An Empirical Evaluation of Social Influence Metrics

by

Nikhil Nanda Kumar

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved June 2016 by the
Graduate Supervisory Committee:

Paulo Shakarian, Chair
Arunabha Sen
Hasan Davulcu

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

Predicting when an individual will adopt a new behavior is an important problem in application domains such as marketing and public health. This thesis examines the performance of a wide variety of social network based measurements proposed in the literature - which have not been previously compared directly. This research studies the probability of an individual becoming influenced based on measurements derived from neighborhood (i.e. number of influencers, personal network exposure), structural diversity, locality, temporal measures, cascade measures, and metadata. It also examines the ability to predict influence based on choice of the classifier and how the ratio of positive to negative samples in both training and testing affect prediction results - further enabling practical use of these concepts for social influence applications.

To my family, friends and colleagues.

ACKNOWLEDGMENTS

I sincerely thank my advisor Dr. Paulo Shakarian for his continued guidance, support and encouragement during my masters and while writing this thesis. I am extremely thankful for his immense patience in evaluating my work and motivating me to strive for excellence. I believe that the knowledge, skills and expertise I learned as part of CySIS lab will significantly enable all my future endeavours.

I would like to thank Dr. Arunabha Sen and Dr. Hasan Davulcu for being on my committee and for their support to my research and dissertation. Their exceptional cooperation and response facilitated a smooth conduct of my dissertation process.

My big thanks to my collaborators Ashkan Aleali and Ruocheng Guo for the contributions made and the help and support provided. I would also thank all the other CySIS lab members. It has been a great experience working and learning together with all of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

INTRODUCTION

Predicting when an individual will adopt a new behavior is an important problem in application domains such as marketing [25], the spread of innovation [24], countering extremism [1], and public health [5]. As a result, a variety of social network based measurements have been proposed in the literature and shown to predict how likely an individual will adopt a new behavior given information about his immediate social ties. However, when such measures are proposed, they are often evaluated under different conditions - making it difficult to understand which of these measurements should be used in a real-world application. Further complicating the issue is that the choice of classification algorithm and the effect of class imbalance in both training and testing are often not explored in most research.

In this thesis, we study measurements based on neighborhood (i.e. number of influencers [5], personal network exposure [24]), structural diversity [23], locality [29], temporal measures [11], cascade measures [12], and metadata [15]. We examine the probability of an individual becoming influenced based on these measurements (probability of adoption). We also examine the the ability to predict influence based on choice of classifier and the how the ratio of positive to negative samples in both training and testing affect prediction results. Specifically, this research make the following contributions.

1. We review a variety of measurements used to predict social influence and we group them in six categories (Chapter 3).

2. We evaluate how these measurements relate to the probability of a user being

influenced using real-world microblog data (Chapter 4).

3. We evaluate how these measurements perform when used as features in a machine learning approach and compare performance across a variety of supervised machine learning approaches (Chapter 5).

4. We evaluate how the ratio of positive to negative samples in both training and testing affect predictive results (Chapter 5.4).

We note that contribution 4 is of particular importance, as (particularly with microblog data) users are exposed to large number of messages that they do not retweet (negative samples). Hence, in both training and testing, researchers can increase the negative samples utilized by large amounts - hence arbitrarily determining the level of class imbalance. As with this study as a whole, the experiments on data imbalance were to better understand these previous research results in tests that better mimicked real-world scenarios.

1.1 Related Work.

Beyond the work that we shall describe concerning the various measures for social influence we investigate in Chapter 4, there has been some general work in the area of social influence that have taken approaches not necessarily amenable to comparison. For instance, the seminal work of Kempe et al. [16] describe two popular models for information cascades which spawned several techniques to learn the parameters (which also correspond to edge weights in the graph). For example, Saito et al. [20] assigned such probabilities based on an expectation-maximization appproach while Goyal et al. [11] leveraged a variety of simple models based on ideas such as a empirically-learned probabilities and similarity measurements. See [21] for a review of some of this work. There has also been related work on predicting cascades [8, 12, 27]

which are more focused on determining if a trend in social media exceeds a certain size. That said, some of the ideas from these approaches, such as structural diversity [23] are examined here (though this thesis is focused on a different problem). Other work such as Myers et al. [19] studied the external factors influencing information diffusion, Liu et al. [18] and Tang et al. [22] focused their studies on topic influence. Jenders et al. [15] studied a combination of different features including some of the metadata features like mentions and hashtags, along with latent features like sentiments and emotional divergence for predicting the virality of a tweet - many of which we examine in this study as well. Hong et al. [14] have also considered a wide spectrum of features including structural, content and temporal information. However, their study focused more on content-based features and not the structural features considered here - many of which were introduced after that work.

Chapter 2

TECHNICAL PRELIMINARIES

Here we introduce the necessary notation and describe our social network data. We represent a social network as a graph $G = (V, E)$ where V is the set of vertices and E is the set of directed edges that have sizes $|V|, |E|$ respectively. The intuition behind edge (v, v') is that node v can influence v'. This intuition stems from how we create the edges in our network: (v, v') is an edge if during a specified time period there is at least one microblog posted by v that is reposted by v' . For node $v \in V$, the set of in-neighbors is denoted as η_v^{in} , and the set of out-neighbors as η_v^{out} . We use d_v^{in} and d_v^{out} to denote the in-degree and out-degree respectively. We also assume a partition over nodes that specifies a community structure. We assume that such a partition is static (based on the same time period from which the edges were derived) and the function $P(V) : V \to \mathcal{C}$ maps the set of nodes (V) to the set of communities $(\mathcal{C})$, where $\mathcal{C}$ consists of k communities: $\{C_1, C_2, ..., C_k\}$. We utilize the Louvain algorithm [3] to identify our communities in this thesis due to its ability to scale.

Cascades. For a given microblog θ, we define t as the number of time units from the initial post of θ before the microblog was reposted by one of v's incoming neighbors - intuitively the time at which v was exposed to θ. We denote the subset of nodes who originally posted or reposted θ for time period t as V_θ^t. Likewise, the set of reposting relationships within the same time period will be denoted by R_θ^t. Taken together, we have a *cascade*: $D_\theta^t = (V_\theta^t, R_\theta^t)$. Any valid original microblog θ could be treated as a unique identifier for a cascade. Given a microblog θ, v_θ is the originator at instance t_θ^0, which is defined as the origin time when the originator posted the

microblog θ. We denote the size of a cascade at any particular time t as $|V_\theta^t|$. For $v \in V_\theta^t$, the set of all *active* neighbors with respect to θ is defined as $S_\theta^v = V_\theta^t \cap \eta_v^{in}$. We also define the distance $d_\theta^t(v, u)$ as the shortest path length between v and u in D_θ^t.

2.1 Sina Weibo Dataset.

The dataset we used was provided by the WISE 2012 Challenge [1] . It included a sample of microblogs posted on Sina Weibo from 2009 to 2012. In this dataset, we are provided with time and user information for each post and the last repost in a chain which enabled us to derive a corpus of cascades. We create the social network G from the retweeting relationships of microblogs published between May 1, 2011 and July 31, 2011. We use the microblogs published in August 2011 to train and test our approach. Table 2.1 lists the statistics of the dataset we used.

#Users	#Edges	#Reposted tweets	#Reposted Users
5,910,608	52,472,547	2,238,659	394,441

Table 2.1: Graph Statistics

We found that the network derived from the dataset had 7,668,693 users with 55,381,104 edges between them. For this network, the number of active users in August (the time period used to study social influence) is 5,910,608 while 5,664,625 of them have at least have one out-neighbor. During the month of August, there were 22,182,703 retweet chains. From this data, we removed the users who are not present in V; we also removed 2,660,421 empty repost chains caused by this elimination. The dataset does not contain the repost time for the nodes in the middle of chains. We estimated this time for each node in the chain based on the original post time and

[1] http://www.wise2012.cs.ucy.ac.cy/challenge.html

the final repost time. Table 2.1 lists the statistics of this dataset during the period of study.

Among all the retweeted users we further extract the top retweeters defined as those who had at least 100 retweets during the period. This set of high frequency tweeters will be used as a base for deriving the sample set for our experiments. For each user in the above mentioned group, an occurrence of them retweeting a post when they have an active in-neighbor is considered as a positive instance. If any of their followees have tweeted and they haven't retweeted, it is considered as a negative instance.

Chapter 3

MEASUREMENTS TO PREDICT SOCIAL INFLUENCE

In this chapter, we categorize several approaches for predicting social influence.

1. Neighborhood-based measures
2. Structural diversity measures
3. Influence locality
4. Cascade-based measures
5. Temporal measures
6. Metadata

We examine each of these categories in turn.

3.1 Neighborhood-Based Measures.

These are the measures computed using each node and its immediate neighbors. These measures represents the pair wise influence that the neighboring nodes exert on a given node. Retweeting from followees is the primary mode of tweet visibility in a microblogging site, as usually a tweet is visible to a user from its followee subgraph. Specifcially, we study the following

- **Number of active neighbors.** ($|S_v^\theta|$) This represents the count of active neighbors for a node v. In Damon Centola's notable empirical study [5], he noted that additional "social signals" – or active neighbors – significantly increased the likelihood of an individual adopting a new behavior.

- **Personal Network Exposuure (PNE).** ($|S_v^\theta|/d_v^{in}$) Is a measure adopted from the social science community (i.e. see [24]) and has obtained recent interest (i.e. [13]). As per [24], PNE quantifies the extent to which a person is exposed to direct and indirect influence. This value is defined as the ratio of number of active neighbors to total number of neighbors. It is a measure of the fraction of influence an active neighbor u has on v. If v has many in-neighbors aka followees, then u's influence is diluted and PNE represents that dilution.

- **Average in-neighbor count of active neighbors.** ($|\Sigma_{u \in S_v^\theta} d_u^{in}|/|S_v^\theta|$) This is calculated by averaging the number of in-neighbors of each active neighbor of a node. This defines the dilution of the influence path and is similar to the measure, *number of uninfected neighbors* as described in [27]. Other releated studies include Cha et al. [6], where they studied the effect of a social network user's indegree in depth, and observed that high indegree is not necessarily correlated to influence in terms of spawning retweets.

3.2 Structural Diversity Measures.

This group of measurements take into account the structural diversity in the local neighborhood of the node - which refers to the communities present in the neighborhood.

Ugander et al. [23] introduced structural diversity where they studied the effect of number of connected components of a friendship network. Fortunato et al [9] defined communities as the set of graph vertices which are organized into groups that seem to live fairly independently of the rest of the graph. Weng et al. [28] used the community structure to predict the increase in cascade size. We use the modularity maximisation method [7] for detecting communities in our dataset. The Louvian Algorithm [3] which comes under this method is used to derive the communities in

this study due to its ability to scale. We use two community based measures.

- **Active community count.** ($|P(S_v^\theta)|$) This is defined as the number of adjacent communities of a given user v with at least one active neighbor of v. The communities that include active neighbors are more significant in this context than rest of the adjacent communities. Shakarian et al. have studied this measure in their book [21] highlighting the importance of structural diversity.

- **Active community ratio** ($|P(S_v^\theta)|/|P(\eta_v^{in})|$) It is calculated as the ratio of the active community count to the total number of adjacent communities. This is similar to the personal network exposure [24] and represents the dilution of the effect of active community count with respect to other neighboring communities.

3.3 Influence Locality.

We examine the Influence Locality model known as LRC-Q, introduced by Zhang et al. [29]. LRC-Q is defined by the influence locality function Q which is a combination of peer influence factor (g) and structural factor (f). Peer influence factor is obtained as a linear combination of the geometric mean of random walk probabilities of active neighbors and structural factor as a linear combination of the number of circles formed by the active neighbors in the ego network of the user v. These are defined in their paper by the following equations.

$$Q = w \times g + (1 - w) \times f \tag{3.1}$$

$$g = \sqrt[|S_v^\theta|]{\prod_{v_i \in S_v^\theta} (t_\theta^v - t_\theta^{v_i}) \times p_{v_i}} \tag{3.2}$$

$$f = a \log(|S_v^\theta| + 1) + be^{-\mu |C(S_v^\theta)|} \tag{3.3}$$

In the above equations, p_{v_i} is the random walk probability from the active user v_i to the given user v, $C(S_v)$ is the collection of circles formed by the active neighbors, t_θ^v

is the time at which v posted or reposted the microblog θ, μ is the decay factor and, a, b and w are balance parameters. For our experiments we set the value of μ as 1 and, a, b and w to be 0.5, as per the parameter settings of [29].

3.4 Cascade-Based Measures.

This group of measurements take into account the various parameters that are part of a microblog cascade. There has been many studies in the area of predicting the cascades including Bakshy et al. [2] , Cheng et al. [8] and more recently Guo et al. [12]. Unlike our study, there hasn't been many attempts to utilize the cascade parameters in predicting retweet behavior. We study the following measures.

- **Cascade size.** ($|V_\theta^t|$) Cascade size is computed as the count of people who have retweeted a particular microblog θ at time t. This number is usually visible to the microblog user and can have an impact on their retweet behavior.

- **Path length.** ($d_\theta^t(v, v_\theta)$) Path length is the length of a tweet trace path from the original tweeter to a given user in the cascade. Watts et al. [26] were the first to study the path length where they found that many social and technological networks have small path lengths. Kwak et al. [17] studied the path length in twitter, and Weng et al. [28] studied a distance measure called Average step distance which was based on the path length. Our study focuses on the path length with respect to a particular cascade D_θ^t.

3.5 Temporal Measure

Temporal measures were given prominence in many of the prior studies either by itself, or as a factor in combination with other measures. Goyal et al. [11] utilized

the temporal factor and attempted to predict the time by which an influenced user will perform an action. Hong et al. [14] studied a variety of temporal measures and observed that they have a stronger effect on messages with low and medium volume of retweets, compared to highly popular messages. We study the following temporal measure.

- **Retweet Time delay.** (t) This is defined as the time delay between the original tweet and the time when v is exposed to microblog θ. The time at which a tweet was made is another piece of information which people are exposed to while viewing a tweet. This can affect their decision to retweet it or not. This is one of the temporal measures studied by Hong et al. [14].

3.6 Metadata.

These are simple measures derived from the metadata associated with the tweets. We consider the presence or absence of links, mentions and hashtags as measures for our study. Jenders et al. [15] did an extensive analysis of a wide range of tweet and user features regarding their influence on the spread of tweets. They considered the number of mentions and number of hashtags among the obvious tweet features. They observed that tweets containing both hashtags and mentions are more likely to be retweeted than those with out, however as the number of hashtags/mentions in a tweet grows, the expected number of retweets decreases. In this study we only considers their presence or absence as a measure and doesn't go into any deeper analysis.

- **Presence of a link (hasLink).** This is a binary value which represents whether the original tweet had a link. Links are usually shown as part of the

tweet content. Links in tweets is measure similar to mentions and hashtags, but hasn't been studied as extensively as either in the context of social influence.

- **Presence of a mention (hasMention).** A binary value which represents whether the original tweet had a mention. Intuitively, a user might be more willing to retweet if there is a mention of them or someone they know. Similar to [15], Cha et al. [6] analysed the effect of the number of mentions and found that mentions can be an important measure of an individual influence in the social network.

- **Presence of a hashtag (hasHashtag).** A binary value which represents whether the original tweet had hashtags. Hashtags are also a means by which tweets become visible to users and thus has a lot of significance in this regard. A deeper analysis of it's significance like in [15], is beyond the scope of this work and we only focus on how the presence or absence of a hashtag affects the retweeting behavior.

Chapter 4

SOCIAL INFLUENCE MEASUREMENT STUDY

Here, we examine the distribution of the various measurements which were defined in the last chapter. For each of those measures, the values are put into intervals of equal sizes and the fraction of positive samples in the interval is plotted as the probability. The horizontal axis shows the value intervals of the measure, while the vertical one shows the number of occurrences for positive instances with respect to the total in that particular interval. The error bar shows twice the standard deviation of the sample. A detailed analysis of their distribution is given below.

Neighborhood-based measures. Active neighbor count intuitively has a positive correlation with the influence as shown in Figure 4.1. Figure 4.2 shows the active neighbor count for the lower values which also shows similar correlation. This is consistent with the empirical study of [5]. As the number of retweeters among in-neighbors increases, the probability of a person retweeting the particular tweet increases. Figure 4.3 shows that PNE also exhibits positive correlation like active neighbor count. This shows the significance of PNE measure as demonstrated by other studies such as [24] and [13]. Average in-neighbor count of Active Neighbors doesn't show a clear correlation in it's distribution as seen in Figure 4.4.

Structural diversity measures. Number of active communities (Figure 4.5) shows a good positive correlation with the retweet behavior. This result is consistent with the related studies such as [28] and [8]. Active community ratio (Figure 4.6) also demonstrates a reasonable correlation with the positive instances as this measure

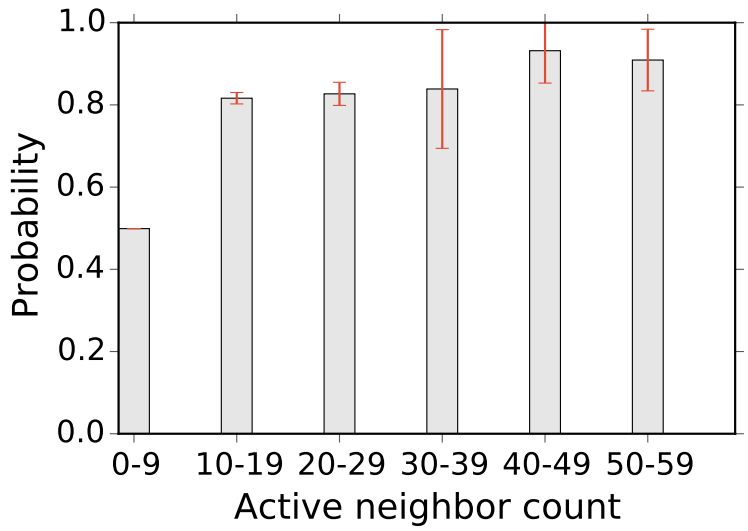


Figure 4.1: Active Neighbor Count. Error Bars Represent Two Standard Deviations.

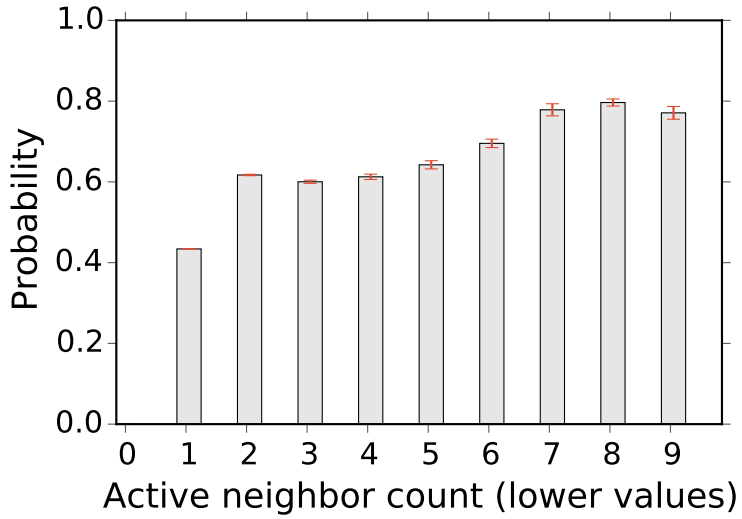


Figure 4.2: Active Neighbor Count (Lower Values). Error Bars Represent Two Standard Deviations.

represents the dilution of community influence based on the total number of adjacent communities.

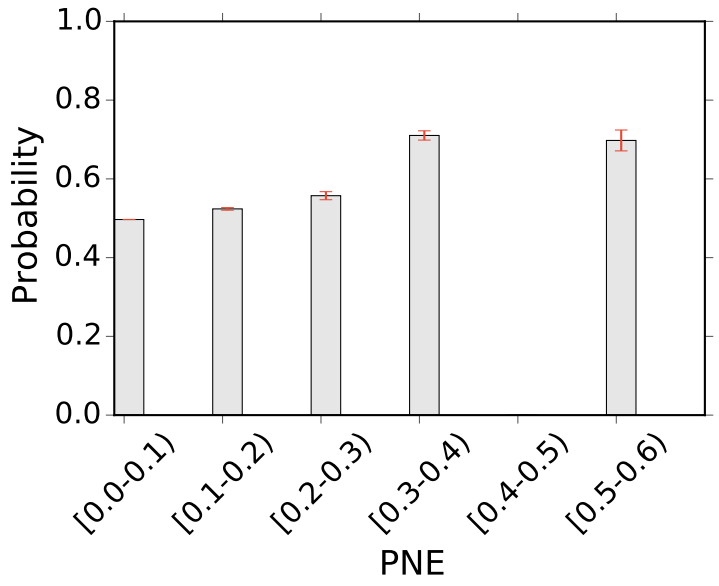


Figure 4.3: PNE. Error Bars Represent Two Standard Deviations.

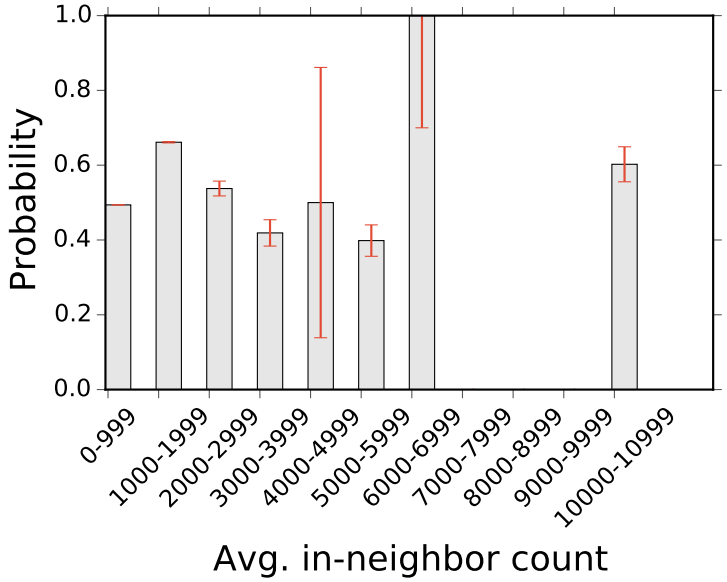


Figure 4.4: Average In-Neighbor Count of Active Neighbors. Error Bars Represent Two Standard Deviations.

Cascade-based measures. Intuitively, cascade size is an important influencer in retweet behavior. If a tweet is reasonably popular it tends to attract further retweets.

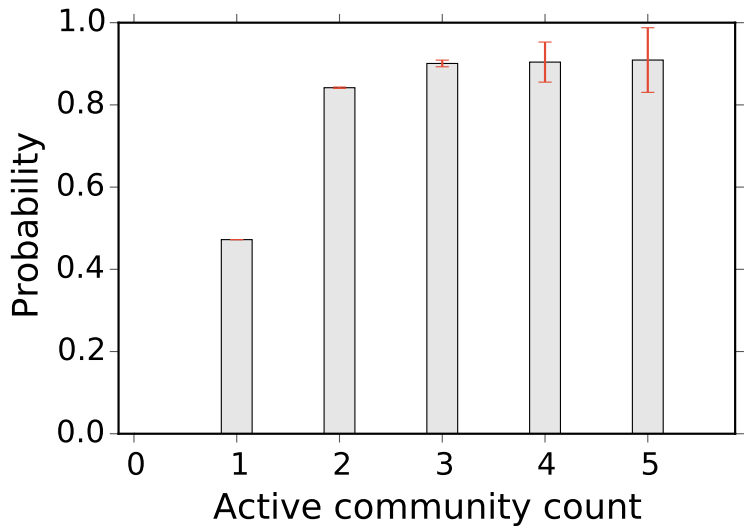


Figure 4.5: Number of Active Communities. Error Bars Represent Two Standard Deviations.

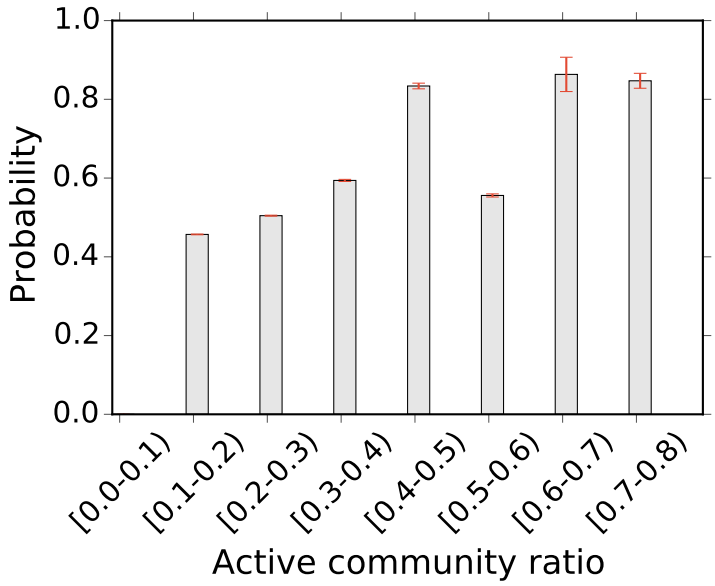


Figure 4.6: Active Community Ratio. Error Bars Represent Two Standard Deviations.

The same is revealed from the distribution in Figure 4.7. This is consistent with the research of [2] and [8] although they studied a different problem. The intuition for

path length is that, as the distance from the original tweeter increases a user is less interested in retweeting the tweet. Our results show (Figure 4.8) that this intuition holds between path length 1 and 2. But, for the remaining intervals, results doesn't correlate well. This can be explained by comparing to the results of [15] where they found similiar pattern while analyzing mentions and hashtags. Further, the results of [8] indicate that information cascade depth is related to popularity. Hence, the microblogs that are far from the original poster may be inherently popular as the information cascade has proceeded to a larger depth.

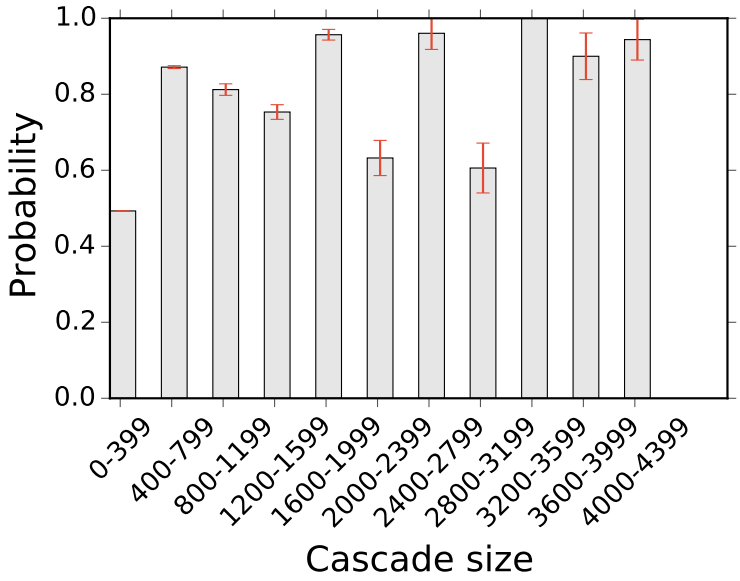


Figure 4.7: Cascade Size. Error Bars Represent Two Standard Deviations.

Temporal. Figure 4.9 shows that retweet time delay have slight inverse correlation with the influence. Intuitively, the influence of a tweet decays with time, and as people are exposed to date/time information in the social network they are less likely to retweet old tweets. This decay factor has been used in works like [11], [29] etc. and above result shows the same.

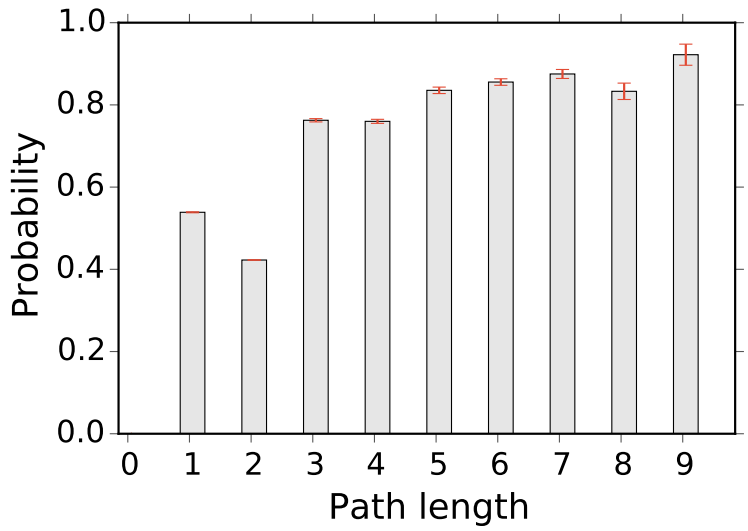


Figure 4.8: Path Length. Error Bars Represent Two Standard Deviations.

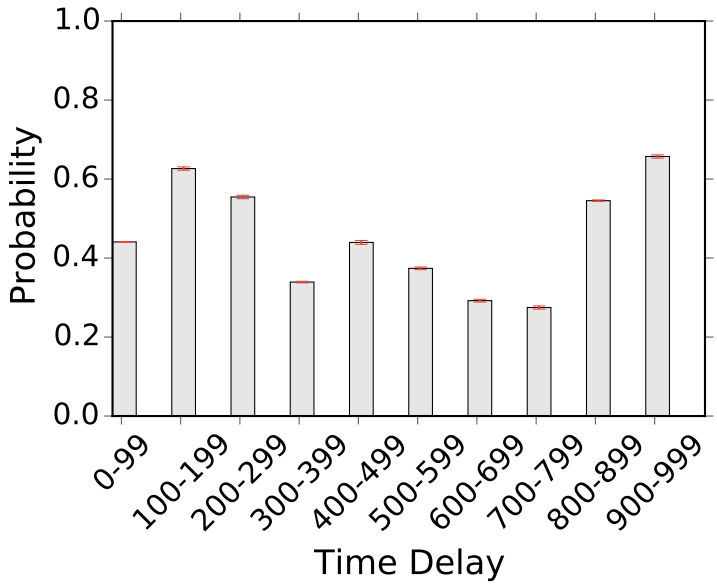


Figure 4.9: Retweet Time Delay. Error Bars Represent Two Standard Deviationss.

Metadata. Table 4.1 shows the conditional probability of positive instances given the meta measure value of 0 and 1, respectively. The values from the table shows

that presence or absence of a link doesn't seem to have much correlation with the influence. It also shows that, the presence of mentions seem have slight negative correlation to influence though there is no actual intuition to base this on. But, this can be explained by the observation in the paper [15] that as the number of mentions in a tweet grows, the expected number of retweets decreases. The presence of hashtag shows an interesting correlation in Table 4.1. This is consistent with the study of [15] and illustrates the significance of hashtags in enhancing the visibility of the tweet and motivating a user to retweet them.

$\vec{V}$	$P(y_i = "pos" \mid V_i = 0)$	$P(y_i = "pos" \mid V_i = 1)$
hasLink	0.51	0.48
hasMention	0.51	0.45
hasHashtag	0.50	0.66

Table 4.1: $\vec{V}$ is a Column of the Design Matrix Corresponding to a Certain Binary Feature,"pos" Represents Positive Label and i is the Index of the Sample

Chapter 5

INFLUENCE PREDICTION

5.1 Methods

We derive our graph G from the dataset as described under Chapter 2. We use the microblogs published in August 2011 to extract the instances to train and test our approach. Positive and negative instances are extracted as described in Chapter 2, and the measures described in Chapter 3 were extracted as features for each of them. This set is used to obtain a random sample with 1:1 negative to positive ratio, which we will use for the classification experiments.

Classification experiments Here we examine our experiments for predicting whether a user under given conditions will retweet or not. As this is a binary classification task we report the performance measurements (precision, recall and unbiased F1) for only the positive (retweeting) class. We also examine the classification performances of various learning algorithms. For each of the experiments we use a training to test set ratio of 70:30 and used a 10 fold cross validation. We use the following classification algorithms for our experiment.

Random Forest (RF). Random Forest [4] is a popular ensemble method used for classification and regression. Ensemble methods use multiple classifier algorithms to obtain better accuracy than that could be obtained using any of the individual classifiers. We use random forest algorithm with bootstrap aggregating, that fits a number of decision trees on different sub-samples of the dataset. Each decision tree provides its own predictions which are then merged obtain a better accuracy.

AdaBoost Classifier (AB). The AdaBoost algorithm [10] proposed by Yoav Freund and Robert Schapire is one of the most important ensemble methods. It is prominent among the boosting techniques [10] which are used in conjuction with other learning algorithms. In this method, the weak learners are combined into a final sum representing the boosted output. We use the particular algorithm called AdaBoost-SAMME [30] and use the decision trees as the base estimator.

Logistic Regression (LR). Logistic regression is a generalized linear model which uses a logistic function to infer the relationship between a dependent variable and one or more independent variables. We utilizes the binomial logistic regression which predicts the probability that an observation falls into one of the two categories. Logistic regression has low varience and is less prone to overfitting.

Naive Bayes Classifier (NB). Naive Bayes is a probabilistic classifier which is based on applying Bayes' theorem with independence assumption between every feature pairs. Naive Bayes classifiers are highly scalable and less prone to the curse of dimensionality, making it one of the top machine learning algorithms. We implement the Gaussian Naive Bayes algorithm for classification where the likelihood of the features is assumed to be Gaussian.

5.2 Measurement Group Comparison

Here we compare the classification performance of the various measurement groups described in Chapter 3. Figures 5.1, 5.2, 5.3 and 5.4 shows the behavior of different feature groups using multiple classifier algorithms. Generally Random Forest provides the best performance among all the classifier algorithms. Neighborhood-based (Nbr)

measures performs quite well in Random Forest, AdaBoost and Logistic regression. This is consistent with what we discussed in Chapter 4. Structural diversity measures shows less performance compared to other groups. This can be attributed to the fact that it is not often used independently in classification, and usually this group performs well in conjunction with other measures such as Neighborhood-based. LRC-Q gives performance measure comparable to the results in [29]. Cascade-based measures are observed to perform reasonably well in Random Forest, Logistic Regression and AdaBoost. This once again illustrates the significance of cascade size and bring into focus the path length measure. Temporal measure performs well in all classifiers except Naive Bayes. Although time based measures are frequently used as a decay factor in conjunction with other measures ([11], [29]), our results show that it could yield high predictive power by itself. Metadata measures shows good and consistent performance across all classifiers. As research by [15] shows, hashtag and mentions have high predictive power with respect to retweet behaviour and our results confirms the significance of this measure along with the hasLinks measure.

We also examine a "Multi-Measurement model" which is a combination of Neighborhood, Structural, Cascade, Temporal and Metadata measures. The Multi-Measurement model shows better performance than individual groups generally among Random Forest, Logistic Regression and AdaBoost classifiers. The other measures such as neighborhood-based, temporal and LRC-Q performs reasonably well compared to rest of the individual future groups. The performance of Multi-Measurement model shows a real value in combining the various features and individual feature groups to improve our ability to predict retweet behavior in real world datasets.

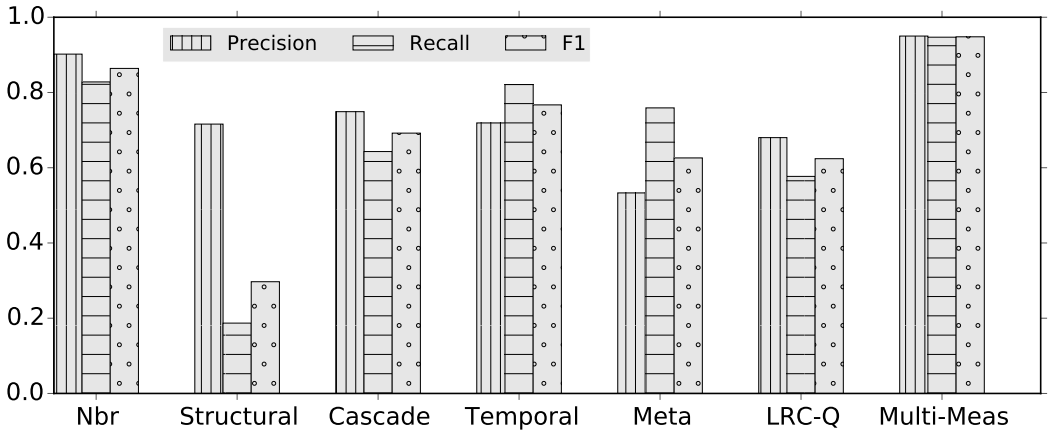


Figure 5.1: Random Forest

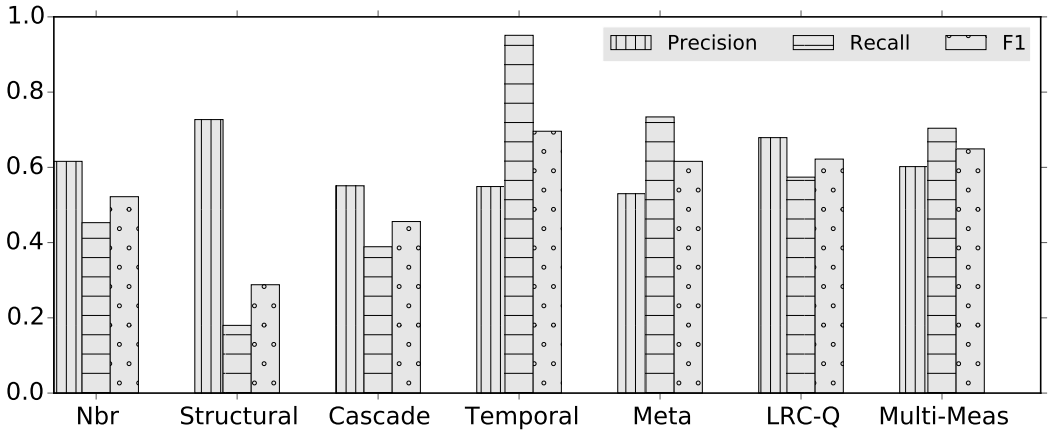


Figure 5.2: Logistic Regression

5.3 Multi-Measurement Model Compared to Influence Locality

We compare our results with the LRC-Q model described in [29]. We experimented with multiple classification algorithms for this task and the best results were obtained using Random Forest classifier. The results obtained using Random Forest (RF), Logistic Regression (LR), Naive Bayes (NB) and AdaBoost (AB) are shown in the Table 5.1 . As LRC-Q uses only a single feature, we only use Logistic Regression for it's evaluation. It can be observed that Multi-Measurement model outperforms

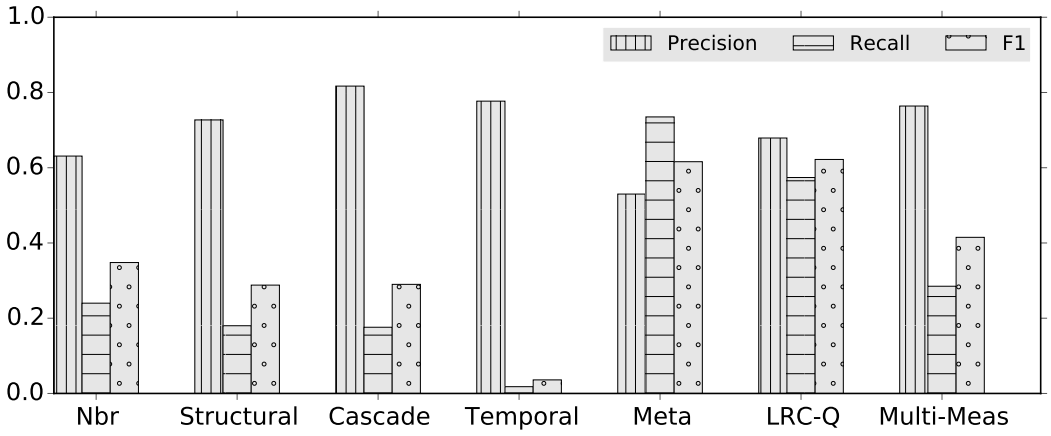


Figure 5.3: Naive Bayes

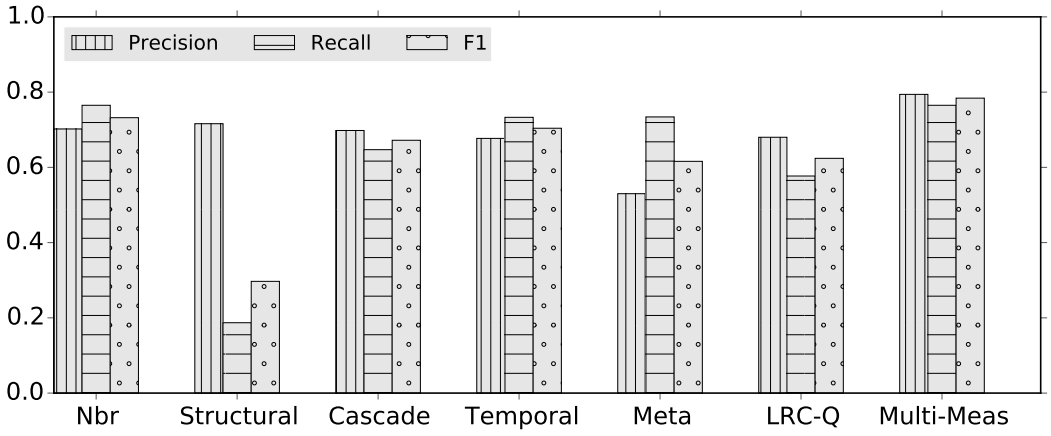


Figure 5.4: AdaBoost

the LRC-Q model in all classifiers except for Naive Bayes. This can be attributed to the fact that while LRC-Q takes into account pairwise and structural influence along with time decay, Multi-Measurement model incorporates more parameters in addition to the above. LRC-Q has combined the pairwise and structural factor into a single feature and uses time measure as a decay factor. The Multi-Measurement model on the other hand treat them individually, along with including different kinds of pairwise influence (such as active neighbor count, personal network exposure and average in-neighbors of active neighbors), considering both direct as well as ratio based mea-

sures for structural diversity, and using temporal measure as an independent feature. In addition to that, this model also includes cascade and metadata based features giving it a broader view of the parameters that can influence an individual's retweeting behavior. This demonstrates that in any attempt of retweet prediction, a broader approach is required, which incorporates multiple measures that are are closely related (within the measurement groups) and those that are mutually exclusive (across groups) to obtain the best prediction in classification.

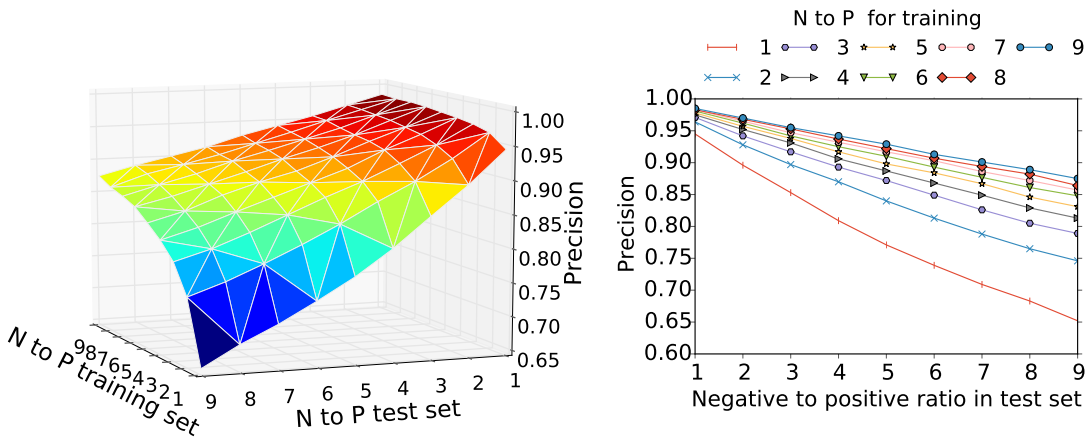
Model	Precision	Recall	F1
LRC-Q (LR)	0.679	0.573	0.622
Multi-Measure (RF)	**0.95**	**0.947**	**0.948**
Multi-Measure (AB)	0.794	0.765	0.784
Multi-Measure (LR)	0.602	0.704	0.649
Multi-Measure (NB)	0.764	0.285	0.415

Table 5.1: Performance of Retweet Behavior Prediction

5.4 Varying Negative to Positive Ratio

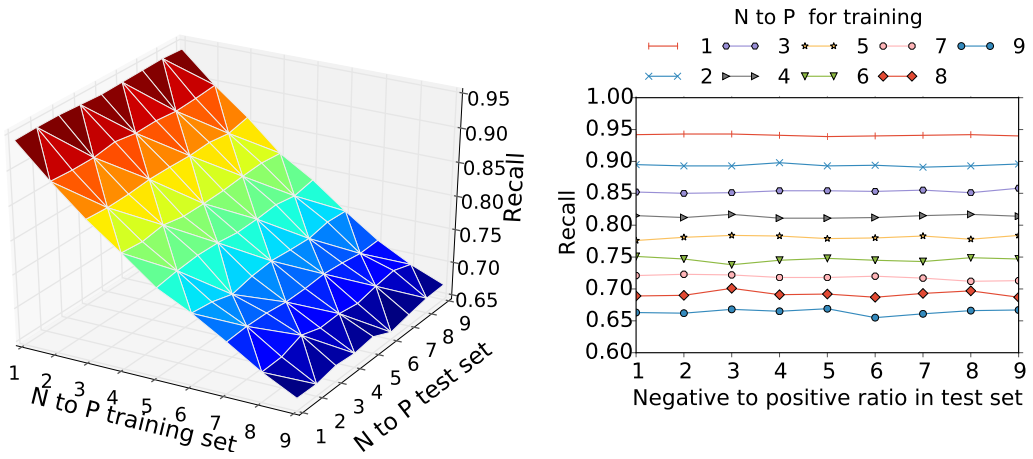
An important question when deploying the aforementioned methods in a real-world application is how to best train the model to cope with data imbalance observed in-practice. As individuals are exposed to an arbitrarily large number of microblogs that they do not rebroadcast, this is a difficult - and unfortunately relatively unstudied problem. Here, we conducted experiments to analyse how classification performance varies with different negative to positive ratio in both training and test set. The surface and linear plots in Figure 5.10 shows the precision, recall and F1 values obtained using Random Forest classifier, when negative to positive ratio is varied

from 1:1 to 9:1. The ratio was varied in both training set and test set to observe the effects on overall performance. Precision is observed to decrease as we increase the size of negative samples in test set while keeping the ratio in training set constant. Recall is observed to remain the same with changing ratio in test set. Change in negative to positive ratio in training set on the hand, shows slight increase in precision where as recall decreases. Results for LRC-Q follows a similar pattern except for the convergence of recall for increased imbalance in training set. From these results, it can be generally observed that 1:1 is the ideal ratio of negative to positive samples in training set for an unknown imbalance in test data.



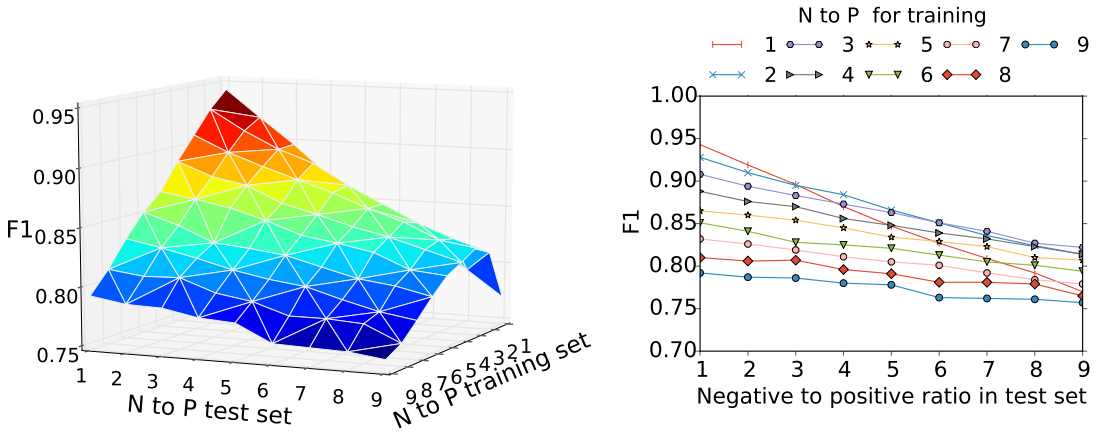
(a) (b)

Figure 5.5: Precision for Classification on Imbalanced Data for Multi-Measurement Model Using Random Forest. a) Surface Plot b) Line Plot



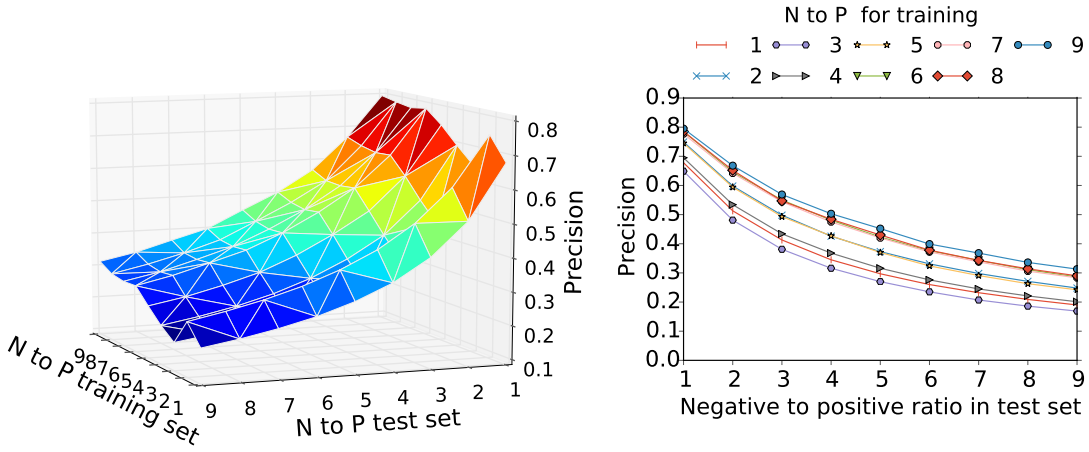
(a) (b)

Figure 5.6: Recall for Classification on Imbalanced Data for Multi-Measurement Model Using Random Forest. a) Surface Plot b) Line Plot



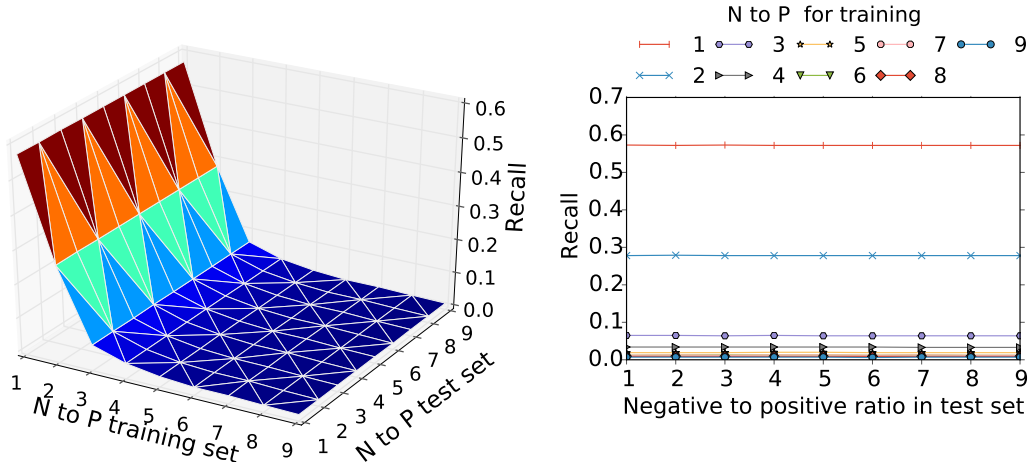
(a) (b)

Figure 5.7: F1 for Classification on Imbalanced Data for Multi-Measurement Model Using Random Forest. a) Surface Plot b) Line Plot



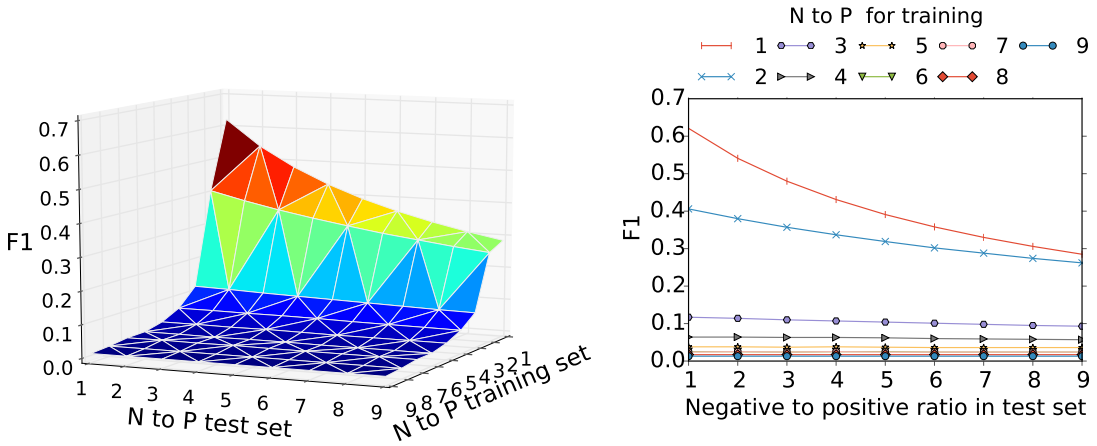
(a) (b)

Figure 5.8: Precision for Classification on Imbalanced Data for LRC-Q Using Logistic Regression. a) Surface Plot b) Line Plot



(a) (b)

Figure 5.9: Recall for Classification on Imbalanced Data for LRC-Q Using Logistic Regression. a) Surface Plot b) Line Plot



(a) (b)

Figure 5.10: F1 for Classification on Imbalanced Data for LRC-Q Using Logistic Regression. a) Surface Plot b) Line Plot

Chapter 6

CONCLUSION

This thesis examines the performance of a wide variety of social network based measurements and study the probability of an individual becoming influenced based on them. In this study, those measures were grouped under various measurement groups to understand their group wise predictive power. We construct our Multi-measurement model using these groups and use various classification algorithms to evaluate it. Our experiments show that Multi-Measurement model outperformed the individual group measures as well as the described baseline approach. These results show that the studied parameters can significantly improve the predictability of real world diffusion process. As there is a major imbalance between positive and negative instances in real world datasets, we also experimented with different negative to positive ratios to identify the one that best suits real world applications. Our experiments show that the 1:1 negative to positive ratio is the most suitable one in this regard. Our results demonstrates that while learning from historical data using classification techniques, a broader approach is required which incorporates and combines diverse measurements that forms the various aspects of an individuals retweet behavior. Such a broader approach can yield better predictive power than any of the individual measures thus bringing it closer to real world application.

REFERENCES

[1] Al-khateeb, S. and N. Agarwal, "Examining botnet behaviors for propaganda dissemination: A case study of isil's beheading videos-based propaganda", in "ICDM Workshops", pp. 51–57 (IEEE, 2015).

[2] Bakshy, E., J. M. Hofman, W. A. Mason and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter", in "Proceedings of the fourth ACM international conference on Web search and data mining", pp. 65–74 (ACM, 2011).

[3] Blondel, V. D., J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks", Journal of Statistical Mechanics: Theory and Experiment **2008**, 10, P10008 (2008).

[4] Breiman, L., "Random forests", Machine learning **45**, 1, 5–32 (2001).

[5] Centola, D., "The Spread of Behavior in an Online Social Network Experiment", Science **329**, 5996, 1194–1197 (2010).

[6] Cha, M., H. Haddadi, F. Benevenuto and P. K. Gummadi, "Measuring user influence in twitter: The million follower fallacy.", ICWSM **10**, 10-17, 30 (2010).

[7] Chen, M., K. Kuzmin and B. K. Szymanski, "Community detection via maximization of modularity and its variants", Computational Social Systems, IEEE Transactions on **1**, 1, 46–65 (2014).

[8] Cheng, J., L. Adamic, P. A. Dow, J. M. Kleinberg and J. Leskovec, "Can cascades be predicted?", in "Proceedings of the 23rd international conference on World wide web", pp. 925–936 (ACM, 2014).

[9] Fortunato, S., "Community detection in graphs", Physics reports **486**, 3, 75–174 (2010).

[10] Freund, Y., R. Schapire and N. Abe, "A short introduction to boosting", Journal-Japanese Society For Artificial Intelligence **14**, 771-780, 1612 (1999).

[11] Goyal, A., F. Bonchi and L. V. Lakshmanan, "Learning influence probabilities in social networks", in "Proceedings of the third ACM international conference on Web search and data mining", pp. 241–250 (ACM, 2010).

[12] Guo, R., E. Shaabani, A. Bhatnagar and P. Shakarian, "Toward order-of-magnitude cascade prediction", in "Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015", pp. 1610–1613 (ACM, 2015).

[13] Halavais, A., K. H. Kwon, S. Havener and J. Striker, "Badges of friendship: Social influence and badge acquisition on stack overflow", in "2014 47th Hawaii International Conference on System Sciences", pp. 1607–1615 (2014).

[14] Hong, L., O. Dan and B. D. Davison, "Predicting popular messages in twitter", in "Proceedings of the 20th international conference companion on World wide web", pp. 57–58 (ACM, 2011).

[15] Jenders, M., G. Kasneci and F. Naumann, "Analyzing and predicting viral tweets", in "Proceedings of the 22nd international conference on World Wide Web companion", pp. 657–664 (International World Wide Web Conferences Steering Committee, 2013).

[16] Kempe, D., J. Kleinberg and É. Tardos, "Maximizing the spread of influence through a social network", in "Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 137–146 (ACM, 2003).

[17] Kwak, H., C. Lee, H. Park and S. Moon, "What is twitter, a social network or a news media?", in "Proceedings of the 19th international conference on World wide web", pp. 591–600 (ACM, 2010).

[18] Liu, L., J. Tang, J. Han, M. Jiang and S. Yang, "Mining topic-level influence in heterogeneous networks", in "Proceedings of the 19th ACM international conference on Information and knowledge management", pp. 199–208 (ACM, 2010).

[19] Myers, S. A., C. Zhu and J. Leskovec, "Information diffusion and external influence in networks", in "Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 33–41 (ACM, 2012).

[20] Saito, K., R. Nakano and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model", in "Knowledge-based intelligent information and engineering systems", pp. 67–75 (Springer, 2008).

[21] Shakarian, P., A. Bhatnagar, A. Aleali, R. Guo and E. Shaabani, *Diffusion in Social Networks* (Springer, 2015).

[22] Tang, J., J. Sun, C. Wang and Z. Yang, "Social influence analysis in large-scale networks", in "Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining", pp. 807–816 (ACM, 2009).

[23] Ugander, J., L. Backstrom, C. Marlow and J. Kleinberg, "Structural diversity in social contagion", Proceedings of the National Academy of Sciences **109**, 16, 5962–5966 (2012).

[24] Valente, T. W., *Network models of the diffusion of innovations*, Quantitative methods in communication (Hampton Press, Cresskill, N.J., 1995), thomas W. Valente. Includes bibliographical references (p. 153-163) and indexes.

[25] Watts, D. and J. Peretti, "Viral marketing for the real world", Harvard Business Review (2007).

[26] Watts, D. J. and S. H. Strogatz, "Collective dynamics of small-worldnetworks", nature **393**, 6684, 440–442 (1998).

[27] Weng, L., F. Menczer and Y.-Y. Ahn, "Virality prediction and community structure in social networks", Scientific reports **3** (2013).

[28] Weng, L., F. Menczer and Y.-Y. Ahn, "Predicting successful memes using network and community structure", in "Eighth International AAAI Conference on Weblogs and Social Media", (2014).

[29] Zhang, J., B. Liu, J. Tang, T. Chen and J. Li, "Social influence locality for modeling retweeting behaviors.", in "IJCAI", vol. 13, pp. 2761–2767 (2013).

[30] Zhu, J., H. Zou, S. Rosset and T. Hastie, "Multi-class adaboost", Statistics and its Interface **2**, 3, 349–360 (2009).