real time system configuration and it's results are discussed in this chapter.

The real time system as shown in Figure 5.1 is a PC based system which is connected to an EMG machine via a USB port. The input data required by the framework is:

- Raw SEMG data of the test subject, from EMG sensors (machine) via USB port.
- Machine Learning model generated offline from training data.
- Factor Loadings of the training data.
- Min and Max values of training data in feature domain.
- Min and Max values of training data in factor score domain.

The various steps involved in data processing at real time are:

- Raw SEMG data is sampled at 1000Hz.
- Data read at every second is passed through a hp filter of 5 Hz to remove motion artifacts.

- For every 1000 samples all the features are computed.
- The features are smoothed over a variable window length. It is fixed at 5 presently.
- These features are input to the machine learning model for classification.
- The normalization of the feature vector prior to computing factor score is done using the training data max and min values.
- The factor loading is used to compute the factor score of the corresponding 12 dimensional feature vector.
- The Fatigue and Intensity levels are obtained from the factor score. The factor component 1 and 2 give the intensity level and activity level correspondingly.
- These values are scaled with respect to max and min factor score values of the training data in scale of 0 to 1.

The output of the real time framework is as follows:

- Detection of the physiological status with respect to the four states of Fatigue and Intensity of activity.
- Fatigue and Intensity level on a continuous scale of 0 to 1, as bar graphs.
- Real time display of the raw SEMG signal being received from the sensors.
- Display of significant features i.e. RMS, Median Frequency and Entropy.

Results of the Real time system when a subject performed the repetitive gripping activity slowly from low speed to high speed are presented in this section. Figure 5.2 shows the raw SEMG data collected by the real time system. Figures 5.3 and 5.4 show the Fatigue and Intensity gradings generated by the real time system with respect to the training model or reference group and figures 5.5 and 5.6 show the fatigue and intensity levels as computed by the offline software with respect to it's own factor loadings. Thus it is seen that Intensity level follows the speed or the cycle time of the activity for both real time and offline results. The fatigue gradings generated by the framework were evaluated against the subjects' confirmation. Results show that the fatigue level computed at real time follows the offline computation as well. The framework was tested on multiple subjects and found to match with the subject's personal experience of fatigue level.

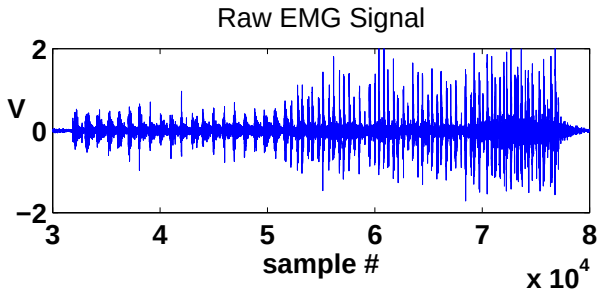


Figure 5.2: Real time raw SEMG signal

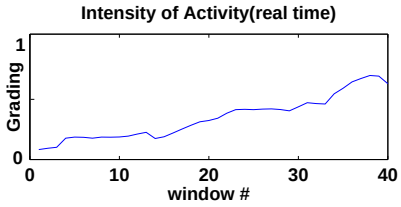


Figure 5.3: Intensity level computed real time

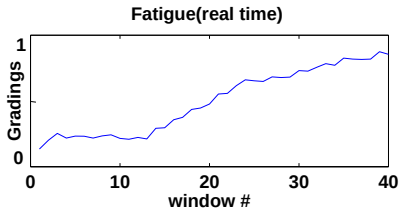


Figure 5.4: Fatigue level computed real time

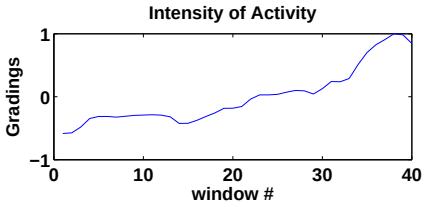


Figure 5.5: Intensity level computed offline

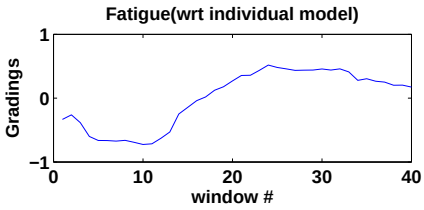


Figure 5.6: Fatigue level computed offline

Some of the critical issues in the real time system are noise in the data, motion artifacts, filtering 1000 samples at a time for reducing motion artifacts induces variation in the data. Also the standardization of the raw signal is not possible as overall mean and standard deviation of the data is not known.

Chapter 6

MARGINAL PROBABILITY BASED BATCH-MODE ACTIVE LEARNING (MP-AL)

Classification has been an active research topic in data mining and machine learning. The availability of a large amount of digital data in recent years, has greatly expanded the opportunities of automated data classification using data mining and machine learning techniques. One of the prerequisites for any data classification framework is the availability of labeled examples. Annotating large quantities of data for developing automated classifiers is a time consuming and expensive process. Hence there is a need to select an optimal set of instances from the pool of unlabeled data for labeling, such that a classifier learned on the selected instances performs well on the unlabeled data and also on unseen data belonging to the same distribution. Randomly selecting a set of unlabeled instances may result in selection of redundant and non-informative instances. *Active learning* methodologies enable selection of a set of most informative unlabeled instances from enormous amount of unlabeled data for manual labeling, with the intention of developing a good classifier with a low generalization error. Specifically, the goal of active learning is to label as little data as possible, to achieve a certain classification performance, thus saving considerable annotation cost for training a good learner.

This chapter presents an efficient batch-mode active learning method based on the principles of matching distributions between the labeled data used for training and the pool of unlabeled data, that needs to be annotated.

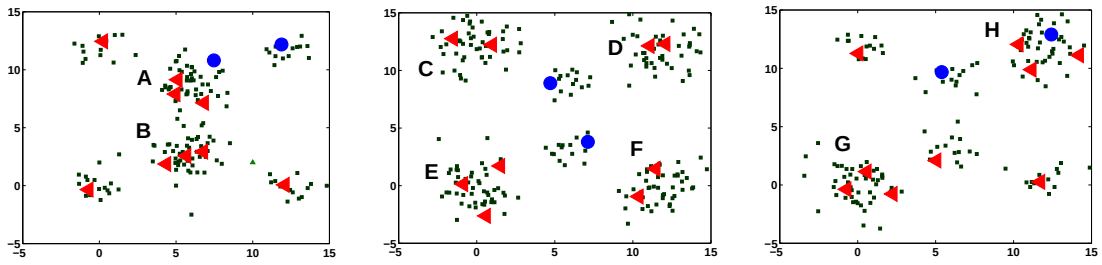


Figure 6.1: Three toy data sets with different data distributions (dark green squares) and corresponding selected sets of query data points (red triangles) based on the proposed algorithm, selected in 3 iterations in batches of 3 data points. The two data points represented by blue circles are randomly selected initially available labeled data points (figures best viewed in color).

6.1 Overview

Active learning methodologies iteratively select the most informative data. Informativeness of a data sample or a set of data samples is measured by their potentiality in increasing the performance of a classifier, once their label is known [94]. Many researchers have addressed the active learning problem in various ways [76]. Most have focused on selecting a single most informative unlabeled instance to query each time. The most popular approaches include query-by-committee [77, 15, 23] where a number of distinct classification models are generated and an instance having the most disagreement among the classification models in predicting the label is selected for querying. Another popular approach is querying an instance with maximum uncertainty of labeling measured by the distance from the classification boundary [11, 75, 86] or by the entropy in the predicted label [44, 43].

Most single instance selection methods require to retrain the classifier with each single instance being labeled. The retraining process between queries can make the process very slow. Furthermore, if a parallel labeling system is available, e.g., multiple annotators working parallely, single instance selection would not be able to make the effective use of the resources. A batch-mode active learning strategy, that selects multiple instances each time is more appropriate under these circumstances. However, the challenge in batch-mode active learning is the formulation of the selection criteria for multiple instance selection. Using a single instance selection strategy to select a batch of queries in each iteration by ranking them based on their individual merit may not give good results, as this strategy fails to take into account the information overlap between multiple instances. Hence principles for batch-mode active learning to select a set of instances simultaneously based on their collective merit, need to be developed to address the multi-instance selection *specifically*.

The batch-mode active learning methods are particularly suitable for large scale applications where the data has high redundancy such as text classifications [31], content based image retrieval [36] and image recognition [35] due to high frame rate. The greatest challenge in selecting a set of instances simultaneously is two fold. The first challenge lies in the formulation of a right objective which will be optimized to select the most informative set of samples and the second challenge is concerned with the computational complexity of the NP hard combinatorial integer programming problem for obtaining a good local solution.

Recently, several sophisticated batch-mode active learning methods based on optimizing an information measure have been proposed. [31] selected a batch of query samples based on maximum mutual information with the unlabeled set of data, while [35] applied Fisher information matrix to select a set of informative instances. [92] selected a set of instances closest to the basis vectors that approximate the set of unlabeled data and [32] proposed a discriminative approach.

This dissertation proposes a novel batch-mode active learning approach that selects a set of query instances such that the marginal distribution represented by the selected instances is closest to the distribution represented by the unlabeled data; hence a classifier learned by minimizing loss on the selected set of labeled data has good generalization capabilities on the unlabeled data and also on the unseen data coming from the same distribution. This statistical assumption is based on the fact that traditional machine learning algorithms provide performance guarantees on a classifier, only when the test data belongs to the same underlying distribution as the training data [88].

In other words, in order to learn a classifier with a budgeted number of labeled data the method selects a set of samples Q from the unlabeled set of data, denoted by U, such that the probability distributions represented by $L \cup Q$ and $U \backslash Q$, where L is the set of available labeled data, are similar to each other. Reducing distribution differences between the training and test data has been previously used in the context of *transfer learning* applications [39, 65, 80, 84, 5], however, performing active learning on the basis of the same criterion, is a novel contribution of this work.

The developed method measures the difference in the marginal probability distribution between the two sets of data using the Maximum Mean Discrepancy (MMD) proposed by [8, 30, 81]. Maximum Mean Discrepancy is a statistical test based on the fact that two distributions are different if and only if there exists at least one function in a characteristic RKHS [81] having different expectations on the two distributions. MMD has been proven to be very effective in finding samples that were generated from the same distribution and outperforms its best competitors. MMD has been widely and successfully used in *transfer learning* applications [39, 65] to ensure similarity in marginal distribution between training and test data. This measure is used in an optimization formulation to select a subset Q out of all candidate subsets, based on minimum distribution difference between $L \cup Q$ and $U \backslash Q$. To the best of our knowledge, this is the first work that uses MMD in the active learning context.

Figure 6.1 shows the data points selected by the proposed method (red triangles) under three different distributions of unlabeled data (dark green squares). Six dense regions of two different densities were created. A budget of nine query points was kept, which were selected in batches of three, in three iterations. The active sampling process was started with two randomly selected data samples (blue circles). The results show that the proposed method selects points from every dense region. It is also interesting to note that the number of points that get selected from each dense region is approximately proportional to the density of the region, i.e., comparatively more data samples get selected from denser regions, A, B [Figure6.1 (a)], C, D, E, F [Figure6.1 (b)], and G, H [Figure6.1 (c)], thus preserving the distribution of the unlabeled data. It is also observed that since available labeled data is considered at every iteration, hence diversity with respect to available data is maintained in query selection, as shown in Figures 6.1 (b) where no query data gets selected from the small regions in the center as an instance from those regions is already available in the initial labeled set. More details about the properties of query points are provided in Section 6.2. It is also observed that the developed method decreases MMD monotonically as more data samples are selected from the unlabeled data and the decrease in MMD value corresponds to the increase in classification accuracy on the test set, discussed in detail in Section 6.3.

The subset selection problem is an NP-hard combinatorial integer programming problem. Specifically, the proposed formulation is an integer quadratic programming problem. It is shown that the quadratic formulation can be reformulated as an integer linear programming problem. This dissertation provides two optimization techniques to solve this problem. In the first method, a continuous quadratic programming problem (by relaxing the integer constraint) on a convex function is solved, providing a global solution. This is unlike most of the state-of-the-art batch-mode active learning methods which provide a local solution following a gradient descent method [32, 31] or a greedy algorithm [35, 9, 90]. In the second method, a continuous linear programming problem is solved.

6.2 Batch-mode Active Sampling based on Distribution Matching

Problem Setting and Motivation

The key hypothesis in active learning is that if the learning algorithm is allowed to choose the data from which it learns, it will perform better even with less annotation [76]. Given a parametric classification model, the learning algorithms often learn the parameters θ by maximizing the joint probability $P(X, Y|\theta) = P(X|\theta)P(Y|X, \theta)$ where X and Y are represented empirically by the

training data $X_{tr} = \{x_1, x_2, \cdots, x_n\}$ and their corresponding labels $Y_{tr} = \{y_1, y_2, \cdots, y_n\}$ and $P(X)$ and $P(Y|X)$ denote the marginal and conditional probability distribution of X and Y respectively. Traditional data mining and machine learning algorithms are based on the assumption that the training data (X_{tr}, Y_{tr}) represents the true underlying distributions of X and Y and hence a model learned on this data works well for the test data (X_{tst}, Y_{tst}) which is also drawn i.i.d. from the same distribution. When the distributions on the training and test set are different the classification model learned on the training data performs poorly on the test data due to model mismatch.

The proposed active learning method addresses this issue by iteratively selecting a set of query instances from the unlabeled data such that the distribution represented by the queried and labeled data (X_{tr}, Y_{tr}), is similar to the probability distribution of the unlabeled data set. In other words, in order to learn a classifier with a budgeted number of labeled data, the proposed method iteratively selects a set of samples S from the unlabeled set of data, denoted by U, such that the joint probability distributions $P(X, Y)$ represented by $X_{tr} = L \cup S$ and $X_{tst} = U \backslash S$, where L is set of available labeled data, are similar to each other. Since the labeling function or the conditional probability $P(Y|X)$ remains the same for both S and $U \backslash S$ as they are drawn from the same underlying distribution, the problem reduces to selecting S such that the marginal probability $P_{S \cup L}(X)$ is similar to $P_{U \backslash S}(X)$. In this dissertation, the difference in the marginal probability distribution between the two sets is measured empirically using Maximum Mean Discrepancy (MMD) [8, 30, 81]. The difference between the empirical means of two distributions after mapping onto a reproducing kernel Hilbert space, called Maximum Mean Discrepancy, has been shown to be an effective measure of the difference in their marginal probability distributions. A basic review on MMD is presented below.

Maximum Mean Discrepancy (MMD)

Let $X = \{x_1, \cdots, x_m\}$ and $Z = \{z_1, \cdots, z_n\}$ be two sets of samples drawn randomly from a target population. Let p and q be the probability distributions defined on the basis of sample sets X and Z respectively. The Maximum Mean Discrepancy (MMD) proposed by Borgwardt et al [8, 30, 81] is a statistical tool that provides a method for testing whether two distributions p and q from which X and Z have been drawn respectively are similar or not.

The principal underlying the Maximum Mean Discrepancy is to find a function that assumes different expectations on two different distributions so that when evaluated empirically on

samples drawn from the different distributions it would tell us whether the distributions are similar or not. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathcal{R}$ and let X, Z, p, q be defined as above. Then the Maximum Mean Discrepancy and its empirical estimate are defined as:

$$\text{MMD}[\mathcal{F}, p, q] := \sup_{f \in \mathcal{F}} \left(E_p[f(x)] - E_q[f(z)] \right). \tag{6.1}$$

$$\text{MMD}[\mathcal{F}, X, Z] := \sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} f(x_i) - \frac{1}{n} \sum_{i=1}^{n} f(z_i) \right). \tag{6.2}$$

Intuitively if $\mathcal{F}$ is 'rich enough', MMD$[\mathcal{F}, X, Z]$ will vanish if and only if $p = q$. A class of functions for which MMD may easily be computed, while retaining the ability to detect all discrepancies between p and q without making any simplifying assumptions is the complete inner product space $\mathcal{H}$ (i.e., a reproducing kernel Hilbert space (RKHS) [82]) of functions $f : \mathcal{X} \to \mathcal{R}$, where $\mathcal{X}$ is a nonempty compact set and for all $x \in \mathcal{X}$, the linear point evaluation functional mapping $f \to f(x)$ exists and is continuous. In this case, $f(x)$ can be expressed as an *inner product* via

$$f(x) = \langle \phi(x), f \rangle_{\mathcal{H}}. \tag{6.3}$$

where $\phi : \mathcal{X} \to \mathcal{H}$ is known as the feature space map from $\mathcal{X}$ to $\mathcal{H}$ [8]. When $\mathcal{F}$ is the unit ball in a characteristic RKHS [81], MMD is defined as the difference between the means of two distributions after mapping onto the characteristic RKHS. An empirical estimate of MMD is then obtained as follows:

$$\text{MMD}[\phi, X, Z] := \left\| \frac{1}{m} \sum_{i=1}^{m} \Phi(x_i) - \frac{1}{n} \sum_{i=1}^{n} \Phi(z_i) \right\|_{\mathcal{H}}^2. \tag{6.4}$$

For more details about Maximum Mean Discrepancy, the related theoretical proofs and comparison with related methods, interested readers may refer to [8, 30, 81].

Proposed Batch Mode Active Sampling

The proposed batch-mode active sampling method, referred to as *Marginal Probability based Active Learning* (MP-AL) iteratively selects batches of query instances which represent best the distribution of the unlabeled instances so that a classifier learned by minimizing risk on the queried data after labeling, has good generalization performance on the unlabeled data set and on future unseen data that comes from the same distribution. The problem is formulated as an integer quadratic programming problem which can be further reformulated as an equivalent integer linear programming problem.

The proposed framework uses MMD to measure the distribution difference between two sets of samples. Let us assume that we have n_u instances of unlabeled data U and n_l instances of labeled data L and we would like to select a batch S of b instances such that the distribution of $L \cup S$ is similar to the distribution of $U \backslash S$. In that case, the MMD between the sets $L \cup S$ and $U \backslash S$ defined by $f(S)$, can be computed using the expression in Equation (8.3), as follows:

$$f(S) = \left\| \frac{1}{n_l + b} \sum_{j \in L \cup S} \Phi(x_j) - \frac{1}{n_u - b} \sum_{i \in U \backslash S} \Phi(x_i) \right\|_{\mathcal{H}}^2 . \tag{6.5}$$

Since we want to select a set S that minimizes the mismatch between $L \cup S$ and $U \backslash S$ this method propose to select a subset S of U that minimizes $f(S)$. Next a binary vector α of size n_u is defined, where each entry α_i indicates whether the data $x_i \in U$ is selected or not. If a point is selected, the corresponding entry α_i is 1 else 0. Thus the minimization problem reduces to finding α that minimizes the cost function $f(S)$:

$$\begin{aligned} \min_{\alpha : \alpha_i \in \{0,1\}, \alpha^T \mathbf{1} = b} \quad & \left\| \frac{1}{n_l + b} \left(\sum_{j \in L} \Phi(x_j) + \sum_{i \in U} \alpha_i \Phi(x_i) \right) - \right. \\ & \left. \frac{1}{n_u - b} \quad \sum_{i \in U} (1 - \alpha_i) \Phi(x_i) \right\|_{\mathcal{H}}^2 . \end{aligned} \tag{6.6}$$

where **1** is a vector of the same dimension as α with all entries 1 and symbol T is used to represent the matrix or vector *transpose* operation. Evidently, the cost function in Equation (6.6) is an alternative (equivalent) representation of the cost function $f(S)$ in Equation (6.5). The first term denotes the mean of the mapped features of the labeled and selected points. Note that if a point x_i is not selected in the current set then α_i will be 0 and this term would not get added in the summation. The second term is mean of the mapped features of the unlabeled data set minus the selected query set. The first constraint ensures that each entry in α is either 0 or 1 and the second constraint ensures that exactly b entries of α are 1, meaning exactly b instances are selected from the unlabeled data set, where b is specified a priori by the user. The above formulation can be represented as:

$$\min_{\alpha : \alpha_i \in \{0,1\}, \alpha^T \mathbf{1} = b} \quad \frac{1}{2} \alpha^T K_{u,u} \alpha - k_{u,u}^T \alpha + k_{l,u}^T \alpha + \text{const.} \tag{6.7}$$

The various terms in the above expression are given as follows. G is denoted as the $(n_u + n_l) * (n_u + n_l)$ kernel Gram matrix over the unlabeled data U and labeled data L, arranged in order, using a kernel function K such that $G(i,j) = K(x_i, x_j)$. Then, $K_{u,u} = G(1 : n_u, 1 : n_u)$, $k_{u,u}(i) = \frac{n_l + b}{n_l + n_u} \sum_{j=1}^{n_u} K_{u,u}(i,j)$, and $k_{l,u}(i) = \frac{n_u - b}{n_l + n_u} \sum_{j=1}^{n_l} G(i, n_u + j)$.

Based on the above expressions, one can draw the following observations regarding the properties of the selected query set:

- The first term ensures that the selected query set has minimum similarity within itself, avoiding *redundancy* in the selected set.
- The second term enforces the selected examples to be similar to the unselected ones, ensuring *representativeness*.
- The third term implies the examples with less similarity with already labeled data are more likely to be selected ensuring *diversity* in the selected set.

Thus the proposed method selects examples which meet all the desirable properties for batch mode active learning. The proposed method can be easily extended to add any other evaluation criteria (M) by adding a corresponding linear term $M^T\alpha$, while still maintaining the quadratic form (see Section 6.4 for more details). Also the proposed method does not depend on the availability of labeled data to initiate the process of selecting a query set, in which case $n_l = 0$, and the third term $k_{l,u}^T\alpha$ in Equation (6.7) vanishes.

The above optimization formulation is an integer quadratic programming problem. Next, it is reformulated as an equivalent integer linear programming (ILP) problem.

Reformulation as an ILP Problem

Due to the binary constraint $\alpha_i \in \{0, 1\}, \forall i$, the linear terms in the objective defined in Equation (6.7) can be absorbed into the quadratic term by subtracting $k_{u,u}$ and adding $k_{l,u}$ terms to the diagonal entries of matrix $K_{u,u}$, forming a D matrix given by: $D(i, j) = K_{u,u}(i, j) - k_{u,u}(i) + k_{l,u}(i)$ for $i = j$ and $D(i, j) = K_{u,u}(i, j)$ otherwise. The optimization problem in (6.7) can be rewritten as:

$$\min_{\alpha:\alpha_i\in\{0,1\},\alpha^T\mathbf{1}=b} \quad \alpha^T D\alpha. \tag{6.8}$$

Next, a binary matrix $Z = (z_{ij})$ with $z_{ij} = \alpha_i\alpha_j$ is defined. Thus the optimization in Equation (6.8) becomes:

$$\begin{aligned} \min_{\alpha,Z} \quad & \sum_{i,j} d_{ij} z_{ij} \\ \text{s.t.} \quad & z_{ij} = \alpha_i\alpha_j, \qquad \alpha_i \in \{0, 1\}\ \forall i, j, \qquad \alpha^T\mathbf{1} = b. \end{aligned} \tag{6.9}$$

The quadratic equality constraint $z_{ij} = \alpha_i\alpha_j$ makes the problem difficult to solve. It is shown that this can be represented by a set of linear inequalities. Since d_{ij} can have both negative and

positive values, the quadratic constraints are rewritten as follows:

$$\begin{aligned} \min_{\alpha,Z} \quad & \sum_{i,j} d_{ij} z_{ij} \\ \text{s.t.} \quad & -\alpha_i - \alpha_j + 2z_{ij} \leq 0 \text{ for } d_{ij} < 0 \\ & \alpha_i + \alpha_j - z_{ij} \leq 1 \quad \text{for } d_{ij} \geq 0 \\ & \alpha_i \in \{0,1\}\ \forall i,j, \quad \alpha^T \mathbf{1} = b. \end{aligned} \tag{6.10}$$

The first constraint ensures that z_{ij} equals zero if the value of α_i or α_j (or both) is zero. If α_i and α_j both equal to one, the value of z_{ij} is free to be either 0 or 1; however, the minimization forces the value of z_{ij} to be 1 since d_{ij} is negative. Similarly, the second constraint when d_{ij} is positive, is derived. The second constraint makes the value of z_{ij} equals to 1 if α_i and α_j both equal to one. If any of the α_i and α_j equals zero, then the value of z_{ij} is free to be either 0 or 1; nevertheless, at optimality, z_{ij} must equal 0 since d_{ij} has positive contribution to the objective, which is a minimization of the cost function. Consequently, the relation between z_{ij} and the pair α_i and α_j at optimality is as follows: $z_{i,j} = 1$, if and only if $\alpha_i = 1$ and $\alpha_j = 1$. Thus, the formulation in Equation (6.10) is equivalent to the original integer quadratic programming problem.

Next, two algorithms are presented to solve the integer quadratic and integer linear optimization problems defined in Equation (6.7) and Equation (6.10) respectively as quadratic programming (QP) and linear programming (LP) problems by relaxing the binary constraints.

Quadratic Programming (QP) Problem

The binary constraint on α_i makes the integer quadratic problem defined in Equation (6.7) NP-hard. A common strategy is to relax the constraints to make it a continuous optimization problem, which can be solved in polynomial time:

$$\min_{\alpha: 0 \leq \alpha_i \leq 1, \alpha^T \mathbf{1} = b} \quad \frac{1}{2} \alpha^T K_1 \alpha - k_2^T \alpha + k_3^T \alpha. \tag{6.11}$$

The minimization problem in (6.11) is a standard quadratic problem (QP) and can be solved efficiently by applying many existing solvers. The 'quadprog' function in MATLAB is used to solve this QP problem. The overall algorithm for selecting the query set S at any iteration is given in Algorithm 4.

Linear Programming (LP) Problem

The integral constraint is relaxed to obtain a linear program (LP) formulation which is a relaxation of the ILP formulation in (6.10) as follows:

$$\begin{aligned} \min_{\alpha, Z} \quad & \sum_{i,j} d_{ij} z_{ij} \\ \text{s.t.} \quad & -\alpha_i - \alpha_j + 2z_{ij} \leq 0 \text{ for } d_{ij} < 0 \\ & \alpha_i + \alpha_j - z_{ij} \leq 1 \quad \text{for } d_{ij} \geq 0 \\ & \alpha_i, z_{ij} \in [0,1] \ \ \forall i, j, \quad \alpha^T \mathbf{1} = b. \end{aligned} \tag{6.12}$$

The LP formulation can be further simplified by incorporating the first constraint into the objective function. Since this is a minimization problem when $d_{ij} < 0$, at optimality, $z_{ij} = \frac{\alpha_i + \alpha_j}{2}$ following the first equality constraint. On the other hand, the second equality constraint for $d_{ij} \geq 0$, may not hold at optimality. Since removing or relaxing a constraint of a minimization program does not reduce the optimal value, the formulation in Equation (6.12) can be reformulated as follows:

$$\begin{aligned} \min_{\alpha, Z} \quad & \frac{1}{2} \sum_{d_{ij}<0} d_{ij}(\alpha_i + \alpha_j) + \sum_{d_{ij} \geq 0} d_{ij} z_{ij} \\ \text{s.t.} \quad & \alpha_i + \alpha_j - z_{ij} \leq 1 \text{ for } d_{ij} \geq 0 \\ & z_{ij} \in [0,1] \text{ for } d_{ij} \geq 0 \\ & \alpha_i \in [0,1] \ \ \forall i, j, \quad \alpha^T \mathbf{1} = b. \end{aligned} \tag{6.13}$$

The problem in Equation (6.13) is a standard linear programming problem, and can be solved efficiently using any standard LP solver. Since the Hessian matrix K_1 in Equation (6.7) is a kernel Gram matrix which is positive semi-definite hence both formulations are convex. `CVX` [29] is used to solve the LP problem. The overall algorithm for selecting the query set S at any iteration using the LP formulation is given in Algorithm 4.

6.3 Experiments and Results

Datasets. The empirical performance of the proposed *MP-AL* algorithm is evaluated using eight datasets from the UCI machine learning repository (both binary and multi-class), and a biological image dataset (Fly-FISH). The biological image dataset consists of 1016 images of 7 developmental stages of *Drosophila*, commonly known as fruit-fly. Each stage forms a class. Each image is represented by 3850 textural features that are extracted using Gabor filters [55].

ALGORITHM 4: MP-AL

Input: L: set of labeled instances; U: set of unlabeled instances; b: batch size;
Output: S: query set;
Compute K_1, k_2 and k_3 as explained in Section 6.2.
if QP Problem **then**
    Compute α by solving (6.11).
end if
if LP Problem **then**
    Form D matrix as explained in Section 6.2.
    Compute α by solving (6.13).
end if
Sort U in descending order of α and select top b instances as S.
Update sets L and U: $L \rightarrow L \cup S$, $U \rightarrow U \backslash S$.

Competing Methods. The performance of the proposed approach is compared with state-of-the-art batch-mode active learning methods which selected a set of instances based on their collective merit including *Matrix* [31], *Fisher* [35] and *Disc* [32]. The developed method is also compared with state-of-the-art batch-mode active learning methods which selected a set of instances based on their individual merit such as *svmD* [9] and *MCS* [90], besides comparing to one transductive experimental design method, referred to as *Design* [92]. The sequential design code is downloaded from the authors' webpage and the method is referred here as *Design*(s), that selects a set of instances sequentially based on their individual merits. A detailed description of each of these methods is provided in Section 8. Comparative performance of a random instance selection algorithm denoted as *Rand*, is also presented for reference.

Experimental Setup. Each dataset is randomly divided into two sets. Batch selection based on active learning methodologies was performed on one set referred to as unlabeled set (65%) and the effectiveness of the selection methodologies was measured based on classification accuracy on the other unseen fixed set (35%) referred to as the test set. Table 6.2 summarizes the sizes of each of the datasets used. Some of the datasets are subsampled due to the computational complexity of the several competing methods. This dissertation considers a hard case of active learning, where the method starts with two randomly selected labeled instances per class. All the algorithms start with the same initial labeled set, unlabeled set and test set. For a fixed batch size b, each algorithm repeatedly selected b instances for labeling at each iteration and evaluated the performance of a classifier learned on labeled instances, on the fixed test set. The size of b was fixed at 10 for all datasets except for Vehicles and Iris, where it was fixed at 5 due to their small sizes. The experiments were repeated 10 times and the average results are reported.

The QP and LP formulations are compared based on the values of the objective function in Equation (6.7) and the classification accuracies obtained by the selected query set on the fixed test set. It is observed that the values of the objective function obtained by LP were slightly lower than QP, though accuracies obtained were similar. It is also noted that the execution time of QP was generally lower than LP. This can be attributed to the larger number of constraints in LP and the specific software package used. Exploring ways to improve the efficiency of the LP formulation is a potential future work. The performance values of the proposed method included in this section are based on the QP formulation. A Gaussian kernel with the parameter value selected via cross validation and Support Vector Machines as classification model is used to evaluate the effectiveness of the queried instances.

Comparative Studies. The comparative performance of the proposed approach on UCI datasets, is shown in Figure 6.2. Results show that the proposed *MP-AL* performed better than the state-of-the-art batch-mode active learning methods for 6 out of 8 datasets and had comparable performance for the remaining two datasets: Musk and Wine. It is observed that for 6 out of 8 datasets the nearest competitors were *Matrix*, *Fisher* and *Disc*, except for Iris and Vehicles where *svmD*, *MCS* and *Design*(s) were nearest competitors.

Another set of experiments were conducted on a multi-class, high dimensional biological image dataset (Fly-FISH) for classifying different developmental stages of *Drosophila*. 511 samples (divided equally among all 7 classes) were randomly sampled. The batch size b and number of iterations were fixed at 10 and 9 respectively. Figure 6.4 reports the results of the proposed method *MP-AL* compared to the other active learning methods. The results show that *MP-AL* outperformed all the other active learning methods followed by *Matrix*, *Fisher* and *Disc*. Table 6.1 presents the results of 2-sided paired t-test of *MP-AL* vs *Matrix*, *Disc* and *Fisher* methods on UCI and Fly-FISH datasets. The accuracies are compared over 10 runs at each evaluation point and the percentage of evaluation points at which *MP-AL significantly* outperforms or under-performs the compared algorithm, denoted as win % and loss% respectively, is presented.

Variation in MMD Vs Number of Selected Samples. The variation in MMD value between the training set and the unlabeled data at each iteration for all the datasets, is evaluated. Figure 6.3 presents the results for some of the representative UCI datasets. Similar patterns were observed for the other datasets. It is observed that the developed algorithm decreases MMD value

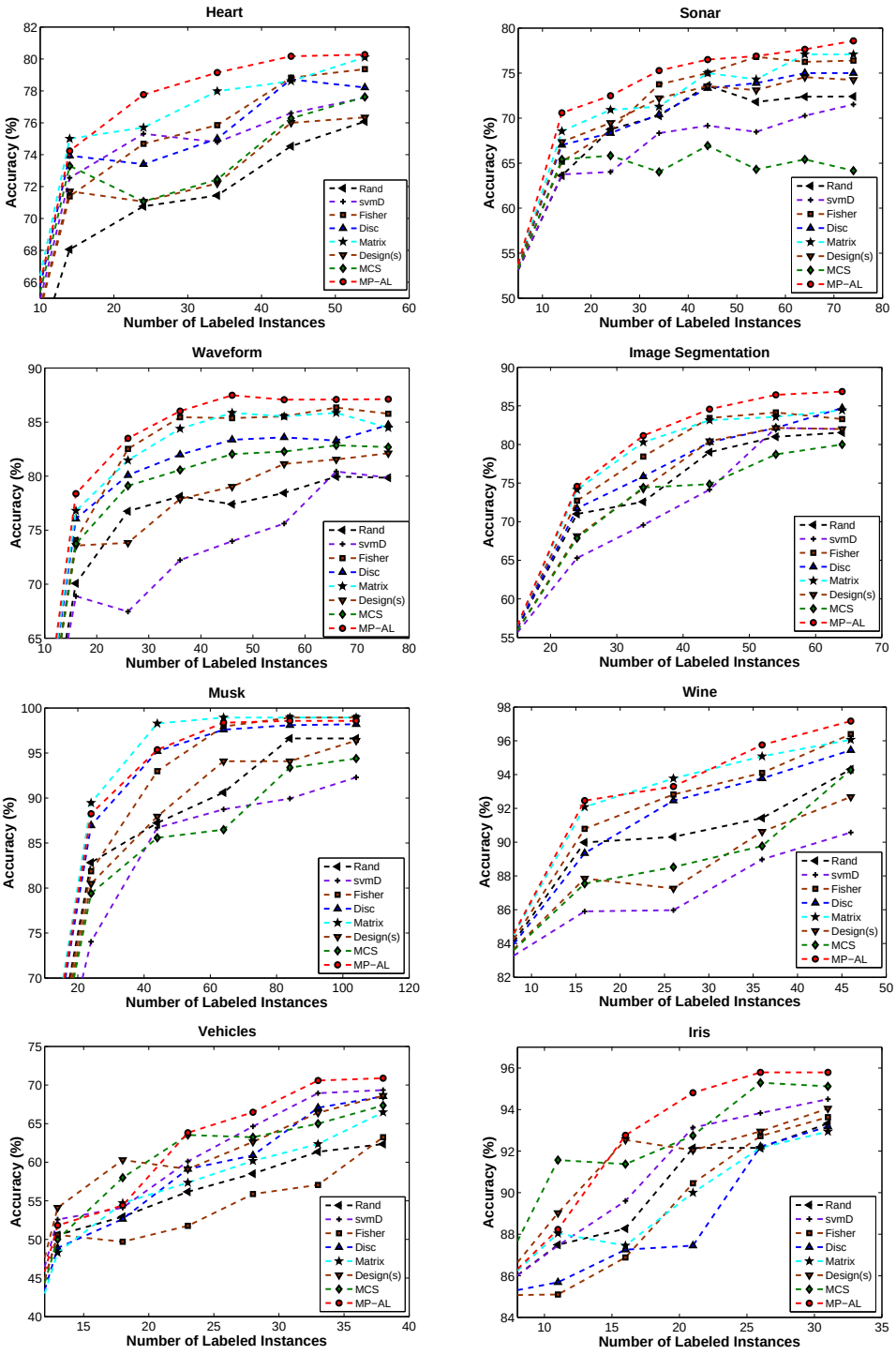


Figure 6.2: Comparative performance of different active learning methods on UCI datasets. Accuracy at the start point, which is same for all methods is not shown in the figures (figures best viewed in color). Results of 2-sided paired t-test at the level of $p < 0.05$ for *MP-AL* vs. *Matrix*, *Fisher* and *Disc* are presented in Table6.1.

Table 6.1: Win-Loss % of MP-AL in 2-sided paired t-test ($p < 0.05$). The fraction not reported (e.g., 60% for MP-AL vs. Matrix on Heart) corresponds to the cases where the two algorithms are not significantly different at the level of $p < 0.05$.

Dataset	MP-AL vs. Matrix		MP-AL vs. Disc		MP-AL vs. Fisher	
	win%	lose%	win%	lose%	win%	lose%
Heart	40.0	0	60.0	0	60.0	0
Sonar	28.5	0	85.7	0	28.5	0
Waveform	14.3	0	71.4	0	28.5	0
Image Segmentation	20.0	0	60.0	0	40.0	0
Musk	0	20.0	0.0	0	20.0	0
Wine	0.0	0	75.0	0	50.0	0
Vehicles	66.6	0	50.0	0	83.3	0
Iris	80.0	0	80.0	0	100.0	0
Fly-FISH	62.5	0	85.7	0	62.5	0

monotonically as more data samples are selected from the unlabeled data and that the decrease in MMD value corresponds to the increase in classification accuracy on the test data as shown in Figure 6.2. The decrease in MMD value during the initial iterations is more than the decrease towards the later iterations, resulting in the higher increase in accuracy values during the initial iterations than later iterations. It is noted that for the Vehicles dataset the accuracy value sharply increases between the second and third iteration points. A sharp decrease in MMD value is also observed at the corresponding iteration points, for the Vehicles dataset.

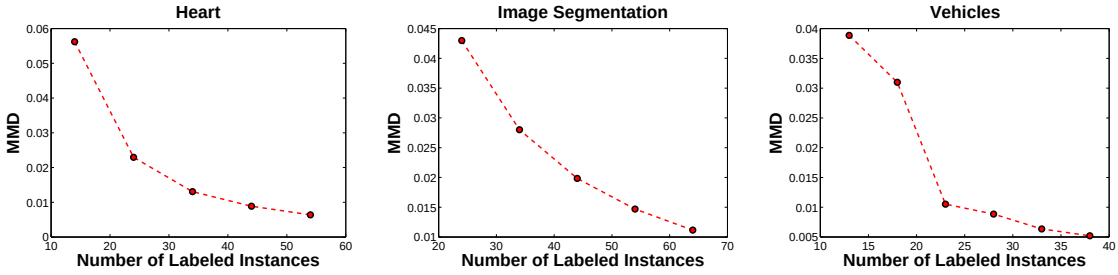


Figure 6.3: MMD value between the training and unlabeled data as more instances are selected by MP-AL.

Table 6.2: Average run time (in seconds).

Dataset	# instances x features	MP-AL	Matrix	Disc	Fisher
Vehicles	94 x 19	0.42	3.35	4.42	1.14
Iris	150 x 5	0.58	6.64	7.97	2.96
Wine	178 x 3	0.59	7.81	10.76	8.38
Sonar	208 x 60	0.76	8.68	12.78	10.04
Image Segmentation	210 x 19	0.76	12.47	22.43	16.51
Heart	270 x 13	1.21	26.14	26.93	18.88
Waveform	350 x 21	2.54	52.21	81.28	38.57
Musk	476 x 167	3.81	153.02	265.71	121.23
Fly-FISH	511 x 3850	6.5	360.70	3330.16	923.30

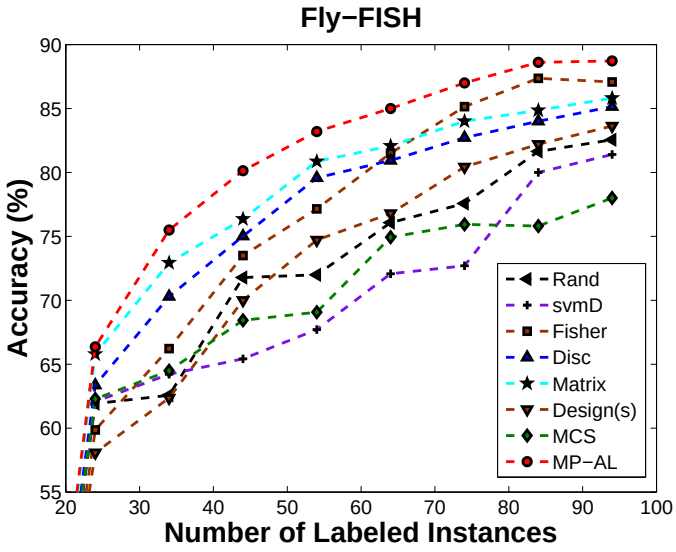


Figure 6.4: Comparative performance of different active learning methods on the Fly-FISH dataset.

Efficiency Comparison. The average time taken to select a batch of unlabeled points by the proposed *MP-AL* versus the nearest competitors *Matrix*, *Fisher* and *Disc* were computed. All algorithms were implemented using MATLAB on a four-core Intel processor with 2.66 GHz CPU and 8 GB RAM. Table 6.2 presents the comparative run times on different UCI and Fly-FISH datasets. The results show that *MP-AL* is much more efficient than the other three batch-mode active learning methods for all datasets. *Matrix* method involved solving a quadratic programming problem multiple times, per batch of query points selection. *Fisher* involved training of a classifier multiple times per query batch selection. The *Disc* method involved training of a classifier followed by solving a quadratic programming problem multiple times per selection of a query batch and the *MP-AL* on the other hand, required solving a quadratic programming problem once per batch of query points selection.

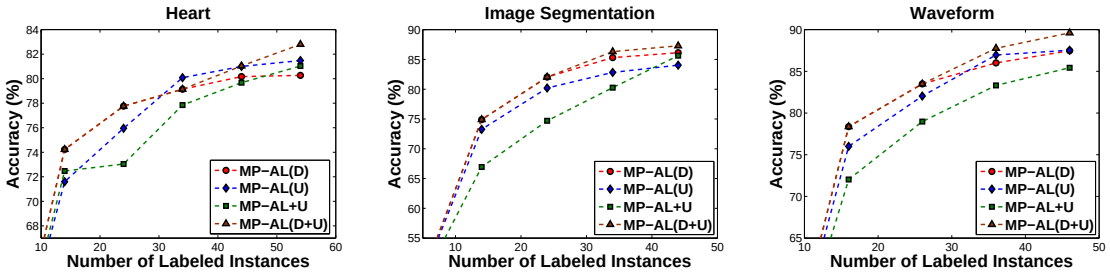


Figure 6.5: MP-AL with Uncertainty on UCI datasets.

6.4 Extensions of MP-AL

Incorporating Uncertainty in MP-AL

The proposed *MP-AL* framework can be readily extended to incorporate uncertainty of prediction of unlabeled data in the query selection process. This dissertation experimented with the following three alternative algorithms: (1) At each iteration a classifier is learned on the available labeled data, and the unlabeled data is ranked based on uncertainty of predictions $M(x_i)$ measured for each unlabeled data x_i using entropy of predicted labels as in [32]. The *MP-AL* method is then applied only on the most uncertain set of unlabeled data (top 70%) instead of on the complete unlabeled data, referred to as *MP-AL*(U). (2) Query selection is based on *MP-AL* at initial iterations and during later iterations based on the *MP-AL*(U), referred to as *MP-AL*(D+U). (3) Add the prediction uncertainty vector M as a separate linear term ($-M^T\alpha$) in the formulation in Equation (6.11), referred to as *MP-AL+U*. The base *MP-AL* method is referred to as *MP-AL*(D). Figure 6.5 presents the comparative results obtained on some of the representative UCI datasets. Similar patterns were obtained on other datasets. It is observed that for initial iterations *MP-AL*(D) and *MP-AL*(D+U) outperform *MP-AL*(U) and *MP-AL+U*. However the performance of these methods improve during later iterations, as the classifier becomes more reliable when learned on a larger number of labeled data. Indeed *MP-AL*(D+U) performs best for most cases as it combines the strengths of both *MP-AL*(D) and *MP-AL*(U).

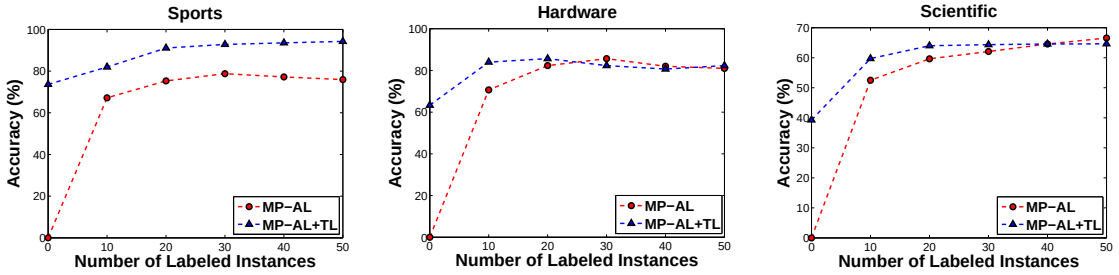


Figure 6.6: MP-AL with transfer learning on 20 Newsgroups dataset.

Transfer Learning in MP-AL

The problem of insufficient labeled data (in a target domain) is addressed in *Active learning* by querying labels of most informative instances. An alternative to address the same problem is by borrowing samples from an already labeled dataset belonging to a related domain (source domain), known as *Transfer Learning* [66]. Different transfer learning methodologies have been

developed to address the distribution difference between a source and a target domain, so that the source domain data can be efficiently used to label the target domain data. Re-weighting source domain data to match the marginal probability distributions is a commonly used strategy [39, 5, 80] in transfer learning. In the experiments an existing re-weighting method [39], was used to re-weight the source domain data to match the distribution of the unlabeled set (target domain). Transfer learning is incorporated in MP-AL framework as follows: At each iteration, the re-weighted source samples are combined with the queried and labeled samples from the unlabeled set (in the target domain) and the classification accuracy on the other unseen fixed test set, is computed, similar to earlier experiments. The proposed transfer learning extension of *MP-AL* on the 20 Newsgroups dataset for document categorization, is described here. Three sets of source domain data vs. unlabeled data (target domain) were built as follows: (1) Sports: rec.sport.hockey vs. rec.sport.baseball; (2) Hardware: comp.sys.mac.hardware vs. comp.sys.ibm.pc.hardware and (3) Scientific: sci.med vs. sci.electronics. The positive class of each source and target domain data consists of 200 documents randomly sampled from the respective categories and the negative class consists of a random mixture of 200 samples from other categories as suggested in [19]. Each document was represented as a binary vector consisting of the 200 most discriminating words determined by Weka's info-gain filter [89], after removing stop words and using a document frequency of 5. The method starts with no selected labeled instances and use a batch size of 10. The experiments were repeated 10 times and the average results are reported. Figure 6.6 shows the accuracies obtained by the extended method denoted as *MP-AL+TL* and by the base *MP-AL* method without using the source domain data, on the 20 Newsgroups dataset. It was observed that during intial iterations *MP-AL* has poorer performance, due to insufficient labeled data. It was noted that combining transfer learning with active learning (*MP-AL+TL*) improves the classification accuracy significantly during the initial iterations. However as the number of labeled data from the unlabeled set (target domain) increases, the performance of *MP-AL* improves and outperforms *MP-AL+TL* for Scientific and Hardware test cases. It has been shown both theoretically and empirically in [4] that if there are enough data from the target domain then no source data are needed, and in fact using additional source data may degrade the performance. It was observed that improvement in classification accuracies due to incorporation of transfer learning is more for Sports and moderate for Scientific and Hardware. This can be attributed to the extent of difference in distribution between the source and target domains in each of these test cases; one way to measure the distribution difference is to compute the MMD value between the source and target

domains. The MMD value is 0.0121, 0.0237 and 0.0239 for Sports, Hardware and Scientific respectively. This is consistent with our observation in Figure 6.6.

Chapter 7

JOINT OPTIMIZATION FRAMEWORK FOR TRANSFER AND BATCH MODE ACTIVE LEARNING

Traditional machine learning methods require sufficient labeled examples in order to construct accurate models. These methods also assume that labeled examples belong to the same underlying distribution as the test data; in other words, both training and test data are drawn i.i.d from the same distribution. However, for real world applications, as in the case of medical diagnosis, video concept detection, sentiment analysis, document classification etc, one may not have sufficient or any labels, belonging to the same underlying distribution as the test data. Typical examples include, labeled diagnosis data from the same medical lab or machine, labeled video clips from the same TV channel, or positive and negative reviews for the same product category, as in the test data. Two machine learning methods namely *transfer learning* [66] and *active learning* [76], address this problem in two different ways. Transfer learning methods try to solve this problem by utilizing labeled data from related domains, which may be available in plenty, e.g., labeled data from another lab or a machine, labeled video clips belonging to other TV channels or positive and negative reviews for another product category. As a different solution, active learning methods focus on selecting a small set of most informative samples from the test data, for which they acquire labels from the domain experts. Hence, under conditions where we have sufficient labeled data from a related domain and a budget to get a fix number of target samples labeled by an expert, a combination of transfer and active learning would provide an optimal solution.

This chapter develops a novel optimization problem for combining transfer and active learning methods leading to an efficient annotation framework for biomedical data.

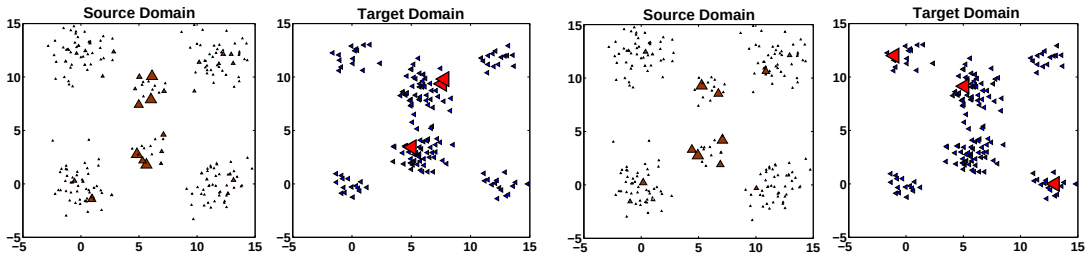


Figure 7.1: Source and target domains with different data distributions and corresponding selected set of query data points from target domain (red triangles) and weights of source instances shown by the size of the source data points based on (a) two stage approach of domain adaptation and active learning and (b) proposed single stage approach of domain adaptation and active learning (figures best viewed in color).

7.1 Overview

Availability of additional labeled data from a related source domain would increase the reliability of the classifier used in active learning; at the same time, availability of informative labeled data from target domain would enable efficient transfer of knowledge from source domains. However, not much work has been reported in the literature along this direction. Existing work [78, 68] does not integrate the transfer and active learning methodologies into a single consolidated framework. Transfer and active learning are performed in two stages, which may cause redundancy or information overlap between the instances selected from the source and target domain data (illustrated in the second paragraph of the following page). Besides, the transfer learning or domain adaptation is just performed initially once and is not dynamically updated at every iteration of active learning, as more informative samples are queried and labeled from target domain data.

This dissertation proposes a novel transfer and active learning method that addresses the above mentioned issues. The proposed method re-weights source samples and selects a batch of query samples from the target domain, such that the marginal distribution represented by the data set consisting of re-weighted source samples, labeled target domain data (if any) and the selected query samples from the target domain, is closest to the distribution represented by the set of unlabeled target domain data, with the purpose of learning a classifier with low generalization error. This is achieved by solving a convex optimization problem, which minimizes the difference in a marginal probability measure between the two data sets. The optimization problem is found to minimize the similarities between the source data samples with larger weights and the selected target samples, potentially avoiding information overlap between them. This process is repeated at every iteration to update transferred knowledge dynamically (see Section 7.2).

To illustrate the problem of information overlap, a source and target domain data with different marginal probability distributions, as shown in Figure 7.1 were created. Six dense regions of different densities for each of the domains, were created. Figure 7.1 (a) shows the re-weighted source domain (size of the triangles is proportional to the weights) and the query set (batch size = 3) selected from target domain data by following a two step methodology, i.e., domain adaptation followed by active learning as in [68]. And Figure 7.1 (b) shows the re-weighted source domain data and the query set selected from target domain data by the proposed framework. One can

observe that similarity in instances that got higher weight in source domain and that got selected from target domain, in the two stage approach (Figure 7.1 (a)), is significantly higher compared to the case when the domain adaptation and active learning are done simultaneously (Figure 7.1 (b)), leading to considerable information overlap amongst the instances, in the former case.

This is the one of the first work that performs transfer and batch-mode active learning simultaneously via a convex optimization problem. Batch-mode active learning selects a 'set' of most informative instances [31, 32, 35, 92]. Reducing marginal distribution with the target domain data via re-weighting source instances has been previously used in the context of *transfer learning* applications [39, 65, 80, 84, 5], however, performing active learning jointly on the basis of the same criterion, has been a novel contribution of this work.

The difference in the marginal probability distribution between the two sets of data was measured using the Maximum Mean Discrepancy (MMD) proposed by Borgwardt et al. [8, 30, 81].

The subset selection problem is an NP-hard combinatorial integer programming problem. The proposed formulation is an integer quadratic programming problem. A continuous quadratic programming problem (by relaxing the integer constraint) on a convex function is solved. The proposed formulation is also easily extendable to multi-source settings and is easily configurable for only transfer or active learning with corresponding parameter changes.

7.2 Joint Optimization based on Marginal Distribution Matching

Different transfer learning methodologies have been developed to address the distribution difference between a source and a target domain, so that the source domain data can be efficiently used to label the target domain data. Re-weighting source domain data to match the marginal probability distributions is a commonly used strategy [39, 5, 80] in transfer learning. Transfer learning was incorporated in the MP-AL framework as follows:

Let us assume that we have n_s instances of re-weighted source domain data S_a, n_u instances of unlabeled target domain data U and n_l instances of labeled target domain data L and we would like to select a batch Q of b instances such that the marginal distribution of $S_a \cup L \cup Q$ is similar to the marginal distribution of $U \backslash Q$. The marginal distribution difference between these two sets can be defined as follows:

$$\tilde{f} = \left\| \frac{1}{n_s + n_l + b} \sum_{j \in S_a \cup L \cup Q} \Phi(x_j) - \frac{1}{n_u - b} \sum_{i \in U \backslash Q} \Phi(x_i) \right\|_{\mathcal{H}}^2, \tag{7.1}$$

where $\phi : \mathcal{X} \rightarrow \mathcal{H}$ is known as the feature space map from $\mathcal{X}$ to $\mathcal{H}$ [8]. Since we want to select a set Q that minimizes the distribution mismatch between $S_a \cup L \cup Q$ and $U \backslash Q$ the developed method selects a subset Q of U that minimizes $\tilde{f}$. Next, a binary vector α of size n_u where each entry α_i indicates whether the data $x_i \in U$ is selected or not, is defined. If a point is selected, the corresponding entry α_i is 1 else 0. Another vector β of size n_s where each entry β_i indicates the weight of the data $x_i \in S$ is also defined. Then, the problem reduces to finding α and β that minimizes the cost function $\tilde{f}$:

$$\begin{aligned} \min \quad & \left\| \frac{1}{n_s + n_l + b} \left(\sum_{i \in S} \beta_i \Phi(x_i) + \sum_{j \in L} \Phi(x_j) + \sum_{i \in U} \alpha_i \Phi(x_i) \right) - \frac{1}{n_u - b} \sum_{i \in U} (1 - \alpha_i) \Phi(x_i) \right\|_{\mathcal{H}}^2, \\ \text{s.t.} \quad & \alpha_i \in \{0, 1\}, \beta_i \in [0, 1], \alpha^T \mathbf{1} = b, \end{aligned} \tag{7.2}$$

where $\mathbf{1}$ is a vector of the same dimension as α with all entries 1 and symbol T is used to represent the matrix or vector *transpose* operation. Evidently, the cost function in Equation (7.2) is an alternative (equivalent) representation of the cost function $\tilde{f}$ in Equation (7.1). The first term in Equation (7.2) denotes the mean of the mapped features of re-weighted source data, labeled target data and selected target data. Note that if a point x_i is not selected in the current set then α_i will be 0 and this term would not get added in the summation. The second term is mean of the mapped features of the unlabeled data set minus the selected query set. The first constraint ensures that each entry in α is either 0 or 1 and the third constraint ensures that exactly b entries of α are 1, meaning exactly b instances are selected from the unlabeled data set, where b is specified a priori by the user. The above formulation can be represented as:

$$\begin{aligned} \min \quad & \frac{1}{2} \alpha^T K_{u,u} \alpha + \frac{1}{2} \beta^T K_{s,s} \beta + \beta^T K_{s,u} \alpha - k_{u,u}^T \alpha - k_{s,u}^T \beta + k_{u,l}^T \alpha + k_{s,l}^T \beta. \\ \text{s.t.} \quad & \alpha_i \in \{0, 1\}, \beta_i \in [0, 1], \alpha^T \mathbf{1} = b. \end{aligned} \tag{7.3}$$

The various terms in the above expression are given as follows. n_s, n_u and n_l are denoted as the number of source domain data S, unlabeled target domain data U and labeled target domain data L respectively, G as the $(n_s + n_u + n_l) * (n_s + n_u + n_l)$ kernel Gram matrix over S, U and L, arranged in order, using a kernel function K such that $G(i, j) = K(x_i, x_j)$. Then,

$K_{s,s} = \frac{1}{c^2} G(1 : n_s, 1 : n_s)$, $K_{u,u} = G(n_s + 1 : n_s + 1 + n_u, n_s + 1 : n_s + 1 + n_u)$,

$K_{s,u} = \frac{1}{c} G(1 : n_s, n_s + 1 : n_s + 1 + n_u)$, $k_{u,u}(i) = \frac{n_l + n_s + b}{c^2 (n_u - b)} \sum_{j=1}^{n_u} K_{u,u}(i, j)$,

$k_{s,u}(i) = \frac{n_l + n_s + b}{c^2 (n_u - b)} \sum_{j=1}^{n_u} K_{s,u}(i, j)$, $k_{s,l}(i) = \frac{1}{c^2} \sum_{j=1}^{n_l} G(i, n_s + n_u + j)$,

$k_{u,l}(i) = \frac{1}{c} \sum_{j=1}^{n_l} G(i + n_s, n_s + n_u + j)$ and $c = \frac{n_l + n_s + n_u}{n_u - b}$.

Based on the above expressions, one can draw the following observations regarding the properties of the re-weighted source data and selected query set:

- The first term ensures that the selected query set has minimum similarity within itself, avoiding *redundancy* in the selected set.
- The second term ensures that the re-weighted source instances have minimum similarity within themselves, again avoiding *redundancy* in the re-weighted source set.
- The third term ensures that the selected query has minimum similarity with re-weighted source data thus avoiding *information overlap* between the query and the re-weighted source set.
- The fourth term enforces the selected query instances to be similar to the unselected ones, ensuring *representativeness*.
- The fifth term enforces the re-weighted data to be similar to the unselected ones, ensuring *representativeness* of the target domain data.
- The sixth term implies that the examples with less similarity with already labeled data are more likely to be selected ensuring *diversity* in the selected set.
- The seventh term similarly implies that the re-weighted source data have less similarity with already labeled target data, again ensuring *diversity* in the re-weighted source set.

Thus the proposed method selects examples which meet all the desirable properties for transfer and active learning i.e., representativeness, diversity and avoiding redundancy. The joint optimization framework can potentially avoid the information overlap between the transferred knowledge and the queried knowledge from the target domain, leading to effective transfer and active learning.

The binary constraint on α_i makes the integer quadratic problem defined in Equation (7.3) NP-hard. Similar to *MP-AL* formulation in Equation (6.11), the above integer constraint is relaxed to make Equation (7.3) a continuous optimization problem. Since transfer and active learning are performed simultaneously using a joint optimization framework, the proposed formulation is

referred to as *Joint Optimization based Transfer and Active Learning* (JO-TAL).

$$\min_{X:X_i\in[0,1],X^T\mathbf{B}=b} \quad 0.5X^THX + f^TX$$

$$\text{where } X = \begin{pmatrix}\beta\\ \alpha\end{pmatrix} \quad H = \begin{pmatrix}K_{s,s} & K_{s,u}\\ K_{s,u}^T & K_{u,u}\end{pmatrix} \quad f = \begin{pmatrix}k_{s,l} - k_{s,u}\\ k_{u,l} - k_{u,u}\end{pmatrix}$$

$$B = \begin{pmatrix}O\\ I\end{pmatrix} \qquad I = \mathbf{1}_{n_u\times 1} \qquad O = \mathbf{0}_{n_s\times 1} \tag{7.4}$$

The standard QP can be solved efficiently by applying many existing solvers. The key steps at each iteration are provided in Algorithm 5. The source weight vector, β_{new}, is initialized at the end of first iteration with the vector β. During subsequent iterations the source weights are added (step 7), thus reinforcing source weights or transferred knowledge dynamically, at every iteration. The proposed formulation can be easily extended to include additional evaluation

ALGORITHM 5: JO-TAL

1: **Input:** S: source domain data; L: set of labeled target domain data; U: set of unlabeled target domain data; b: batch size; β_{new} : source weights (input for iteration no. > 1);
2: **Output:** β_{new}: source weights (updated), Q: target query set;
3: Compute H and f as explained in Section 6.2 and 7.2.
4: Compute β and α by solving (7.4).
5: $Q \rightarrow$ top b instances of U, sorted in descending order of α.
6: Update L, U : $L \rightarrow L \cup Q$, $U \rightarrow U \backslash Q$.
7: Update β_{new} : $\beta_{new} \rightarrow \beta_{new} + \beta$ (for iteration no. > 1).

criteria (E_u) by adding a corresponding linear term $E_u^T\alpha$ in Equation (7.4), while still maintaining the quadratic form. In order to incorporate uncertainty of predictions by the existing classifier, a term ($-E_u^T\alpha$) was added, where $E_u(x_i)$ is entropy of predicted labels computed as in [32] for each unlabeled data x_i in target domain, with a weighting factor of n_l/n_u.

7.3 Experiments and Results

Competing Methods. To evaluate the performance of *JO-TAL* the following competing methods were created.

2-Stage based Transfer Active Learning (2S-TAL): In this method, domain adaptation is performed on source data using a transfer learning methodology in the first stage, and then a classifier is learned on the domain adapted source data to use it to actively select most informative instances from the target domain in the second stage, as in [78]. In the experiments described here, an instance re-weighting domain adaptation methodology [[39]] is used and a query set is selected from unlabeled target domain data based on an existing batch-mode active learning

method [[9]] using the classifier learned on the re-weighted source domain data. This method is different from *JO-TAL* as the transfer and active learning are performed in two stages.

Joint Optimization based Transfer Learning (JO-T-Rand): In this method transfer learning is performed on the source domain data and the target domain data is selected randomly for labeling. However, the source weights (β) are computed considering the randomly selected labeled data from target domain, i.e., by minimizing MMD between the re-weighted source (S_a), labeled target domain data (L) and the unlabeled target domain data (U), i.e. between sets $S_a \cup L$ and U. This is achieved by modifying the proposed QP formulation given in Equation (7.2), considering $\alpha = 0$ and $b = 0$.

2-Stage based Transfer Learning (2S-T-Rand): In this method, domain adaptation is performed on source data using a transfer learning methodology [[39]] in the first stage, and instances from the target domain are randomly selected in the second stage.

This dissertation also extended the proposed *JO-TAL* to incorporate an entropy term as explained in Section 6.2 and referred it as *JO-TAL-Ent*.

Finally, the re-weighted source samples are combined with the queried or randomly selected and labeled samples from the unlabeled set of target domain data and the classification accuracy is computed on the fixed test set, from target domain, as explained in Section 6.3.

The above methods provide a basis for comparing the performance of joint optimization method and the traditional method of performing active learning and transfer learning in two stages, under both conditions of actively selected and randomly selected target domain data. The performance of *JO-TAL* is also compared with only active learning using *MP-AL*.

Experimental Procedure. Same as in Section 6.3, the only change in these experiments was that the algorithm started with *no* labeled instances from the unlabeled set. Each algorithm repeatedly performed transfer and active learning, at each iteration, and evaluated the performance of the resulting classifier on the fixed test set, as defined in Section 6.3. Each of the biological image datasets is used as source while the other being target and vice-versa in these experiments.

Comparative Studies. Figure 7.2(a) shows the re-weighted source domain or Fly-FISH instances (size of the triangles is propotional to the weights) and the query set selected from the target domain or BDGP dataset by following the two stage strategy *2S-TAL*. And Figure 7.2(b) shows the

re-weighted source or Fly-FISH instances and the query set selected from target domain or BDGP dataset by the proposed joint optimization framework *JO-TAL*. Figures 7.2(c) and (d) show the results for the cases with BDGP as the source and Fly-FISH as target or test dataset. One can observe that there is considerable amount of information overlap between the instances that get higher weight in source domain and those that get selected from target domain, in the two stage approach (as shown in Figure 7.2(c)). However, when domain adaptation and active learning were done simultaneously, it can be noted that information captured by these two sets is complementary in nature with less information overlap. Also the instances selected by *JO-TAL* are better able to represent the marginal distribution of the target domain data as shown in Figure 7.2(b).

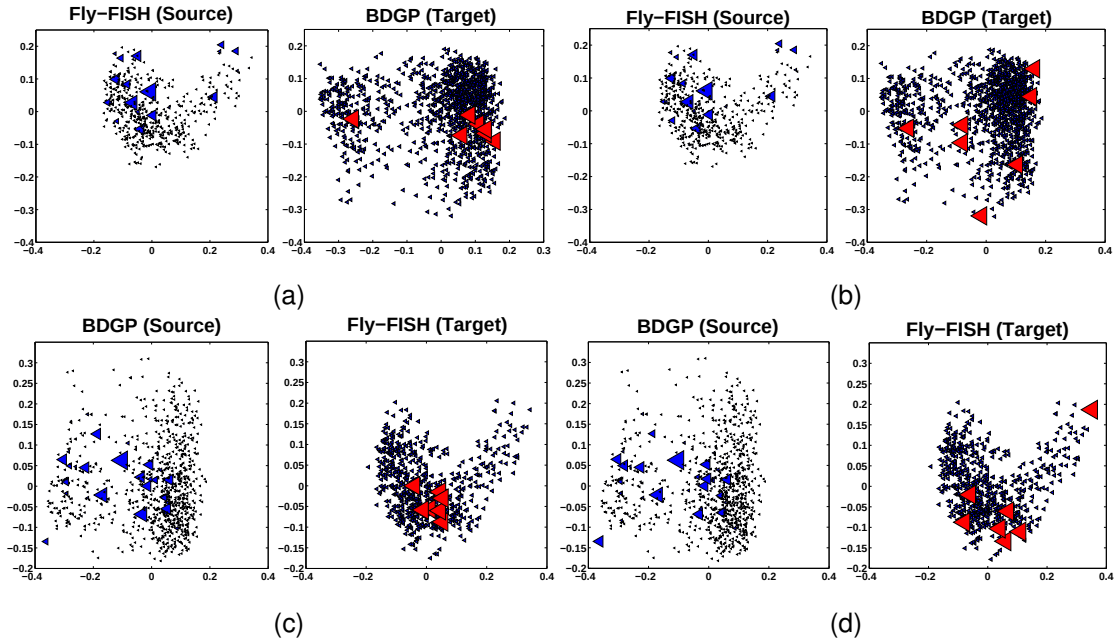
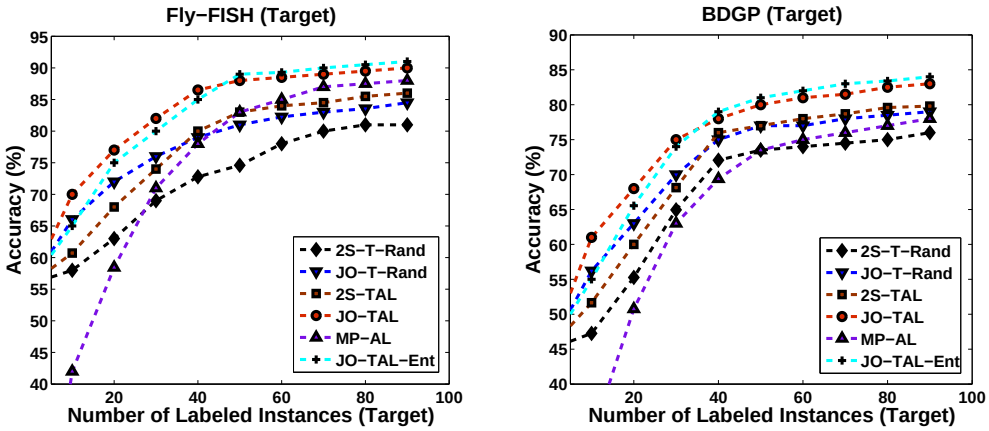


Figure 7.2: Re-weighted instances from Fly-FISH (shown by the size of the data points) and query data points from BDGP (red triangles) based on (a) *2S-TAL* and (b) *JO-TAL*. Re-weighted instances from BDGP and query data points from Fly-FISH (red triangles) based on (c) *2S-TAL* and (d) *JO-TAL*. Figures best viewed in color.

The comparative performance of the proposed method *JO-TAL*, on Fly-FISH and BDGP datasets is shown in Figure 7.3; Figure 7.3 (a) shows the results with Fly-FISH as test (target) set and Figure 7.3 (b) shows the comparative performance with BDGP as test set. The results show that *JO-TAL* performs better than *2S-TAL*. This can be attributed to the efficient transfer and active learning, by selecting complementary samples from the target domain as shown in Figure 7.2. Also *JO-T-Rand* performs better than *2S-T-Rand*. It is however interesting to note that *JO-T-Rand* performs better than *2S-TAL* during initial iterations. However *2S-TAL* improves during the later

iterations with more actively sampled data from the target domain. It can also be noted that the performance of *JO-TAL-Ent* improves towards later iterations, as the classifier becomes more reliable when learned on more labeled data from the test or target domain. The results also show that incorporating transfer learning to active learning (*MP-AL*) improves performance significantly during initial iterations. Thus the proposed joint optimization framework provides a viable solution to the problem of biological image annotation, by effectively using related image databases to develop a classifier for a new database. The performance of *JO-TAL* approach was also evaluated on a benchmark 20 Newsgroups dataset and Sentiment Analysis dataset (see the detailed results in Appendix C).



(a) (b)

Figure 7.3: Comparative performance of proposed joint transfer and active learning method *JO-TAL* with respect to commonly used two stage transfer and active learning method *2S-TAL* on (a) Fly-FISH and (b) BDGP datasets with increasing number of labeled instances from target domain. *JO-T-Rand* and *2S-T-Rand* are joint and two stage transfer and active learning methods with randomly selected target data.

Chapter 8

RELATED WORK

This chapter presents the literature survey on the areas related to the developed methods in this dissertation. A comparison of these methods with the developed methods is also presented.

8.1 Domain Adaptation

Domain adaptation methods have been applied in several applications to reduce the differences in distribution between the training and test data sets. Many existing methods perform domain adaptation based on only marginal probability differences between the two data distributions. Shimodaira et al. [80] biased the training samples by a test-to-training ratio to match the marginal distribution of the test data - in other words, the authors gave a higher weight to the training samples which appeared more frequently in test data. Sugiyama et al. [84] tried to reduce the gap in marginal probabilities by minimizing the KL-divergence between test and re-weighted training data. Bickel et al. [5] achieved the same objective by discriminating training against test data with a probabilistic classifier. Huang et al. suggested a method called *Kernel Mean Matching (KMM)* [39], where the authors re-weighted the samples in source domain so as to *minimize* the marginal probability difference using Maximum Mean Discrepancy (MMD) [8] as the difference measure. Pan et al. [65] suggested a method based on feature mapping for reducing the marginal probability differences between the source and target distributions by minimizing MMD, called as *Transfer Component Analysis (TCA)*.

The proposed domain adaptation frameworks differ from all these methods in two ways: (1) they are based on conditional probability and marginal probability differences (2) they are based on multiple source domains (CP-MDA, 2SW-MDA and HC-MDA).

Several algorithms have been developed in past to combine knowledge from multiple sources. Luo *et al.* used consensus maximization as the basis of combining multiple source data [58]. Mansour *et al.* based the transferability of knowledge on a distribution weighted combination of the hypothesis generated by the independent sources [59]. The theoretical proof of both frameworks are based on strong assumptions on the predictive power of the individual source domains on the target domain data. In [79], a clustering based knowledge transfer was proposed for applications with different class labels across source and target domains, unlike the application addressed in this dissertation.

The proposed frameworks are related to two multi-source domain adaptation frameworks including Domain Adaptation Machine (DAM) [18] and Locally Weighted Ensemble (LWE) [24]. The proposed framework differs from DAM in the way the weights are computed for different auxiliary sources. In DAM, the weight assigned to each auxiliary source is obtained by measuring the marginal probability distribution difference between the target domain and the particular auxiliary source only, using the empirical estimate of the difference based on the Maximum Mean Discrepancy measure [8]. The proposed frameworks however computes weights for the auxiliary source data considering predominantly conditional probability distribution of the target data. The weights for all sources are computed in a joint optimization framework (CP-MDA and 2SW-MDA), which takes the interaction among multiple auxiliary sources into account.

Locally Weighted Ensemble (LWE), suggested by Gao et al. [24] addressed conditional probability difference by comparing the clustering manifold of the test or target domain data around a target data sample with the manifold formed by the labels generated by the training or source domain data. The proposed frameworks differ from LWE [24] in that in LWE, the label y of an unlabeled target domain data x is computed using a local weighting ensemble (LWE) scheme as follows. Different from the proposed weighting scheme where we compute all weights in a joint framework (CP-MDA, 2SW-MDA), the weight for each auxiliary classifier is computed independently [24].

This dissertation also compares the developed algorithms with representative single-source domain adaptation algorithms such as Kernel Mean Matching (KMM) proposed by Huang *et al.* [39], Transfer Component Analysis (TCA) proposed by Pan *et al.* [65] and KMapEnsemble (KE) proposed by Zhong *et al.* [95]. KMM re-weights the samples in the source domain so as to minimize the marginal probability difference between the source and target domain using Maximum Mean Discrepancy (MMD) as the measure. TCA is based on feature mapping so as to reduce the marginal probability differences between the source and target distributions again using MMD as the measure. KE differs from the first two algorithms, in which it addresses the conditional probability differences by sample selection after performing a feature mapping step to reduce the marginal probability differences.

There has been little effort in addressing both marginal and conditional probability distribution differences. Most of these approaches involve a domain mapping to reduce the marginal probability differences, followed by sample selection in the mapped domain to address

the conditional probability differences. One such method proposed by Zhong et al. [95] called *KMapEnsemble (KE)*, performs domain adaptation using Kernel Discriminant Analysis (KDA), followed by clustering based sample selection. One drawback with this approach is the increase in entropy of labels due to the mapping of the distributions in a common representation [42].

In this dissertation the proposed methods are compared with five different domain adaptation methods published in literature: KMM, TCA, DAM, LWE and KE. Three methods are based on marginal probability differences e.g., *DAM*, *KMM* and *TCA* , one which is based only on conditional probability differences e.g., *LWE* and another which addresses both marginal and conditional probabilities e.g., *KE*.

A brief description of each of these methods that are used for comparison, is presented in this chapter.

Kernel Mean Matching (KMM)

This method re-weights source domain samples such that the differences in marginal probability distribution between the re-weighted source domain data x_i and the target domain data x_i' defined by Maximum Mean Discrepancy (MMD), is minimized. The minimization problem is defined as follows:

$$\min_{\beta} \left\| \frac{1}{m} \sum_{i=1}^{m} \beta \phi(x_i) - \frac{1}{m'} \sum_{i=1}^{m'} \phi(x_i') \right\|_H^2 \tag{8.1}$$

where $\phi(x)$ is a Gaussian kernel function in universal reproducible kernel Hilbert space(RKHS) H [82], m is the number of samples in the source domain, m' is the number of samples in target domain and β is the $m \times 1$ weight vector that minimizes the cost function. A classifier for the target domain is then learned on the re-weighted samples from source domain.

Transfer Component Analysis (TCA)

This method learns a kernel in the mapped domain such that the marginal distribution differences between the source and target domains is reduced using MMD as follows:

$$\min_{\phi} \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_{S_i}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_{T_i}) \right\|_H \tag{8.2}$$

where $X_S = \{x_{S_1} \cdots x_{S_{n_1}}\}$ and $X_T = \{x_{T_1} \cdots \cdots x_{T_{n_2}}\}$ are source and target domain data respectively. ϕ is the feature map induced by a universal kernel such that $\phi : X \to H$ and $\phi(x_{S_i})$ and $\phi(x_{T_i})$ are the corresponding features mapped into the RKHS. Pan et al. proposed learning a feature map such that MMD between X_S and X_T is minimized.

Kernel Ensemble (KE)

This method, suggested by Zhong et al., follows a two step strategy. The first step obtains a common feature mapping for the training and test data by performing Kernel Discriminant Analysis (KDA) [95] using the few labeled test data available. This feature mapping is used to map the training (source) and unseen test data (target) into a common domain. And the second step applies a cluster based instance selection method to select training samples which have similar conditional probability distribution as the test data.

Domain Adaptation Machine (DAM)

The difference between the proposed method and the Domain Adaptation Machine (DAM) method suggested by Duan et al. [18] mainly lies in the way the weights are computed for different auxiliary classifiers to generate the psuedo labels for the unlabeled target domain data. As per DAM, the weights assigned to the hypothesis provided by each auxiliary source is obtained by measuring the difference in marginal probability distribution between the target domain and the particular auxiliary source using an empirical estimate of the difference defined by Maximum Mean Discrepancy measure [8]. As per this measure the difference between the two distributions $X_S = \{x_{S_1} \cdots\cdots x_{S_{n_1}}\}$ and $X_T = \{x_{T_1} \cdots\cdots x_{T_{n_2}}\}$ with distributions P and Q is given by

$$Dist(X_S, X_T) = \left\| \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_{S_i}) - \frac{1}{n_2} \sum_{i=1}^{n_2} \phi(x_{T_i}) \right\|_H \tag{8.3}$$

where x_{S_i}, x_{T_i} are data from X_S and X_T, and H is a universal RKHS [82] and ϕ is the feature map induced by a universal kernel such that $\phi : X \to H$ and $\phi(x_{S_i})$ and $\phi(x_{T_i})$ are the corresponding mapped features into the RKHS. The weights for each auxiliary source is obtained by

$$\gamma_S = \frac{\exp(-\beta(Dist(X_S, X_T))^2)}{\sum_s \exp(-\beta(Dist(X_S, X_T))^2)}$$

The proposed framework computes weights based on smoothness assumption on target domain data thus considering both marginal and conditional probability differences with the target data. Also the weights are computed in a joint optimization framework and not one at a time, to take care of their mutual dynamics. Besides the proposed framework provides for assigning differential weights to the psuedo labels generated by the auxiliary classifiers, thus biasing the target classifier towards the true labels of the labeled target domain data.

Locally Weighted Ensemble (LWE)

The LWE framework computes the label y of an unlabeled target domain data x, using k source models and an ensemble weighting scheme as follows:

$$P(y|x) = \sum_{i=1}^{k} w_{M_i,x} P(y|M_i, x) \tag{8.4}$$

where $P(y|M_i, x)$ is the prediction made by one of the k models M_i for target data point x and $w_{M_i,x}$ is the weight of the model M_i at point x obtained as follows:

$$w_{M_i,x} = \frac{s(G_{M_i}, GM_T : x)}{\sum_{i=1}^{k} s(G_{M_i}, GM_T : x)} \tag{8.5}$$

where G_{M_i} and G_T are the graphs around point x. G_T is built by connecting points belonging to the same cluster as x in the test data, and G_{M_i} is built by connecting points belonging to the same class as x, as determined by the i-th model, M_i. $s(G_{M_i}, GM_T : x)$ is the measure of similarity between these two graphs at point x in test data. For more details the reader is requested to refer [24].

The proposed methods are different from DAM, KMM, TCA and LWE as they address both marginal and conditional probability differences between the source and target domains. And unlike KE, ODA addresses both these distribution differences based on a sample selection strategy, rather than feature mapping, and it does so using an unified optimization framework rather than addressing these differences using a two-step strategy. Also KE method does not preserve the topology of the input high dimensional data in the new mapped domain and second it expects the mapped data to have a clustering manifold, where as in the proposed TPDA method the original topology of the data is preserved, and the mapped data is not assumed to be of any specific manifold.

There was some classification work dealing with physiological signals using neural networks [54] and linear discriminant analysis [47]; they achieved moderate generalization performance across subjects. This dissertation reports the first systematic empirical analysis of domain adaptation methods to address the distribution differences due to the subject based variability in physiological signals such as electromyogram (EMG).

8.2 Electromyogram signals (EMG) as fatigue indicator

The electromyography (EMG) is a biosignal recording of the skeletal muscle activity of the body. It is routinely used by clinicians for analysis of the skeletal muscle activity. EMG may be

recorded from the surface of the skin without any invasion of the body known as surface EMG (SEMG). There are number of techniques that can be used to objectively determine the level of fatigue in a subject. The most reliable of these is the direct measurement of chemical properties in the muscle of the subject. Since this is an invasive technique it is inappropriate for routine utilization, away from the clinical environment. SEMG provides a non invasive way to identify fatigue. Indication of localized muscle fatigue has been frequently based on the observed shift of the power spectral density of the SEMG [51], [13], [25], [56], [72]. Several parametric measures of SEMG signal have been used as a relative indicator of the muscle fatigue phenomenon for an individual subject. These include the root mean square (rms), instantaneous frequency, zero crossing rate, mean-frequency, and median-frequency.

In general there is a large variation in these measures due to variance in SEMG power spectrum and it's shift for different subjects. Hence most of the work done in past towards quantification of fatigue from SEMG has been very subject specific. Contessa et al. [13] collected data from 4 subjects - once each at the beginning and conclusion of a fatiguing exercise and observed a significant difference between the data patterns collected from different subjects. Gerdle et al. [26] observed variations in the root mean square (RMS) of the EMG signal across different subjects even when the subjects performed the same activity under similar experimental conditions. Investigations into the use of physiological data for recognizing the different emotional states of a person have reported a subject independent classification accuracies of the order of 70% [54], [47].

Researchers have attempted to track the changes in electromyographic (EMG) features of the involved muscles that can correlate with an increase in muscle fatigue [22, 6, 72, 51, 13]. Typically, during a sustained submaximal fatiguing contraction the mean amplitude of the EMG increases while the power spectrum of the EMG signal shifts to lower frequencies [71]. These changes are consistent with the recruitment of additional motor units and a decrease in the conduction velocity of action potentials along the muscle fibers [67], respectively. It has been proved in literature that decrease in muscle pH during fatiguing contractions, decreases the conduction velocity of action potentials, resulting in compression of the power spectrum towards lower frequencies [67]. However, there is often a large amount of inter-subject variability in such features [13], [26] and these phenomena can also be affected based on the type of contraction performed or electrode location [28]. These generally unpredictable variations across subjects

make it difficult to develop a general framework that can grade several EMG features from a subject without prior data from that same subject.

Because of the large inter-subject variability in EMG features it may not be possible to reduce myoelectric information into a single EMG parameter. Moreover, only EMG amplitude parameters may not be able to reliably track the myoelectric manifestations of fatigue [52]. However, it may be possible to track the myoelectric manifestations of fatigue by observing multiple EMG amplitude and spectral parameters, collectively. Hence, multiple EMG features were used to develop the subject-independent generalized frameworks. Many studies have only used two or three EMG features, namely root mean square, mean frequency and median frequency [26, 6, 72].

Although studies have attempted to track multiple features from the EMG signal during fatiguing contraction [83, 87] [27], these frameworks require baseline data from the test subject. Thus, these frameworks are not subject-independent. Guler et al. [27] classified EMG signals using two alternative methods: (1) a Support Vector Machine (SVM) and (2) a Multilayer Perceptron (MLP). They first extracted FFT coefficients from the EMG signals, and then performed PCA for data reduction. Subasi et el [83] classified isometric SEMG signals using an Artificial Neural Network (ANN) and ICA for dimension reduction. Torvik et al. [87] proposed a new model (called OCAT) for distinguishing between static fatigue and non-fatigue state EMG signals, on the basis of shifts in certain fractal frequencies, and the peak frequency. Others [69] have suggested to have developed a subject-independent fatigue grading index [generalized mapping index (GMI)] based on a mapping function derived from multiple features. These claims are based on the fact that GMI has better signal-to-noise ratio, compared to commonly used estimators of fatigue such as mean frequency of EMG signals. In the proposed EMG feature grading framework, the subject based variability of the suggested measure was quantified and is shown to be an order less than the subject based variability of the commonly used features such as mean frequency and median frequency. Moreover, GMI is based on four time domain features, that are used to train a Artificial Neural Network based on an assumption of linear relationship between muscle fatigue and time. Where as, the proposed method derives its grading index based on both time and frequency domain features and do not assume any specific trend in progression of fatigue.

The proposed EMG feature grading framework uses factor analysis to learn the relation between the EMG features during a fatiguing contraction, rather than learning the values of the

features. Latent factors, obtained as a result of factor analysis on multiple features, represent the relation between the features. An analysis of the subject-based variability of each of the commonly used frequency and time domain features and latent factors revealed that the latent factors are more robust across multiple subjects.

8.3 Batch-mode Active Learning

Matching marginal probability distribution between the training and test data, has been widely used for transfer learning applications. [80] biased the training samples by their test-to-training ratio to match the marginal distribution of the test data. [84] proposed to reduce the gap in marginal probabilities by minimizing the KL-divergence between test and weighted training data and [5] discriminated training against test data with a probabilistic model that accounts for the marginal probability difference between training and test distribution. There are several other methods which are also based on marginal probability differences using Maximum Mean Discrepancy [8] as a measure such as Kernel Mean Matching [39] and Transfer Component Analysis [64]. The proposed framework, *MP-AL*, however differs from all these methods, as it matches marginal probability distributions, not for transfer learning, but for active learning applications.

This dissertation compared the performance of the proposed *MP-AL* method with state-of-the-art batch-mode active learning methods including *Matrix* [31], *Disc* [32] and *Fisher* [35] which selected a set of instances that are together maximally informative, similar to the proposed approach. Besides, it compared with state-of-the-art batch-mode active learning methods which selected a set of instances in each iteration based on their individual merits such as *svmD* [9] and a multiple criteria based instance selection method [90], referred to in this dissertation as *MCS* for convenience. *MP-AL* method was also compared to one transductive experimental design method [92], referred to as *Design*, which is based on regression models. A brief review of each of these methods is presented below.

Disc

The *Disc* method selects a set of instances by maximizing the likelihood of labeled and selected instances, while minimizing the uncertainty of unlabeled instances, based on a classifier learned on the labeled and selected instances. The formulation selects S which posses one set of label

configuration $\mathbf{y}_S^*$ that maximizes the following equation:

$$\begin{aligned} f(S) = \quad & \max_{\mathbf{y}_S} \sum_{i \in L^t \cup S} logP(y_i|x_i, w^{t+1}) + \\ & \alpha \sum_{j \in U^t \setminus S} \sum_{y=\pm 1} P(y|x_j, w^{t+1}) logP(y|x_j, w^{t+1}) \end{aligned} \tag{8.6}$$

The problem formulation is non-convex and a local solution is obtained using the gradient descent method.

Matrix

The *Matrix* method selects a batch of queries S in each iteration by maximizing a mutual information criterion between the selected and labeled instances ($L' = L \cup S$) and unlabeled instances ($U' = U \setminus S$), based on multivariate Gaussian distribution as follows:

$$S^* = \arg \max_{|S|=b, S \subseteq U} ln|\Sigma_{L'L'}| + ln|\Sigma_{U'U'|}. \tag{8.7}$$

where $\Sigma_{U'U'}$ and $\Sigma_{L'L'}$ are covariance matrices of U' and L' computed using Gaussian kernel. Similar to the *Disc* method, this formulation is non-convex and a local solution is obtained using the gradient descent method. Similar to proposed approach, this method does not depend on any classifier model. This strategy selects S that maximizes the log determinants of the covariance matrices L' and U'.

Fisher

The *Fisher* [35] method selects samples using Fisher information as the criterion. It selects a set of instances such that the difference in Fisher information between the selected set and unlabeled examples is minimum. The formulation is solved using a greedy algorithm. This method depends upon a classifier model to compute the Fisher information. The proposed approach is similar to the *Matrix* method as it does not depend on any classifier model, unlike *Disc* and *Fisher* methods.

$$\begin{aligned} x^* \quad & = \arg\max_{x \notin L} (f(L \cup x) - f(L)) \\ where \quad & f(L \cup x) - f(L) \\ = \quad & g(x, L) + \sum_{x' \notin (L \cup x)} g(x', L) g(x\, L \cup x)(x^T x')^2 \\ where & \\ g(x, L) \quad & = \frac{\pi(x)(1 - \pi(x))}{\sum_{x' \in L} \pi(x')(1 - \pi(x'))(x^T x')^2} \\ and & \\ \pi(x) \quad & = p(-|x) = \frac{1}{1 + exp(w^T x)} \end{aligned} \tag{8.8}$$

svmD

The *svmD* method [9] selects a set of uncertain and diverse instances for query by ranking each instance in the unlabeled data based on their distance from the margin and maximum angle with the already labeled samples. The angle between two samples is measured using cosine of the angles between the hyperplanes corresponding to the samples and is given by $\frac{x_i.x_j}{\sqrt{\|x_i\|\|x_j\|}}$ as follows:

$$i^* = \arg\min_{i \in U} \left(\lambda |g(x_i)| + (1-\lambda) \max_{j \in L} k(x_i, x_j) \right). \tag{8.9}$$

where $k(x_i, x_j) = \frac{x_i.x_j}{\sqrt{\|x_i\|\|x_j\|}}$.

Similar to *svmD*, *MCS* [90] evaluates instances based on their individual merit using multi-criteria, but added a third term to measure the representativeness of each unlabeled data based on average cosine similarity with the unlabeled data.

QUIRE

Besides, the above batch-mode active learning methods, there exists a state-of-the-art, single instance selection method, *QUIRE* [40] which also ranks instances based on multi-criteria such as uncertainty in prediction and representativeness, but measured them using discriminative models learned using labeled and unlabeled data instead of cosine based similarities. The informativeness of an instance is measured by its prediction uncertainty based on labeled data, while the representativeness is measured by its prediction uncertainty based on unlabeled data. The instance with minimum confidence of prediction is selected for query. The problem is formulated as below

$$\mathcal{L}(L, U, x_i) = \min_{y_u \in \{\pm 1\}^{n_{u-1}}} \max_{y_i = \pm 1} \min_{f \in H} \frac{\lambda}{2} |f|_H^2 + \sum_{i=1}^{n} l\left(y_i, f(x_i)\right). \tag{8.10}$$

where y_u are the predicted labels of unlabeled data U, based on the classifier learned on the labeled data.

The proposed *MP-AL* method based on distribution matching, addresses both diversity and representativeness, besides addressing redundancy as well. It is also different from *svmD*, *MCS*, [90] and *QUIRE* as it selects a batch of instances simultaneously which are together maximally informative based on their collective merit.

Transductive Methods based on Optimum Experimental Design Techniques.

The *Design* method [92] proposes an experimental design in a *transductive* setting, where the focus is on the predictive performance on known test data. This method selects a set of instances closest to the basis vectors that approximate the set of unlabeled data. The problem is formulated into a regularized least square problem and is independent of evaluation classification models. This method does not consider already labeled data, when selecting the next set of query, unlike the proposed method.

However, there are some more transductive methods in literature which are based on optimum experimental design (OED) techniques [93, 10, 34]. The typical OED criteria minimize the variance of the parameter estimates or predicted value. [93] selected those instances which can be used to best reconstruct the whole data set. [10] proposed a method based on a data manifold adaptive kernel space and [34] suggested a subspace learning method.

Zhang et al. [93] proposed an active learning method which selects a set of most representative points whose coordinates can be used to best reconstruct the whole data set. This method is based on local manifold structure, hence given the local reconstruction coefficients for every data point and the coordinates of the selected points, a transductive learning algorithm called Locally Linear Reconstruction (LLR) is proposed to reconstruct every other point. The most representative points are defined as those whose coordinates can be used to best reconstruct the whole data set. Cai et al. [10] proposed an active learning method based on a data manifold adaptive kernel space, which reflects the underlying geometry of the data. By minimizing the expected error with respect to a classifier learned in this kernel space, most representative and discriminative data points are selected for labeling. He et al. [34] suggested an active subspace learning method based on the connection between the subspace learning and linear regression. The projection vector is expressed in terms of the data points and their labels and for any data point its expected predictive error is determined by the covariance matrix of the projection vector. Using techniques from experimental design, those data points are selected such that the expected predictive errors over all the other data points are minimized. These methods are different from the proposed method, with respect to their transductive setting and also similar to *Design* method [92], these methods do not consider already labeled data, when selecting the next batch of query instances, at each iteration.

8.4 Transfer and Active Learning

There has not been much prior work towards combining of transfer and active learning methodologoies. A combination of transfer learning with active learning has been presented by Shi. et al. [78]. As per this method, a classifier is learned on the source domain data and another classifier is learned on an initial pool of labeled target domain data. Label for an unlabeled instance is predicted by both the classifiers. Then based on a decision function evaluated based on the confidence of predictions of each of the classifiers, the algorithm decides whether to accept the label provided by the classifiers or get it queried for manual annotation, using an already published active learning methodology. One drawback of this approach is the requirement of an initial pool of labeled target domain data used to train an initial target classifier. Without this initial target classifier, no transfer learning is possible in this setting. Another drawback is that the source data is used without any domain adaptation.

Another method, suggested in literature by Rai et al. [68], uses multiple classifiers to perform transfer and active learning. Firstly, it trains a classifier to distinguish between source and target domain data. Then those target domain data which has been classified as source domain data are labeled using a classifier trained only on source domain data. Other target domain data are classified by a classifier trained on domain adapted source and labeled target domain data, and the uncertainty of labeleing is measured using entropy of labeling. Those target domain data with most uncertainty of labeling are queried. A drawback of this method is that the domain adaptation is done once initially and as more labeled data is obtained from target domain, the domain adaptation is not refined. Besides, similar to the previous method, this method too performs active and transfer learning independent of each other using existing active and transfer learning algorithms.

Unlike both these methods, the proposed *JO-TAL* method performs transfer and active learning in a combined framework and domain adaptation is updated at every iteration as more and more target domain data gets queried. Also *JO-TAL* similar to *MP-AL* does not require initial pool of labeled data from target domain for it's operation. However, there is a combined transfer and active learning method suggested by Chen et al. [12], based on the assumption that the target domain may have unique features, and hence selects a subset of shared source and target features based on a non-convex optimization problem.

Chapter 9

CONCLUSION

Machine learning is one of the fastest growing areas of computer science research. Search engines, recommendation systems, spam filtering and fraud detection, computational social science, market prediction, speech and handwriting recognition, natural language processing, vision systems, DNA sequence analysis and health care are just some of the applications in which machine learning is routinely used. Despite the wide application of machine learning, efficient deployment of its principles on real world data is still a challenge. Some of the real world challenges include missing data, corrupted data, heterogeneous data, noisy data and last but not the least, distributional differences in data. This dissertation successfully addressed the key aspects related to distributional differences in data. Specifically, this dissertation addressed the distributional differences in physiological and biomedical data caused due to subject based and technology based variabilities, using novel transfer learning and active learning methods. This dissertation has several novel contributions in the area of domain adaptation and transfer learning and in the area of batch-mode active learning, besides proposing a first joint framework for both methodologies.

9.1 Discussion

Despite the empirical success of these methods, there remains much future work to be done in the area of domain adaptation for developing person adaptive systems, to the point of being useful in commercial systems. Some of the aspects being scalable and efficient algorithms for selection of data sources with similar distributions, the design of the similarity measure and the methods for knowledge transfer or domain adaptation in general. This dissertation was originally motivated for developing adaptive systems for physiological and biomedical image data, which seeks to transfer knowledge from annotated data belonging to multiple subjects or entities (source), to develop a classifier for a new test subject or entity (target), for whom little annotated data is available, by adapting the distribution of the source data as per the target data distribution.

While successful, each of these methods suffers from a fundamental drawback. They rely on the availability of a pool of data from the test subject or target domain, though unlabeled, for domain adaptation: either for source selection as in CP-MDA (Section 3.1), instance re-weighting as in 2SW-MDA (Section 3.2) and HC-MDA (Section 3.3), instance selection as in ODA (Section 3.4) or for feature mapping as in TPDA (Section 3.5). This dependency introduces an initial delay

in the computational framework, making them unsuitable for applications that require near real-time transfer of knowledge or domain adaptation. Faster methods of domain adaptation, with minimum data from the target domain are necessary for real-time operations. Regardless of the mechanism, annotation being a time consuming and expensive process, specifically for physiological and biomedical data, as it is done by experts, knowledge transfer from available pool of already annotated data, is a lofty goal for machine learning, worthy of future study. Person adaptive systems based on knowledge transfer show promise for dramatically improving the performance and practical application of machine learning.

9.2 Future Work

This dissertation establishes a foundation for several avenues of future work in machine learning and data mining by enabling novel and efficient applications of machine learning algorithms. Some of the evolving applications being scalable person adaptive systems, dynamic knowledge transfer and 'Big data analytics' for health care and business intelligence. The novel concept of joint transfer and active learning has paved the way for efficient learning models with improved performance and new learning theories analyzing these frameworks. Some of these future directions are outlined in this section.

Pre-sampling of Sources to Increase Scalability: In order to scale the domain adaptation algorithms for cases where the number of subjects or sources is very large, a preliminary sampling to reduce the number of subjects or the sources would greatly reduce the computational burden and enable real-time operation. Sampling of instances of each source for the algorithms involving re-weighting of source instances, as in 2SW-MDA, HC-MDA and ODA methods, would additionally enhance the scalability of the algorithms, specially for the cases with a large number of source instances.

Dynamic Knowledge Transfer: Another potential direction for future work is dynamic knowledge transfer. This refers to the applications where the distribution of the test data does not remain static, hence domain adaptation needs to be performed dynamically, with the changing distribution of the test data. In all the cases discussed in this dissertation, domain adaptation is done once and the classifier remains fixed once learned on the domain adapted data. Dynamic knowledge transfer is still an open problem. A typical application is stock market data.

Transfer-Active Learning for 'Big data analytics': Recently, processing of very high volume of high dimensional data of a variety of types to uncover hidden patterns, unknown correlations and

other useful information, has become a very hot area and is referred to as 'Big data analytics'. The primary goal of big data analytics is to analyze huge volumes of transaction data as well as other data sources that may be left untapped by conventional business intelligence such as web server logs, Internet click stream data, social media activity reports, mobile-phone call detail records and information captured by sensors. Presently, the focus of 'Big Data analytics' is the development of the software/hardware platforms that would enable storage, processing and fast access of the huge volume of unstructured data in the form of text (email messages etc) and images. The intelligent processing of this data would soon become the next big challenge of 'Big data analytics'. The transfer and active learning algorithms presented in this dissertation can be extended to address 'Big data analytics' settings, capable of transferring knowledge across heterogeneous and unstructured data sources, having different feature spaces. An interesting research direction would be the application of these frameworks to other real world problems such as missing and corrupted data. It would also be very interesting to explore the application of these frameworks to multi-label data.

Transfer-Active Learning for Improved Adaptive Systems: It has been proven theoretically that active learning requires a smaller number of labeled samples compared to passive learning for achieving the same performance level [33, 16]. However, there are certain significant aspects of learning algorithms, one of them being self verification. It turns out that self-verification is not possible under active learning settings, due to scarcity of labeled data. A potential future work would be to explore whether transfer learning can help bridge this gap by enabling self-verification in active learning algorithms and obtain the same dramatic improvements over passive learning as can be achieved by their non-self-verifying counterparts.

Just as transfer learning may be used to address deficiencies in active learning, similarly active learning can be exploited to address some of the problems of transfer learning. Experiments with transfer learning as well as existing literature show that transfer of knowledge from a related data source does not always improve the performance of a classifier for a target domain. This is referred to as *negative* transfer [70]. The *negative* transfer can be due to inadequate source knowledge exploration or due to missing information. Similar to application of transfer learning to improve active learning , a potential future work would be to explore whether active learning can be useful in detection and analysis of the causes of negative transfer, in transfer learning. Thus

the frameworks based on transfer and active learning would enable adaptive systems to succeed in situations where they would normally fail due to lack of appropriate or sufficient knowledge.

Theoretical Analysis of Transfer-Active Learning: Existing literature presents theoretical performance bounds for transfer learning [4, 59] and theoretical analysis of sample complexity with respect to active learning [33, 16]. There is not much existing work that analyzes the performance of a classifier based on both transfer and active learning. A potential future work is to analyze the performance of such a classifier and also its sample complexity compared to a classifier based only on active learning or transfer learning.

Chapter 10

PUBLICATIONS

10.1 Journal Publications

1. **Rita Chattopadhyay**, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, Jieping Ye. Batch Mode Active Sampling based on Marginal Probability Distribution Matching. **Invited Paper**. To be published in *ACM Transactions on Knowledge Discovery from Data (TKDD)- Special Issue on the Best of SIGKDD 2012*. 2013.

2. **Rita Chattopadhyay**, Mark Jesunathadas, Brach Poston, Marco Santello, Jieping Ye, Sethuraman Panchanathan. A Subject-Independent Method for Automatically Grading Electromyographic Features during a Fatiguing Contraction. *IEEE Transactions on Biomedical Engineering (TBME)* Jun; 59(6), pages 1749-1757. 2012.

3. **Rita Chattopadhyay**, Qian Sun, Sethuraman Panchanathan, Wei Fan, Ian Davidson, Jieping Ye. Multi-Source Domain Adaptation for Early detection of Fatigue using Surface Electro-myogram Signals. **Invited Paper**. *ACM Transactions on Knowledge Discovery from Data (TKDD)- Special Issue on the Best of SIGKDD 2011* 6(4):18:2012.

10.2 Refereed Conference Publications

1. **Rita Chattopadhyay**, Zheng Wang, Wei Fan, Ian Davidson, Sethuraman Panchanathan, Jieping Ye. Batch Mode Active Sampling based on Marginal Probability Distribution Matching. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. pp. 741-749. Full presentation. **1 of 8 Best papers out of 744 papers**. 2012.

2. **Rita Chattopadhyay**, Jieping Ye, Sethuraman Panchanathan, Wei Fan, Ian Davidson. Multi-Source Domain Adaptation for Early detection of Fatigue using Surface Electro-myogram Signals. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. pp.717-725. Full presentation (7.84% acceptance rate). **1 of 7 Best papers out of 714 papers**. 2011.

3. Qian Sun, **Rita Chattopadhyay**, Sethuraman Panchanathan, Jieping Ye, A Two-Stage Weighting Framework for Multi-Source Domain Adaptation. In *Proceedings of the Twenty-Fifth Annual Conference on Neural Information Processing Systems (NIPS)*(21.78% acceptance rate). 2011.

4. **Rita Chattopadhyay**, Narayanan C Krishnan, Sethuraman Panchanathan. Hierarchical Domain Adaptation for SEMG Signal Classification across Multiple Subjects. In *Proceedings of IEEE Engineering in Medicine and Biology Society (EMBC)*. pp. 7853-7856. 2011.

5. **Rita Chattopadhyay**, Shayok Chakroborty, Vineeth Subramanian and Sethuraman Panchanathan. Optimization-based Domain Adaptation towards Person-Adaptive Classification Models. In *IEEE International conference on Machine Learning and Applications (ICMLA)*. Vol. 1, pp.476-483. 2011.

6. **Rita Chattopadhyay**, Narayanan C Krishnan, Sethuraman Panchanathan. Topology Preserving Domain Adaptation for addressing Subject based Variability in SEMG signal. In *AAAI 2011 Spring Symposium on Computational Physiology (AAAI Spring Symposium)*. 2011.

7. **Rita Chattopadhyay**, Gaurav Pradhan, Sethuraman Panchanathan. Subject Independent Computational Framework for Myoelectric Signals. In *IEEE International Instrumentation and Measurement Technology Conference (IIMTC)*. pp. 1-4. 2011.

8. **Rita Chattopadhyay**, Gaurav Pradhan, Sethuraman Panchanathan. Towards Fatigue and Intensity Measurement framework during continuous repetitive activities. In *IEEE International Instrumentation and Measurement Technology Conference (IIMTC)*. pp. 1341-1346. 2010.

9. **Rita Chattopadhyay**, Jieping Ye, Sethuraman Panchanathan. A Generalized Machine Learning framework based on Transfer Learning techniques for early detection of Fatigue using Surface Electro-myogram signals (SEMG). In *WIML workshop organized by NIPS*. 2010.

10. Gaurav Pradhan, **Rita Chattopadhyay**, Sethuraman Panchanathan, Processing Body Sensor Data Streams for Continuous Physiological Monitoring, In *ACM SIGMM conference on Multimedia Information Retrieval 2010 (ACM SIGMM)*. pp. 479-486. 2010.

11. **Rita Chattopadhyay**, Gaurav Pradhan, Sethuraman Panchanathan. A Generalized Machine Learning framework for Continuous Monitoring of physiological conditions based on Fatigue and Intensity of activity in Daily Living. In *WIML workshop organized by NIPS*. 2009.

10.3 Talks

1. Presented research work: 'Subject based variability in Surface Electromyogram Signals', *Doctoral Forum, SIAM Conference on Data Mining (SDM)* 2011.

2. Guest lecture on *Introduction to Transfer Learning* to the class 'CSE 591: Machine Learning and Applications' at Arizona State University. 2011.

REFERENCES

[1] SVM light, http://svmlight.joachims.org (2002) by T. Joachims.

[2] P. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3:463–482, 2002.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 7:2399–2434, 2006.

[4] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Vaughan. A theory of learning from different domains. *Journal of Mach Learn*, 79:151–175, 2010.

[5] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *JMLR*, 10:2137–2155, 2009.

[6] B. Bigland-Ritchie, E. F. Donovan, and C. S. Roussos. Conduction velocity and EMG power spectrum changes in fatigue of sustained maximal efforts. *J Appl Physiol*, 51(5):1300–1305, 1981.

[7] J. Blitzer, M.Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL*, 2007.

[8] K. Borgwardt, A. Gretton, M. Rasch, H. Kriegel, Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):49–57, 2006.

[9] K. Brinker. Incorporating diversity in active learning with support vector machines. In *ICML*, 2003.

[10] D. Cai and X. He. Manifold adaptive experimental design for text categorization. *IEEE Trans. on Knowledge and Data Engineering*, 24(4):707–719, April 2012.

[11] C. Campbell, N. Cristianini, and A. Smola. Query learning with large margin classifiers. In *ICML*, 2000.

[12] M. Chen, K. Weinberger, and J. Blitzer. Co-training for domain adaptation. In *NIPS*, 2011.

[13] P. Contessa, A. Adam, and C. J. D. Luca. Motor unit control and force fluctuation during fatigue. *Journal of Applied Physiology*, 9(5):337–350, July 2009.

[14] E. Cutrell and D. Tan. Bci for passive input in hci. In *CHI*, 2008.

[15] I. Dagan and S. Engelson. Committee-based sampling for training probabilistic classifiers. In *ICML*, 1995.

[16] S. Dasgupta. Coarse sample complexity bounds for active learning. In *NIPS*, 2005.

[17] H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

[18] L. Duan, I. W. Tsang, D. Xu, and T. S. Chua. Domain adaptation from multiple sources via auxiliary classifiers. In *ICML*, pages 289–296, 2009.

[19] E. Eaton and M. desJardins. Set-based boosting for instance-level transfer. In *IEEE International Conference on Data Mining Workshops*, 2009.

[20] R. M. Enoka and J. Duchateau. Muscle fatigue: what, why and how it influences muscle function. *Journal of Physiology*, 2008.

[21] R. S. et al. Sensorimotor eeg patterns during motor imagery in hemiparetic stroke patients. *International J. Bioelectromagnetism*, 2007.

[22] G. Filligoi and F. Felici. Detection of hidden rhythms in surface EMG signals with a non-linear time-series tool. *Medical Engineering & Physics*, 21(6-7):439–448, Jul 1999.

[23] Y. Freund, H. Seung, E. Shamir, and N. Tishby. Selective sampling using query by committee algorithm. *Mach. Learn.*, 28, 1997.

[24] J. Gao, W. Fan, J. Jiang, and J. Han. Knowledge transfer via multiple model local structure mapping. In *KDD*, pages 283–291, 2008.

[25] A. Georgakis, L. Stergioulas, and G. Giakas. Fatigue analysis of the surface EMG signal in isometric constant force contractions using the averaged instantaneous frequency. *Biomedical Engineering, IEEE Transactions on*, 50(2):262–265, 2003.

[26] B. Gerdle, B. Larsson, and S. Karlsson. Criterion validation of surface EMG variables as fatigue indicators using peak torque: a study of repetitive maximum isokinetic knee extensions. *Journal of Electromyography and Kinesiology*, 10(4):225–232, Aug 2000.

[27] N. F. Güler and S. Koçer. Classification of EMG signals using PCA and FFT. *Journal of Medical Systems*, 29(3):241–250, June 2005.

[28] M. González-Izal, I. Rodríguez-Carreño, A. Malanda, F. Mallor-Giménez, I. Navarro-Amézqueta, E. M. Gorostiaga, and M. Izquierdo. semg wavelet-based indices predicts muscle power loss during dynamic contractions. *Journal of Electromyography and Kinesiology*, 20(6):1097–1106, 2010.

[29] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, 2007.

[30] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel method for the two-sample-problem. In *NIPS*, 2007.

[31] Y. Guo. Active instance sampling via matrix partition. In *NIPS*, 2010.

[32] Y. Guo and D. Schuurmans. Discriminative batch mode active learning. In *NIPS*, 2007.

[33] S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 39(1):333–361, 2011.

[34] X. He and D. Cai. Active subspace learning. In *ICCV*, 2009.

[35] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Batch mode active learning and its application to medical image classification. In *ICML*, 2006.

[36] S. Hoi, R. Jin, J. Zhu, and M. Lyu. Semi-supervised svm batch mode active learning for image retrieval. In *CVPR*, 2008.

[37] X. Hu, P. Yu, Q. Yu, W. Liu, and J. Qin. Classification of surface emg signal based on energy spectra change. *International Conference on BioMedical Engineering and Informatics*, 2008.

[38] H. M. Huan Liu. *Computational Methods of Feature Selection*. Chapman And Hall/CRC, 6000 Broken Sound Parkway NW, Suite 300, Boca Ratan, FL 33487-2742, 2008.

[39] J. Huang, A. J. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *NIPS*, 2007.

[40] S. Huang, R. Jin, and Z. Zhou. Active learning by querying informative and representative examples. In *NIPS*, 2010.

[41] S. W. J. Nocedal. Numerical optimization. *Springer*, 1999.

[42] J. Jiang. *A literature survey on domain adaptation of statistical classifiers*. Citeseer, 2008.

[43] F. Jing, M. Li, H. Zhang, and B. Zhang. Entropy based active learning with support vector machines for content based image retrieval. In *ICME*, 2004.

[44] A. Joshi, F. Porikli, and N. Papanikolopoulos. Multi-class active learning for image classification. In *CVPR*, 2009.

[45] J. C. L. Joshua B. Tenenbaum, Vin de Silva. A global geometric framework for nonlinear dimensionality reduction. *SCIENCE*, 2000.

[46] S. Kakade and A. Tewari. Lecture notes of CMSC 35900: Learning theory, Toyota Technological Institute at Chicago. Spring 2008.

[47] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.

[48] M. Knaflitz and P. Bonato. Time-frequency methods applied to muscle fatigue assessment during dynamic contractions. *Journal of Electromyography and Kinesiology*, 9(5):337–350, Oct 1999.

[49] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902Ű–1914, 2001.

[50] G. A. Koumantakis, F. Arnall, R. G. Cooper, and J. A. Oldham. Paraspinal muscle EMG fatigue testing with two methods in healthy volunteers. reliability in the context of clinical applications. *Clinical Biomechanics*, 16(3):263–266, Mar. 2001.

[51] D. Kumar, N. Pah, and A. Bradley. Wavelet analysis of surface electromyography. *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, 11(4):400–406, 2003.

[52] J. L. Dideriksen, D. Farina, and R. M. Enoka. Influence of fatigue on the simulated relation between the amplitude of the surface electromyogram and muscle force. *Phil. Trans. R. Soc. A*, 368:2765–2781, 2010.

[53] E. Lecuyer, H. Yoshida, N. Parthasarathy, C. Alm, and T. Babak. Global analysis of mrna localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 2011.

[54] E. leon, G. Clarke, V. Callaghan, and F. Sepulveda. A user independent real time emotion recognition system for software agents in domestic environment. *Engineering Applications of Artificial Intelligence*, 20(3):337–345, 2007.

[55] C. Liu and H. Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Trans. Image Processing*, 11:467–476, 2002.

[56] M. Lowery, C. Vaughan, P. Nolan, and M. O'Malley. Spectral compression of the electromyographic signal due to decreasing muscle fiber conduction velocity. *Rehabilitation Engineering, IEEE Transactions on*, 8(3):353–361, 2000.

[57] R. Lowry. Concepts and applications of inferential statistics. *Website for Statistical Computation*.

[58] P. Luo, F. Zhuang, H. Xiong, Y. Xiong, and Q. He. Transfer learning from multiple source domains via consensus regularization. In *CIKM*, 2008.

[59] Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation with multiple sources. In *NIPS*, 2009.

[60] P. Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculte des sciences de ToulouseSciences de Toulouse*, IX(2):245Ű–303, 2000.

[61] C. McDiarmid. *On the method of bounded differences.*, volume 5. Cambridge University Press, Cambridge, 1989.

[62] L. A. Merkle, C. S. Layne, J. J. Bloomberg, and J. J. Zhang. Using factor analysis to identify neuromuscular synergies during treadmill walking. *Journal of Neuroscience Methods*, 82(2):207–214, aug 1998.

[63] R. Merletti and P. Parker. Electromyography: Physiology, engineering, and non-invasive applications. *IEEE Press Eng in Med and Biol Soc*, 2004.

[64] S. J. Pan, J. T. Kwok, and Q. Yang. Transfer learning via dimensionality reduction. In *AAAI*, 2008.

[65] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. In *IJCAI*, 2009.

[66] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 2009.

[67] L. R. Brody, M. T. Pollock, S. H. Roy, C. J. De Luca, and B. Celli. pH-induced effects on median frequency and conduction velocity of the myoelectric signal. *Journal of Applied Physiology*, 71:1878–1885, 1991.

[68] P. Rai, A. Saha, H. Daumé III, and S. Venkatasubramanian. Domain adaptation meets active learning. In *NAACL-HLT Active Learning for NLP Workshop*, 2010.

[69] D. Rogers and D. MacIsaac. Training a multivariable myoelectric mapping function to estimate fatigue. *Journal of Electromyography and Kinesiology*, 20, 2006.

[70] M. Rosenstein, Z. Marx, and L. Kaelbling. To transfer or not to transfer. In *NIPS 2005 Workshop on Inductive Transfer: 10 Years Later*, 2005.

[71] T. Rudroff T, D. Staudenmann, and R. Enoka. Electromyographic measures of muscle activation and changes in muscle architecture of human elbow flexors during fatiguing contractions. *J Appl Physiol*, 104(6):1720–1726, 2008.

[72] P. S. Sung, U. Zurcher, and M. Kaufman. Reliability difference between spectral and entropic measures of erector spinae muscle fatigability. *Journal of Electromyography and Kinesiology*, 2009.

[73] A. D. Santos, B. Poston, M. Jesunathadas, L. R. Bobich, T. M. Hamm, and M. Santello. Influence of fatigue on hand muscle coordination and EMG-EMG coherence during Three-Digit grasping. *Journal of Neurophysiology*, 104(6):3576 –3587, Dec 2010.

[74] B. Schölkopf and A. J. Smola. *Learning with Kernels Support Vector Machines, Regularization, Optimization and Beyond*. The MIT Press, 2002.

[75] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML*, 2000.

[76] B. Settles. Active learning literature survey. In *Computer Sciences Technical Report 1648*. University of Wisconsin-Madison, 2009.

[77] H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *ACM Workshop on Computational Learning Theory*, 1992.

[78] X. Shi, W. Fan, and J. Ren. Actively transfer domain knowledge. In *ECML/PKDD*. Antwerp, Belgium, 2008.

[79] X. Shi, W. Fan, Q. Yang, and J. Ren. Relaxed transfer of different classes via spectral partition. In *KDD*, 2009.

[80] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. In *JSPI*, 2000.

[81] B. Sriperumbudur, A. Gretton, K. Fukumizu, B.Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, 2010.

[82] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *JMLR*, 2:67–93, 2001.

[83] A. Subasi and M. Kiymik. Muscle fatigue detection in emg using time-frequency methods, ica and neural networks. *J Medical Systems*, 34(4):777–785, April 2010.

[84] M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, 2008.

[85] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, and S. Shu. Systematic determination of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 3, 2002.

[86] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *JMLR*, 2:45–66, 2000.

[87] V. Torvik, E. Triantaphyllou, T. Liao, and S. Waly. Predicting muscle fatigue via electromyography: A comparative study. In *25th International Conference on Computers and Industrial Engineering*, pages 277–280, 1999.

[88] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.

[89] I. Witten and E. Frank. In *Data Mining: Practical Machine Learning Tools with Java Implementations*, San Francisco, CA, 2000. Morgan Kaufmann.

[90] N. Wu and J. Zhang. Factor-analysis based anomaly detection and clustering. *Decision Support Systems*, 42(1):375–389, Oct. 2006.

[91] H. Xie and Z. Wang. Mean frequency derived via Hilbert-Huang transform with application to fatigue EMG signal analysis. *Computer Methods and Programs in Biomedicine*, 82(2):114–120, May 2006.

[92] K. Yu, J. Bi, and V. Tresp. Active learning via transductive experimental design. In *ICML*, 2006.

[93] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang. Active learning based on locally linear reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(10):2026–2038, Oct 2011.

[94] T. Zhang and F. Oles. A probability analysis on the value of unlabeled data for classification problems. In *ICML*, 2000.

[95] E. Zhong, W. Fan, J. Peng, K. Zhang, J. Ren, D. Turaga, and O. Verscheure. Cross domain distribution adaptation via kernel mapping. In *KDD*, 2009.

APPENDIX A

THEORETICAL ANALYSIS

A: Proof of Lemma 1

Proof. Define $\Phi(S) = \sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)$. Changing the i-th point in the s-th source affects $\Phi(S)$ by at most $\gamma^s_i = \mu\beta^s\alpha^s_i$, while changing a point in the target affects $\Phi(S)$ by at most $\gamma^s_i = 1/n$ ($s = 0$). Applying McDiarmid's inequality [61] to $\Phi(S)$, the following holds with probability at least $1 - \delta/2$:

$$\Phi(S) \leq E_S[\Phi(S)] + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, using standard techniques used in [2], the expectation is bounded as follows:

$$\begin{aligned}
E_S[\Phi(S)] &= E_S\left[\sup_{h\in\mathbb{H}} E^S_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_S\left[\sup_{h\in\mathbb{H}} E_{\bar{S}}[\hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)]\right] \\
&\leq E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \hat{E}^{\bar{S}}_{\alpha,\beta}(h) - \hat{E}^S_{\alpha,\beta}(h)\right] \\
&= E_{S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&= E_{\sigma,S,\bar{S}}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\left(\mathcal{L}(h(\bar{x}^s_i), \bar{f}_s(\bar{x}^s_i)) - \mathcal{L}(h(x^s_i), f_s(x^s_i))\right)\right] \\
&\leq 2E_{\sigma,S}\left[\sup_{h\in\mathbb{H}} \sum_{s=0}^{k}\sum_{i=1}^{n_s}\sigma^s_i\gamma^s_i\mathcal{L}(h(x^s_i), f_s(x^s_i))\right] \leq 2\Re_S(G) = \Re_S(H),
\end{aligned}$$

where the last step follows from the standard techniques for relating the Rademacher complexities [46], and $\mathbb{G}$ is a class of functions given by:

$$\mathbb{G} = \{x \mapsto \mathcal{L}(h'(x), h(x)) : h, h' \in \mathbb{H}\}.$$

Thus, for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$E^S_{\alpha,\beta}(h) \leq \hat{E}^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Similarly, by defining $\Phi'(S) = \sup_{h\in\mathbb{H}} \hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)$ and bounding the expectation of $\Phi'(S)$, it can be shown that for any $h \in \mathbb{H}$, the following holds with probability at least $1 - \delta/2$:

$$\hat{E}^S_{\alpha,\beta}(h) \leq E^S_{\alpha,\beta}(h) + \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Thus, with probability at least $1 - \delta$:

$$\left|\hat{E}^S_{\alpha,\beta}(h) - E^S_{\alpha,\beta}(h)\right| \leq \Re_S(H) + \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma^s_i)^2\right)\log(2/\delta)}{2}}.$$

Next, $\Re_S(H)$ is bounded as follows [46]:

$$\begin{aligned}
\Re_S(H) &= E_{S,\sigma}\left[\sup_{h\in\mathbb{H}}|\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s h(x_i^s)|\middle| S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}}|\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s|\middle| S=(x_i^s)\right] \\
&= E_{S,\sigma}\left[\sup_{u\in\mathbb{H}_{|S}}|\sum_{s=0}^{k}\sum_{i=1}^{n_s}\gamma_i^s\sigma_i^s u_i^s|\middle| S=(x_i^s)\right] \\
&\leq E_S\left[\max_{u\in\mathbb{H}_{|S}}||u||\sqrt{2\log|\mathbb{H}_{|S}|}\right] \text{ (Massart's Lemma [60])} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}E_S\left[\sqrt{2\log|\mathbb{H}_{|S}|}\right] \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\sqrt{2\log|\prod_{\mathbb{H}}(m)|} \\
&\leq \sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}\sqrt{2d\log\frac{em}{d}},
\end{aligned}$$

where $\mathbb{H}_{|S}$ is the restriction of $\mathbb{H}$ on S, $\prod_{\mathbb{H}}(m)$ is the growth function for $\mathbb{H}$ given by the maximum number of ways m points can be classified by $\mathbb{H}$, and e is the natural number. □

B: Proof of Theorem 1

Proof. Let $h^* = \arg\min_{h\in\mathbb{H}}\{\epsilon_T(h)+\epsilon_{\alpha,\beta}(h)\}$. By the triangle inequality, one has

$$\begin{aligned}
|\epsilon_{\alpha,\beta}(h)-\epsilon_T(h)| &\leq |\epsilon_{\alpha,\beta}(h)-\epsilon_{\alpha,\beta}(h,h^*)|+|\epsilon_{\alpha,\beta}(h,h^*)-\epsilon_T(h,h^*)|+|\epsilon_T(h,h^*)-\epsilon_T(h)| \\
&\leq \epsilon_{\alpha,\beta}(h^*)+|\epsilon_{\alpha,\beta}(h,h^*)-\epsilon_T(h,h^*)|+\epsilon_T(h^*) \\
&\leq \lambda_{\alpha,\beta}+\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right).
\end{aligned}$$

Next, $(1+\mu)\epsilon_T(\hat{h})$ is bounded as follows:

$$\begin{aligned}
& (1+\mu)\epsilon_T(\hat{h}) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(\hat{h})+\epsilon_T(\hat{h})+\mu\left(\lambda_{\alpha,\beta}+\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(\hat{h})+\hat{\epsilon}_T(\hat{h})+\Re_S(H)+\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}+\mu\left(\lambda_{\alpha,\beta}+\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\hat{\epsilon}_{\alpha,\beta}(h_T^*)+\hat{\epsilon}_T(h_T^*)+\Re_S(H)+\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}+\mu\left(\lambda_{\alpha,\beta}+\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & \mu\epsilon_{\alpha,\beta}(h_T^*)+\epsilon_T(h_T^*)+2\Re_S(H)+2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}+\mu\left(\lambda_{\alpha,\beta}+\frac{1}{2}d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right) \\
\leq\ & (\mu+1)\epsilon_T(h_T^*)+2\Re_S(H)+2\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}}+\mu\left(2\lambda_{\alpha,\beta}+d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta},\mathbb{D}_T\right)\right)
\end{aligned}$$

Thus,

$$\epsilon_T(\hat{h}) \leq \epsilon_T(h_T^*) + \frac{2\Re_S(H)}{1+\mu} + \frac{2}{1+\mu}\sqrt{\frac{\left(\sum_{s=0}^{k}\sum_{i=1}^{n_s}(\gamma_i^s)^2\right)\log(2/\delta)}{2}} + \frac{\mu}{1+\mu}\left(2\lambda_{\alpha,\beta} + d_{\mathbb{H}\Delta\mathbb{H}}\left(\mathbb{D}_{\alpha,\beta}, \mathbb{D}_T\right)\right) \tag{1.1}$$

□

Note that our proof follows a similar procedure in [4]. The main differences include (1) this thesis employs the weighted Rademacher complexity, which provides a tighter bound than the one in [4] based on the VC dimension; (2) the empirical minimizer $\hat{h}$ of our joint error function includes two terms involving both source and target domain data with a differential weight μ, while the one in [4] involves one term only. For the special case when $\mu = 1$ and α_i^s's are given a uniform weight, i.e., $\alpha_i^s = 1/n_s$, our bound in (3.17) is strictly tighter than the one in [4] (due to the $1/2$ factor in the last term). In the general case with different choices of μ and α_i^s's, our bound can be further improved.

C: More details on the datasets and parameters used for the implementation of different methods

The statistics of the test datasets used is summarized in Table 1.1.

Dataset	Number of domains	Dimension	Number of classes
20 Newsgroups	13	100	2
Sentiment Analysis	4	200	2
Surface Electromyogram (SEMG)	8	12	4

Table 1.1: Statistics of the test datasets

The categories of 20 Newsgroups dataset that were used in the experiments as source and target domains are as listed in Table 1.2.

20 Newsgroups	
Categories	inst
comp.os.ms-windows.misc	100
comp.sys.ibm.pc.hardware	100
comp.sys.mac.hardware	98
comp.windows.x	100
rec.motorcycles	100
rec.sport.baseball	100
rec.sport.hockey	100
sci.electronics	100
sci.med	100
sci.space	100
talk.politics.miseast	94
talk.politics.misc	78
talk.religion.misc	64

Table 1.2: Summary of categories (domains)

A Gaussian kernel with $\sigma = 10$ was used to compute the α values for each source. The weighted hypothesis for each source was learned using Support Vector Machines implemented in the LibSVM package, with a linear kernel and a regularization penalty $C = 10$. The β weights were computed based on a binary similarity matrix, i.e., $W_{ij} = 0$ if the i-th data point is among the N nearest neighbors of the j-th data point or the j-th data point is among the N nearest neighbors of the i-th data point; this work sets $N = 10$. TCA is implemented with a linear kernel and KMM with

a Gaussian kernel as they gave the best results. All parameters were tuned using 10-fold cross-validation.

D: Additional empirical results

Figure 1.1 shows the α-weighted data samples in both source domain D1 and source domain D2 of the toy data shown in Figure 3.4. It is observed that data samples having similar marginal probabilities in both the domains get higher weight, shown by the size of the points. The size of the points are proportional to their weights. One can observe that since at this stage the source data is re-weighted based only on marginal probability distribution difference, hence some of the data samples from source domain D1 having conflicting conditional probabilities with target domain data also get higher weight as they share similar marginal probability distributions.

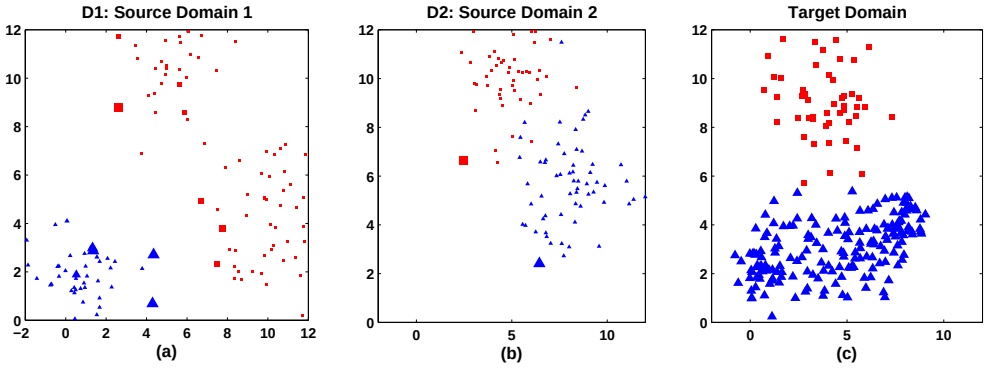


Figure 1.1: Data samples in source domains D1 and D2 re-weighted by α_i^s. Results show that points from source domain D1 also get large weights due to the similarity in marginal probabilities (the size of a point is proportional to its weight).

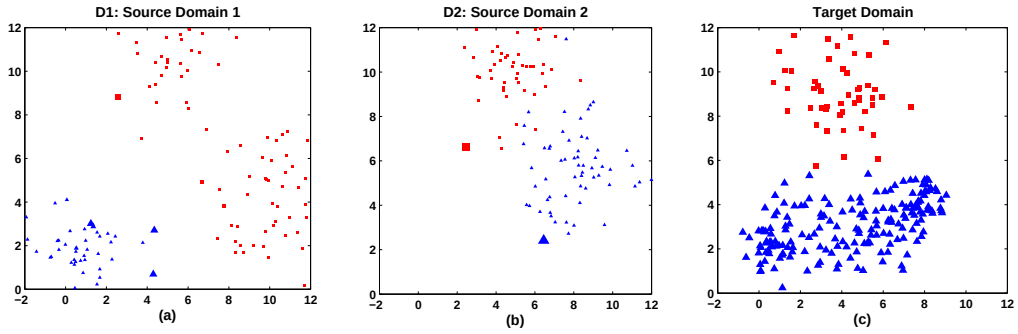


Figure 1.2: Data samples in the source domains D1 and D2 re-weighted by both α_i^s and β^s. One can observe that the points with conflicting conditional probabilities get moderated by β^s (the size of a point is proportional to its weight).

Figure 1.2 shows the results of applying β-weights to the data samples in both source domain D1 and source domain D2 of the toy data. The results show that the data samples in source domain D1 with conflicting conditional probabilities get reduced when moderated with β weights, as source domain D2 is more similar to target data in conditional probability distribution than the source domain D1.

Figure 1.3 shows the performance of 2SW-MDA on toy dataset shown in Figure 3.4 with varying μ. The result is consistent with the theoretical result established in this dissertation.

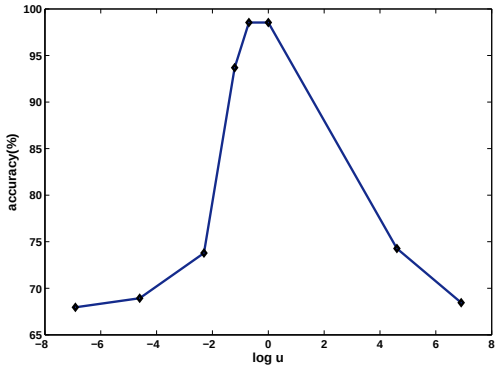


Figure 1.3: Performance of proposed 2SW-MDA method on the toy dataset shown in Figure 3.4 with varying μ - Accuracy (%).

Figure 1.4 shows the results of applying the proposed 2SW-MDA method on another set of toy dataset consisting of two source domains and a target domain with different marginal and conditional probability differences. One can observe that the distribution D1 which has conflicting conditional probabilities with target domain data gets under-weighted by the proposed weighting scheme and hence transfer happens mostly from the source distribution D2, which shares similar marginal and conditional probability differences with the target domain. β value of 0.17 for D1 and 0.83 for D2 are obtained.

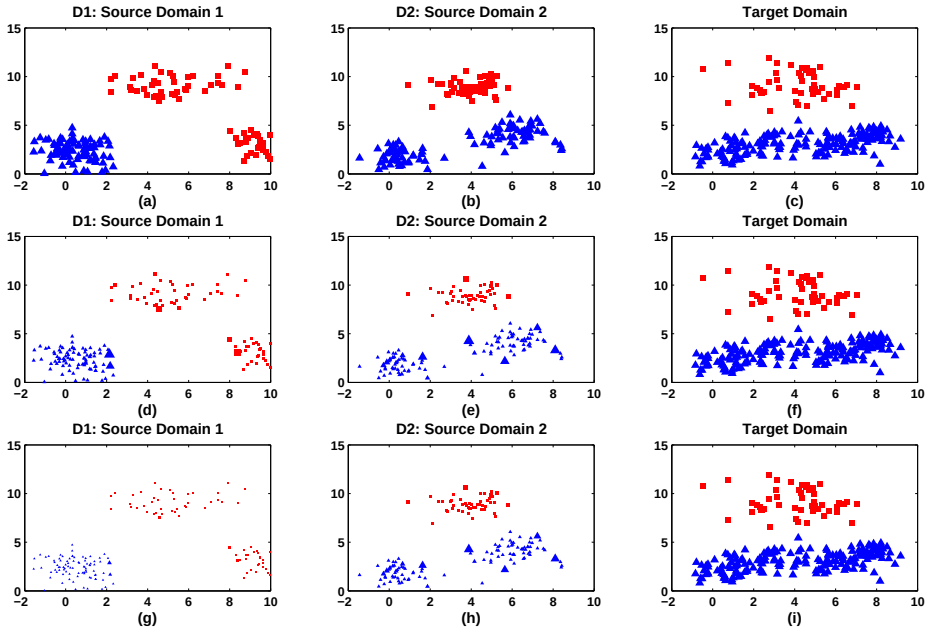


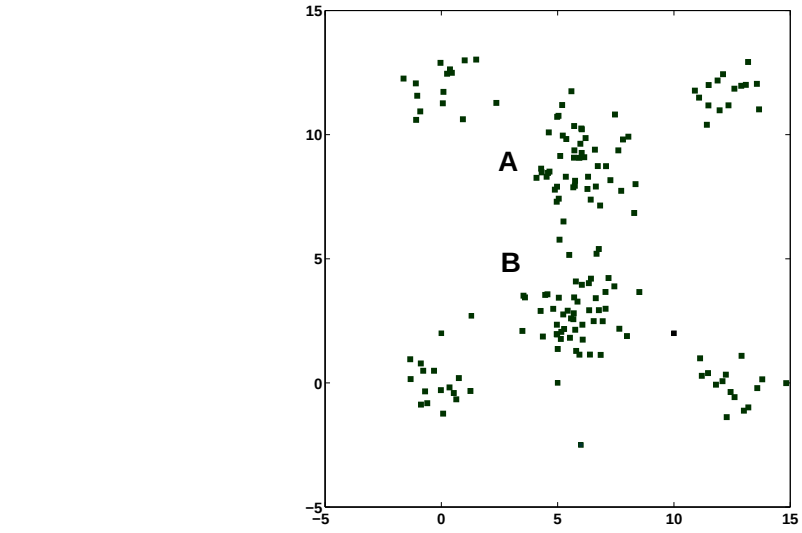
Figure 1.4: Results on another toy dataset: First row shows the original distribution of two source domains D1 and D2 and a target domain. The second and third rows show the results of applying α and β weights, respectively. The results show that source domain data samples with similar marginal and conditional probabilities get higher weight. The β values for D1 and D2 are 0.17 and 0.83 respectively, individual accuracies being 61.65% and 89.51% and proposed method gives 98.51%.

APPENDIX B

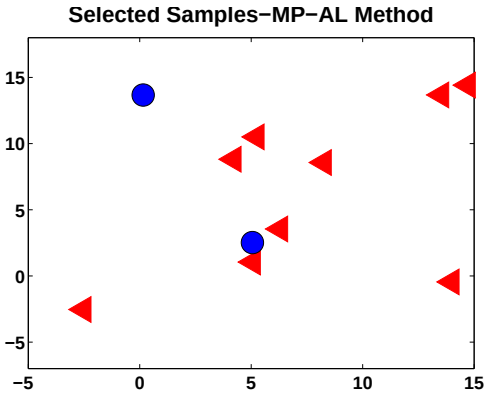
MORE RESULTS ON MARGINAL PROBABILITY BASED BATCH-MODE ACTIVE LEARNING

A: More results on *MP-AL* method.

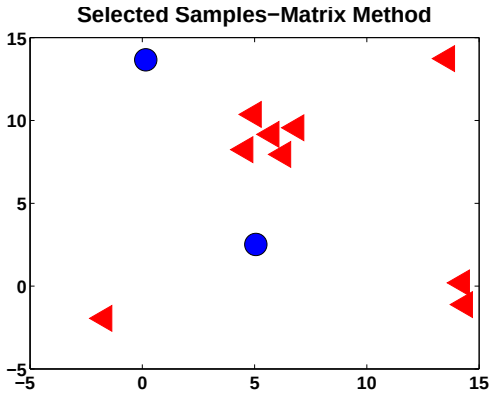
Figure 1.5 presents the nine query samples selected by *MP-AL*, *Matrix*, *Disc* and *Fisher* methods (red triangles) with a batch size of 3 in first 3 iterations from the unlabeled data shown in Figure 1.5(a). Blue circles are initially available randomly selected labeled data. One observes that *Matrix* similar to *MP-AL*, select samples from all dense regions, however it does not preserve the distribution of the unlabeled data, where as *Disc* and *Fisher* methods does not necessarily select samples from all dense regions. Hence the query samples selected by these methods, do not necessarily represent the distribution of the unlabeled data.



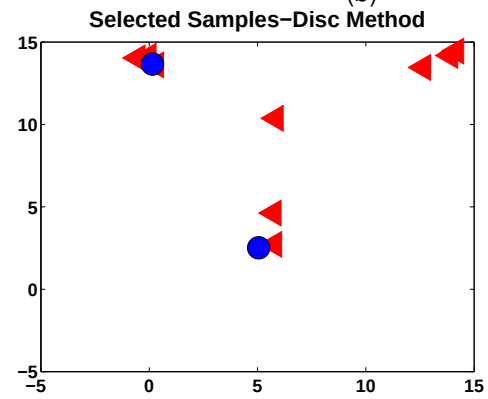
(a)



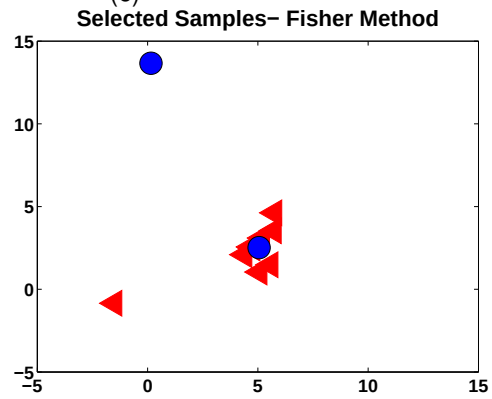
(b)



(c)



(d)



(e)

Figure 1.5: Query Samples Selected by Different Methods (red triangles) with a batch size of 3 in first 3 iterations from the unlabeled data shown in (a). Blue circles are initially available randomly selected labeled data.

APPENDIX C

MORE RESULTS ON JOINT TRANSFER AND BATCH-MODE ACTIVE LEARNING

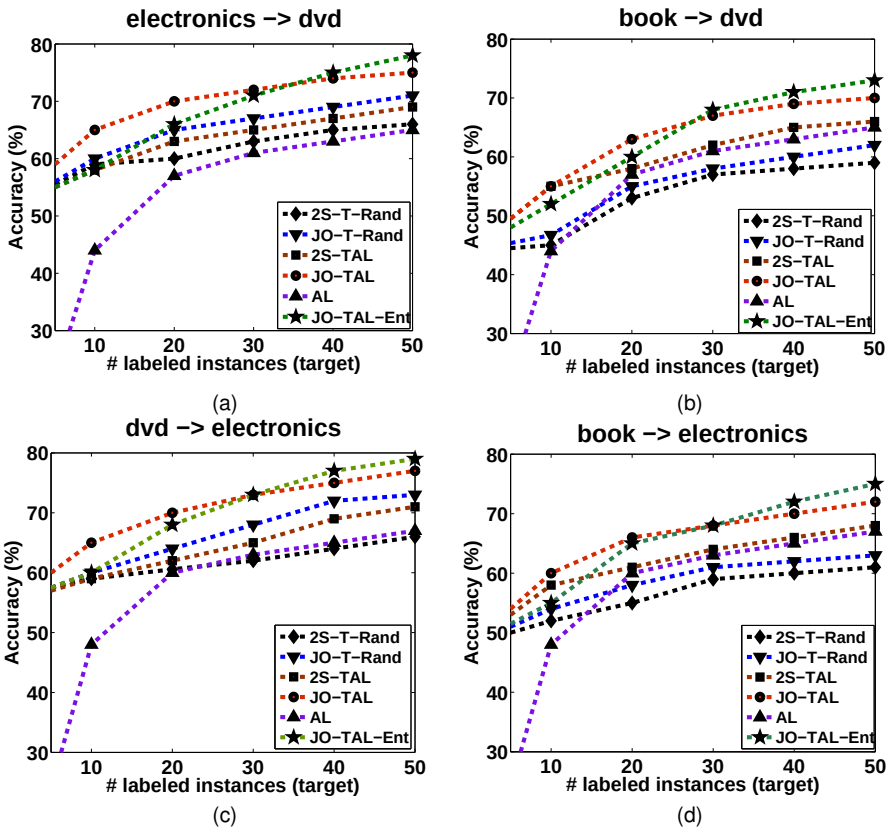


Figure 1.6: Comparative performance on Sentiment Analysis data set.

A: More results on 20 Newsgroups dataset

The competing methods and the experimental procedure remains the same as described in Section 6.3. Figure 1.7 shows the comparative performance of *JO-TAL* on all three sets of 20 Newsgroups dataset. The results show that *JO-TAL* performs better than *2S-TAL* as in the case of biomedical images. Furthermore, for Scientific and Hardware categories *JO-T-Rand* performs comparable to *2S-TAL*. This shows that performing transfer learning taking into account the randomly selected samples from target domain may be more effective than performing transfer and active learning in two stages, which may possibly cause overlap of information between the selected samples from target and source domain as illustrated in Figure 7.2. One can also observe that improvement in classification accuracies due to incorporation of transfer learning is more for Sports and moderate for Scientific and Hardware. This can be attributed to the extent of difference in distribution between the source and target domains in each of these test cases; one way to measure the distribution difference is to compute the MMD value between the source and target domains. The MMD value is 0.0121, 0.0237 and 0.0239 for Sports, Hardware and Scientific categories respectively. This is consistent with our observation in Figure 1.7.

Figure 1.8 shows the change in MMD values between the re-weighted source, selected target domain data and unlabeled target data at every iteration. One observes that as in the cases of biomedical and synthetic data, the MMD value monotonically decreases for all three categories.

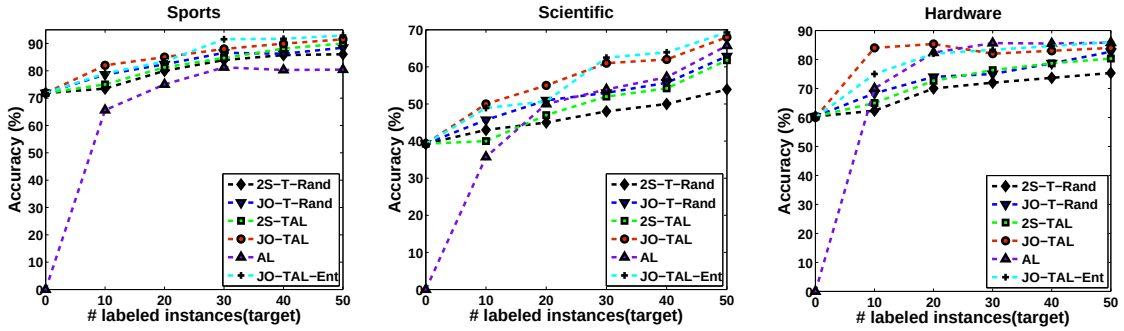


Figure 1.7: Comparative performance on 20 Newsgroups dataset.

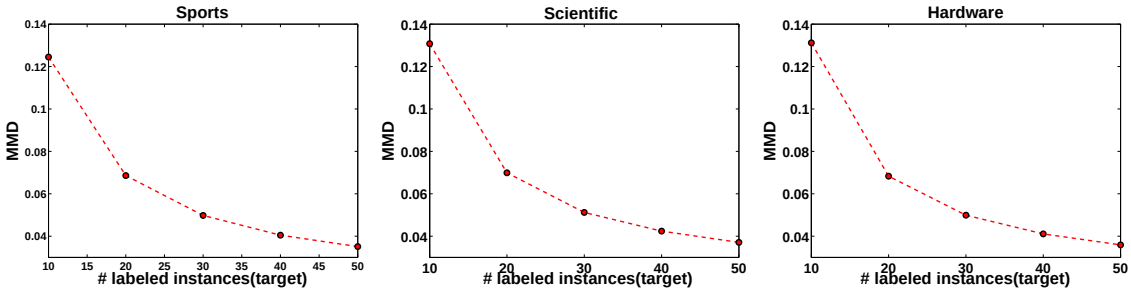


Figure 1.8: Change in MMD with increasing number of labeled data.

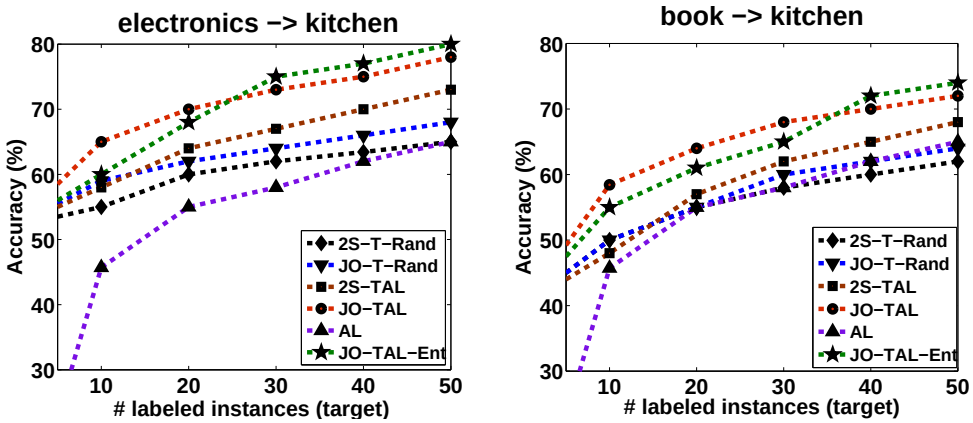


Figure 1.9: Comparative performance on Sentiment Analysis data set.

B: More results on Sentiment Analysis data set

Figures 1.9 (a) and 1.9 (b) show the comparative performance of *JO-TAL* on Sentiment Analysis data set. The first and second names in the title of the figures refer to the source and target domains respectively. The results show that *JO-TAL* performed better than *2S-TAL* by 7% and 5% for the cases with electronics and book data sets as source domains, while documents belonging to the category of kitchen forming the target domain, respectively. Please note that similar to 20

Newsgroups data set, the performance of *JO-TAL-Ent* improved towards later iterations. One can also observe that incorporation of transfer learning has improved the classification accuracy on kitchen data set by 13% and 9% with electronics and book as source data sets respectively. This can be explained by the differences in their MMD values, as in the case of 20 Newsgroups data sets, which are, 0.0145 and 0.0349 for electronics vs. kitchen and book vs. kitchen data sets respectively.

Figure 1.6 shows the comparative performance of *JO-TAL* on Sentiment Analysis data set. The first and second names in the title of the figures refer to the source and target domain respectively. Figures 1.6 (a) and 1.6 (b) show the results with electronics and book data sets as source domains, while documents belonging to the category of dvd forming the target domain, respectively. Results show that for both cases *JO-TAL* and *JO-TAL-Ent* performed better than *2S-TAL* by 7% to 10% [1]. One can observe that incorporation of transfer learning has improved the classification accuracies on dvd data set by 13% and 8% with electronics and book as source domain data, respectively. This can be explained by the MMD values, which are 0.0321 and 0.0290 for book vs. dvd and electronics vs. dvd data sets respectively. Lower MMD value between electronics and dvd, signify more relatedness in the data distribution than in the case of book vs. dvd.

One observes very similar phenomenon in Figures 1.6 (c) and 1.6 (d), with electronics as target and book and dvd being the source domains respectively. In both the cases, *JO-TAL* and *JO-TAL-Ent* performed better than *2S-TAL* by 8% to 10%. Besides, *JO-T-Rand* performed better than *2S-TAL* by 3% with dvd as source domain data. The results show that incorporation of transfer learning has improved the classification accuracy by 18% and 6% with dvd and book as source domain data respectively. This again is consistent with the distribution differences measured by their respective MMD values, which are 0.0329 and 0.0290 for book vs. electronics and dvd vs. electronics data set respectively.

Similar results have been obtained with other six combinations of the Sentiment Analysis data sets, such as electronics vs. book, kitchen vs. book, kitchen vs. dvd, kitchen vs. electronics, dvd vs. kitchen and dvd vs. book.

[1] All differences in accuracies are measured at number of labeled instances from target = 50.