Master’s Thesis
im Studiengang "Angewandte Informatik"

Analyzing User Behavior in Online Social Networks

Xu Luo

am Lehrstuhl für Computer Networks

Bachelor’s and Master’s Theses of the Center for Computational Sciences at the Georg-August-Universität Göttingen

28.09.2015
I hereby declare that I have written this thesis independently without any help from others and without the use of documents or aids other than those stated. I have mentioned all used sources and cited them correctly according to established academic citation rules.

Göttingen, 28.09.2015
# Contents

1 Introduction
   1.1 Thesis Motivation .................................................. 4
   1.2 Thesis Organization .................................................. 4

2 Related Work .............................................................. 7

3 Web and Mobile Users Behavior Comparison ................................. 9
   3.1 Data Collection and Analysis ........................................ 9
   3.2 Collective Normal User Behavior .................................... 10
   3.3 Comparison of the 10% of Users who have the most Followers .......... 17
   3.4 Comparison of the 10% of Users who Tweet the most .................. 21
   3.5 Comparison of Users in The Japan, UK, and USA Behavior ............... 23
      3.5.1 Behavior of Users in Japan .................................... 23
      3.5.2 Behavior of Users in UK ....................................... 26
      3.5.3 Behavior of Users in USA ...................................... 30

4 Data Fusion and Stratified Sampling ...................................... 33
   4.1 Data Fusion ............................................................ 33
   4.2 Stratified Sampling .................................................... 34
   4.3 Feature Draw .......................................................... 35

5 Prediction Method ......................................................... 37
   5.1 General Modeling Framework ........................................ 37
   5.2 Logistic Regression .................................................... 39
      5.2.1 Odds ............................................................. 39
      5.2.2 Odds Ratio ...................................................... 39
      5.2.3 Logistic Model ................................................... 39
   5.3 SVM ................................................................. 40
      5.3.1 Linearly-separable Data ......................................... 41
      5.3.2 Finding the Boundary ............................................ 41
# CONTENTS

5.3.3 Nonlinear Decision Boundary .......................................................... 42
5.3.4 Kernel Trick and Kernel Function ..................................................... 43
5.3.5 Gaussian Kernel .................................................................................. 45

6 User Type Prediction ................................................................................. 47
   6.1 Logistic Regression with Single Feature ............................................. 49
   6.2 SVM with Single Feature .................................................................... 51
   6.3 Logistic Regression with Data Fusion ................................................. 53
   6.4 SVM with Data Fusion ........................................................................ 54

7 Conclusion and Future Work .................................................................... 57
   7.1 User Behaviors .................................................................................... 57
      7.1.1 Tweet Size Distribution ............................................................... 57
      7.1.2 Tweet Interval .............................................................................. 57
      7.1.3 Online Activity ............................................................................ 58
      7.1.4 Retweets ...................................................................................... 58
      7.1.5 Tweet Type Distribution ............................................................... 58
   7.2 Prediction ............................................................................................ 58
      7.2.1 Logistic Regression .................................................................... 59
      7.2.2 SVM ......................................................................................... 59
   7.3 Future Work ....................................................................................... 60

List of Tables .............................................................................................. 61
List of Figures .............................................................................................. 63
Bibliography ................................................................................................. 65
Chapter 1

Introduction

Online social networks today represent some of the world’s most popular social platforms, have changed our everyday forms of communication, created a convenient way to interact with friends, and provided an important means of information sharing on the Internet. Twitter is a powerful and popular social network, through which users can send and receive short public messages, including images, videos and other sources that create a method of micro-blogging, on Twitter, opinions and comments can be freely expressed and lifestyle concepts can easily be shared. There are over one billion Tweets flow through Twitter every week, a wide range of people, organizations and institutes from all over the world use Twitter to share, announce and publish their present events from their computer, smartphone or tablet. Mobile devices and Apps are gradually becoming important tools for personal and work life, and provide more flexibility and dynamic connections regardless of the distance or familiarity.

Our work refers to the data mining of big data in Twitter, where data mining is a process of information discovery from incomplete, noisy, messy or random data in a practical application or database, where data is taken from a big dataset in order to find acceptable, understandable, applicable information, which is then combined into the familiar cognitive mode for decision making. However, with the increased amount of information the probability of finding valuable information is decreased. In our work, we focus on two types of users: mobile users and web users, analyze and discuss their different but related activity patterns, for example the growing popularity of using mobile devices to connect to this network has revealed patterns of different activity in comparison to browser users, such as differences in login time, frequency and activity. From these different user behaviors relevant patterns and trends can be defined, and through these defined behavior patterns an effective mathematical model with apposite Eigenvectors can be created to predict user type. This is meaningful for understanding the correlation in seemingly random data, and allows us to track user communication patterns of hybridity in social network.
1.1 Thesis Motivation

Most of the research work done until now in the field of social networks focused on analyzing single location or single user group behavior. In contrast, our work concentrates on a comparison of user’s behaviors in different groups, thus the 10% of users who have the most followers and 10% of users who Tweet the most are discussed as a horizontal comparison, and the user behavior from the USA, UK, and Japan is discussed as a longitudinal comparison. Through this work we can analyze the comparison between the behavior of normal users and specific users. One challenge lies in defining the user type, as there are over 7,000 devices or application types in the dataset, and to distinguish between these devices is a significant preparation for later analysis of the work.

Through these behaviors the Eigenvector can be selected and aggregated in a single matrix for user type prediction, however, the second challenge is that computing all the data from the dataset to create a prediction model is not practical due to the computing-complexity and time-consumption, thus another motivating theme is the comparison of prediction models. logistic regression and SVM with an effective data sampling method which can approximately represent the original data is of significant work for prediction. Another challenge is to improve the prediction method accuracy for the selected simple data through a feasible approach without increasing the complexity, therefore selecting and combining apposite Eigenvectors as input data is a significant prerequisite for prediction.

1.2 Thesis Organization

The rest of thesis is organized as follows.

Chapter 2 gives an overview of existing data analysis for social networks and data sampling approaches for prediction.

In Chapter 3 the behavior of the different users groups is discussed. The first behavioral comparisons are observed in normal users, and include time spent online, Tweets size length, and Retweet Interval, etc. The 10% of users who have the most followers and the 10% of users who tweet the most are selected as specific user groups, to be further discussed in the second behavioral comparison. The last comparisons use the users from countries with the greatest Twitter activity: the USA, the UK, and Japan.

In Chapter 4 the data fusion method is discussed, where the different data are sampled into a single data matrix with stratified aggregation.

In Chapter 5 prediction methods are discussed, or in the other words, a general framework for modeling, explanation of logistic regression and SVM, and for how these methods work with the parameters to separate and classify data.
In Chapter 6 the prediction results are shown, where through the users’ behavior a specific mathematical model can be built to predict the users’ device type, where methods of random sampling and stratified sampling may be considered, and logistic regression and SVM will be used as the prediction method to analyze their accuracy. A comparison using logistic regression and SVM with single feature and data fusion is necessary for observation of efficiency of the different prediction methods’.
Chapter 2

Related Work

Twitter is currently one of the most popular social network in existence and as such has attracted much attention in past years. Through the rise of the mobile phone new devices and platforms have become available, exemplified by instances such as when 270,000 iPhone sold in the first 30 hours of the launch weekend in the U.S market in July 2007 [1], mobile devices have changed the way people organize themselves and their relationships [2]. Because of this, there is already a body of social network research related to the analysis of Twitter user’s behavior through the exponential expansion of the Twitter population. Java et al. [3] have published a preliminary analysis of Twitter in 2007, and much work has focused on the difference in the information and search usage behavior between mobile and desktop user’s [4].

Ghose and Han [7] quantified how mobile users Internet usage relates to some specific unique characteristics of the mobile Internet such as browsing, downloading, and sharing content, using mobile devices [5], and their behavior in the dominant application categories, online game and email [6]. However, few of these works are based on big data, which contains huge amounts of valuable and diverse information. Although a single user is seen as a discrete point in the network, there are some groups or communities making networks which are logically modular with certain characteristics. Analysis of the patterns of how different users perform complex information activities is a significant line of related work.

Gao [8] investigated location-based social networking and showed that these users bridge the gap between real-word and online social media. Rafa Absar et al. [9] showed that there are many similarities between social and non-social information behavior in terms of satisfaction and motivation. There is also some work published on online social networking which focuses on mobile user’s behavior, for example, Long Jin et al. [10] has studied user behavior from four perspectives: connection and interaction, traffic of activity, mobile social behavior, and malicious behavior. Zhu et al. [11] has designed a social application in Symbian system in order to analyze the user’s social behavior in mobile social networks, which shows mobile users activity online time.
duration is approximately equal to the user’s daily life-work pattern. Social network investigation refers not only to analyzing mobile user behavior, but also to the study of specific user behavior, such as location-based social networking, for example Harada et al. [12] conducted research on some special users, referring here to elderly participants, to analyze their needs and issues based on mobile multi-touch devices. Comparing the activity of these users with normal social networking will demonstrate common characteristics, with which we can construct trends of user activity from the uncertainty of social network data. However, few of these works can properly be used as a comparison to a mobile and web user behavior study on Twitter.

Our work is also related to the area of prediction, accurate prediction requires a representative sample as a prior condition. We focus on stratified sampling method, which is selected by Mckay [13] to compare with two other methods for selecting values for input variables in the analysis of output from a computer code, the result of which showed that stratified sampling with proportional allocation offers an improvement over random sampling. Podgurski [14] used stratified sampling to estimate software reliability and has significantly reduced the number of program executions needed. James L. Crowley [15] and Triadaphillou [16] have introduced a framework for dynamic world modeling with data fusion for industrial fermentation processes through data fusion, using data fusion is valuable in order to observe whether increasingly accurate predictions can be formed by applying multiple Eigenvector aggregation.

Another line of related work is to compare two prediction methods: Logistic Regression [18, 19] and SVM [20]. Logistic regression applies a continuous function to fit the input data, where the prediction result is a probability. In contrast, SVM attempts to classify two classes of data using a hyperplane. There are some work of comparison, such as Huang [17] has studied prediction methods and analyzed the results of logistic regression and SVM. However, none of these works can properly be compared with using stratified sampling to predict the data using a single Eigenvector or multiple Eigenvectors.
Chapter 3

Web and Mobile Users Behavior Comparison

In this section, the user size distribution, Tweet interval, online activity, activity on ‘Special days’, retweet distribution and Tweets’ properties will be analyzed and compared in the collected dataset. This analysis indicates that the mobile and web users’ have a strong connection, but different social behavior. To analyze the statistics, all the users’ Tweets have been collected and standardized as a percentage of the total values.

3.1 Data Collection and Analysis

Data has been collected from more than 250,000 Twitter users. Each user’s file contains four subfiles, namely profile information, tweets, friends and followers- the dataset comprises of 84,625,897 friends and 78,986,681 follows in total. Through analysis of the dataset some characteristics of the user activity can be described, for instance user online time, retweets activity, tweets type and so on. However, no personal message related information has been collected.

<table>
<thead>
<tr>
<th>Column</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>*._followers</td>
<td>Followers list</td>
</tr>
<tr>
<td>*._friends</td>
<td>Friends list</td>
</tr>
<tr>
<td>*._profile</td>
<td>Basic user data</td>
</tr>
<tr>
<td>*._tweet</td>
<td>Tweets list</td>
</tr>
</tbody>
</table>

Table 3.1: User Data File Structure

Although Twitter is an American company, this has not made it difficult for various countries’ users to use it to send and read messages. Figure 3.1 shows the proportion of nationalities of registered users in Twitter. The users live on different continents and come from 35 countries, the three regions where the most users live in are: USA, UK, and Japan. Obviously the majority of
users come from the USA, which has 47%, followed by the UK with 26%, and then Japan with 8%; these three together account for more than 80% of users. An interesting phenomenon is that in this location distribution there are no Chinese users- this does not mean that the Chinese are not interested in social networking or microblogging, quite the contrary, but there is Weibo in China, which is an active social networking provider.

There are four files which can define a Twitter user, the first two files contain user’s social connections: “*_followers” means the followers list are allocated in this file, and the user’s friends are in “*_friends” file. The descriptions of user’s information are saved in “*_profile” and shown in following table3.2.

All users’ Tweets are saved as a file “*_Tweet”, the structure of which contains eight parts which exactly define the single Tweet, the structure is showed in the following table 3.3.

### 3.2 Collective Normal User Behavior

Figure 3.2 shows all users’ behaviors of Tweets Size, Interval Size, Online Time, and Retweets. Distribution can help us understand the writing behavior of different users. The y-axis represents the percentage of total Tweets that are of that length. The curve of the mobile users’ distribution is similar to the log-normal distribution, which is an asymmetric probability distribution. It indicates that the mobile users prefer short Tweets, potentially because of the expensive data volume price
### Table 3.2: Profile Structure

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>User’s location</td>
</tr>
<tr>
<td>Timezone</td>
<td>User’s timezone</td>
</tr>
<tr>
<td>Website</td>
<td>User’s website</td>
</tr>
<tr>
<td>favouritesCount</td>
<td>Number of favourites</td>
</tr>
<tr>
<td>followersCount</td>
<td>Number of counts</td>
</tr>
<tr>
<td>friendsCount</td>
<td>Number of friends</td>
</tr>
<tr>
<td>statusesCount</td>
<td>Number of Statuses</td>
</tr>
<tr>
<td>protectedUser</td>
<td>True or False</td>
</tr>
</tbody>
</table>

### Table 3.3: Tweet Structure

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweet’s ID</td>
<td>11 digit</td>
</tr>
<tr>
<td>ReTweets</td>
<td>No Retweets: -1</td>
</tr>
<tr>
<td></td>
<td>11 digit’s retweets ID</td>
</tr>
<tr>
<td>Length</td>
<td>Tweets Size Length: 1-140</td>
</tr>
<tr>
<td>Device</td>
<td>User’s device</td>
</tr>
<tr>
<td>Time</td>
<td>Day Month Date</td>
</tr>
<tr>
<td>Hour: Minute: Second</td>
<td></td>
</tr>
<tr>
<td>Timezone</td>
<td>CEST</td>
</tr>
<tr>
<td></td>
<td>CET</td>
</tr>
<tr>
<td>Year</td>
<td>2006-2011</td>
</tr>
<tr>
<td>End</td>
<td>No Retweets: -1</td>
</tr>
<tr>
<td></td>
<td>Retweets: n</td>
</tr>
</tbody>
</table>
CHAPTER 3. WEB AND MOBILE USERS BEHAVIOR COMPARISON

Figure 3.2: Normal Users’ Behavior
3.2. COLLECTIVE NORMAL USER BEHAVIOR

during this study period [21]. The majority of the Tweets are concentrated on the left side of the curve where the peak value reaches 1.3%, because short Tweets are more popular and easier to be delivered. In comparison, the distribution for web users is similar to a cumulative distribution and normally the size of Tweets are homogeneously distributed, which means the users write not only short but also long Tweets, for instance the probability is almost equal for that of a 30 character Tweet to be written as for a Tweet of 120 characters when written through a web browser. This phenomenon indicates that the web users writing behavior may be more focused on the enjoyment of writing long Tweets, as opposed to worrying about the limited data volume.

Interval shows the time interval between a first and second Tweet in mobile and web users. The X coordinate represents the time interval; the y coordinate expresses the percentage of the entire probability distributions. The curve strongly indicates that both types of users can be intuitively understood as a multiplicative inverse function, so that the two curves are infinitely close to the x axis. Compared to the web users, the majority of mobile phone users send a second Tweet after a relatively short time interval, usually within an hour. For web users, the frequency of sending Tweets is a more linear interval compared to the mobile phone users. 35% of web users would update the second Tweet within one hour, about 30% of users within two hours and about 20% within three hours. Both types of users rarely send a Tweet again after five hours, so the tails of both curves become slowly infinitely close to the x-axis.

Online Time describes quite clearly two different types of user social behavior. The lowest numbers of Tweets were sent late at night because most users are sleeping. From early morning around 8:00, the quantity of Tweets begins to increase slowly, and it is almost the same Tweet percentage as at midnight (0:00), in the other words at this time most users are less active, perhaps reading news, posting photos, or sharing music with their friends. But starting from 9 o'clock the situation changes, as the percentage of web user activity reaches a small peak for a short time. One explanation is that during this time, when the web users arrive at work, between 9:00 and 11:00, they have to first engage with their daily tasks; therefore their Twitter activity is reduced. As lunchtime approaches, web users Twitter usage begins to perk up again, lasting until around 15:00. Perhaps some web users do not need to work after 15:00, which means that they don’t return to their computer, hence the curve once again falls to a lower point, but again at 20:00 the usage has reached the highest point of the day.

Because mobile device users have naturally greater flexibility, they can write or share Tweets anywhere, anytime without geographic limitation. As a researchable specific feature, “Retweet” allows friends or fans to share the user’s Tweet. At the same time, additional text can also be added, which makes Retweet’s role in microblogging more personal.

Retweets Activity shows how mobile users’ and web users’ behavior can appear really quite different. During the daytime the percentage of cell phone users Retweet activity is almost similar to a slow growth linear curve, where there exists hardly any visible dips or peak. This shows that
the probability of Retweets from mobile phone users is in a very stable distribution throughout the
day.

On the other hand, the probability distribution for web users’ Retweets trend is similar to a negative
skew, with most of the probability at the end of the figure, that is, the distribution is skewed to the
left. In order to classify the different types of Tweets, the following Pie charts and Heat maps in
the following four figures have been formed, so that the Tweet type distribution and the activity at
different time points between the two types of users can be easily analyzed.

![Figure 3.3: User Activity in Christmas Day and Easter Sunday](image)

In the Figure 3.3 which depicts Easter Day and Christmas Day, the two user group’s behaviors
seem quite the opposite to that of work days. One reason may be that because of the holidays
the users and their families or friends prefer to talk and gather together, so the reliance on social
networking is reduced. This phenomenon in web users is more obviously due to the reduction of
time spent at the computer due to being with the family, hence their entire activity extent during
the day time has reduced, unlike during workdays when the curve dips and peaks constantly.
Even more surprising is that in the entire 24 hours, both types of users’ activity reaches the lowest
point at noon. Another reason for this may be because the two types of users are enjoying the
holiday, they are more willing to enjoy their free time without using the internet. However, mobile
users in these two special holidays are more active in the evening, when they share the day’s
images, news and messages via Twitter.

Unsurprisingly, most Tweets are about social life, which in mobile phone users accounts for 77%, and in web users for 64%. In other classifications, it could be clearly seen that mobile phone users prefer to use the app to view the data transfer amount generated by their own microblogging, number of page visits, friend amount and other information, which all in all make up about 8 percent. The other themes are ‘Read’, ‘Twitter related management’, ‘clients’, and ‘sharing’, accounting for a combined 13%. The web users are focusing more on third-party software and reading, for example 14% of time is spent on Application, and 8% on blogging.

The above Figure 3.6 shows the web users heat map. Here, users are more interested in traditional ways of using the network, such as writing blogs, or sending and receiving e-mails, these sections represent the main distribution. Music Tweets are almost evenly distributed except for early mornings, meaning there was almost no increase or decrease. The news distribution begins to peak at 8:00 am, surprisingly the distribution during the morning and in the late night is almost at the same peak level, but in the afternoon has reduced somewhat.

The blow Figure 3.6 shows a heat map of “Mobile Users’ Tweets Type distribution” on other microblogging types when the largest proportion: “the social” is removed. When a mobile phone is no longer used just for calls and messages, people can use it to access more features such as listening to music and radio, and playing games. When the mobile device has mobile network abilities, then users’ behavior such as receiving data, information, or entertainment will be changed.
CHAPTER 3. WEB AND MOBILE USERS BEHAVIOR COMPARISON

Figure 3.5: Mobile User Tweets’ Type Distribution

Figure 3.6: Heat Map Zoom In of User Tweets Type Distribution
3.3 Comparison of the 10% of Users who have the most Followers

Let us now consider three special user groups, which are: the 10% users who have got the most followers, the 10% of users who send Tweets to the largest amount of Tweeters and user groups from the typical Twitter using countries like the USA, UK and Japan. Because 75% of all tweets are sent from these three countries, the study of user’ behavior from these three countries is necessary. We have extracted this data from the dataset to compare the difference between these three groups and the ordinary users.

These users are concentrated in the central node of the network, but also in the top of the power law curve. Such users own a major ‘voice’, in other words, they can dominate the public opinion. In comparison to the previous size distribution, these mobile users’ behavior looks very similar to the normal mobile users, in that they prefer to send short Tweets to describe their lives or work. The difference can be seen in the web users, although the average values appear quite similar, (short character Tweets like 20-70 are quite popular), the distribution tended to a downward-sloping curve which occupies a relatively high percentage after 70 characters.

Figure 3.7 has expressed the distribution of the users who send Tweets at different times, similar
to Figure 3.2, the x-axis represents the time, and the y-axis represents the percentage. The profile for Web users’ is similar to Figure 3.2, where a peak occurs at the beginning of work time, but of course the highest peak was in the evening around 20:00, after users are at home and have finished their evening meal. On the other hand, mobile phone users rise steadily without such large fluctuations. Moreover, the activity of these users’ activity exceeds the web users’ activity throughout working hours. Presumably, these users are not too dependent on computer equipment for accessing Twitter, and so prefer to use their mobile phones for accessing social media.

Figure 3.8: Most followed Mobile User Tweets Type Distribution

Figure 3.8 shows that most Tweets are “social” and Twitter Related. And among these particular users, the social-related Tweets have claimed the majority, the other relatively big parts are “Twitter” and “Twitter management”. Although most of the data looks quite similar to the normal user data, some finer details can be observed which depict tiny differences between user types. Apart from the distribution of such “social” values, the other values are distributed more evenly; “microblogging data view”, “sharing the blog”, or “reading news”, “listening to music” or “reading” also taken small sections.
3.3. COMPARISON OF THE 10% OF USERS WHO HAVE THE MOST FOLLOWERS

Figure 3.9: Most followed Web User Tweets Type Distribution

Figure 3.9 shows many web users use software to view, update, and publish their microblogs; this category accounts for nearly 80%. As can be seen in this figure, unlike mobile users, the web users are most concentrated in the social wedge, although there are still considerable distributions elsewhere, hence web users can be regarded as usually preferring to focus on the social aspects of Twitter as opposed to other activities.
Figure 3.10: User’ Tweets Type Distribution Heat Map Zoom In
As can be seen in Figure 3.10, the x-axis represents the type of use and y-axis represents the time of use. Web users’ three most beloved microblogging types are: blogging, sharing and social media which have almost 24 hours activity. The blog has the highest activity, and peaks at about 8,000 Tweets per hour in the afternoon.

The News has high activity in the early morning and in the late night; this also seems to fit to many people’s habits of reading News after getting up or before sleeping. The mobile users have a different pattern of Tweet Types; first the Application occupies a greater distribution, where between 6:00 and 22:00 the average activity is 500 Tweets per hour. Blog is the second most popular activity Type due to the use of Tweets during working hours between 9:00 and 11:0 and 13:00 to 15:00. Analyses is the most popular Type, which shows the users’ often focus on observing their social data, such as Tweets read number, Retweets number, or Comment number.

### 3.4 Comparison of the 10% of Users who Tweet the most

The 10% of the people who Tweet the most also have valuable characteristics.

Figure 3.11: Behavior of Users Who Tweets the Most

The time interval between two Tweets is a little bit different for the two different user types. First of
all, most users release a second Tweet no more than an hour after the first, but there lies a difference in that there are many users who send micro-blogs in the second hour or third hour after the initial tweet, unlike normal users. The web users also reach micro-peaks at 8:00 and 22:00.

Online activity of such users is very interesting. Web users and mobile users’ use grow at almost the same frequency, both mobile users and web users increase rapidly in the early morning, although afternoon activity shows some volatility, but overall the trend has not slowed down. They reach the peak at almost the same time. After this, activity quickly reduces.

Users’ behavior during specific holiday days appear a little different; web users have a very reduced Tweet output. During the average day’s working time, the activity distribution of web users is higher than for many mobile users, and the relative similarity of the curves for web users’ and mobile users’ activity can be compared, and although web users reach a higher peak, they also contain some unusually low values.

In Figure 3.14 shows heat map, web users’ blog and social use take up the greatest portion. A feature to be noted is that the distribution of news, at midnight and noon holds a relatively high proportion of the chart.
3.5 Comparison of Users in The Japan, UK, and USA Behavior

Unlike other categories of mobile users, for most users who send Tweets during the day there exists a very even distribution of blog, music, news, game, photo and radio usage. This shows that these users are not just interested in these content types, but may also potentially work in these industries, such as media information, news dissemination, game reviews and so on, so they will have tendency to use Twitter during working hours in order to describe their relevant jobs. After work, the number of such Tweets rapidly decreases.

3.5.1 Behavior of Users in Japan

In Figure 3.15 the trend in Tweet length is visibly different here. An ordinary distribution of the Tweets length from web users is generally from 20 characters to 120 characters and was very flat. After 50 characters, it could be seen that in both users there exists a similar rate of decrease.
Figure 3.14: User’s Tweets Type Distribution Heat Map Zoom In
3.5. COMPARISON OF USERS IN THE JAPAN, UK, AND USA BEHAVIOR

Figure 3.15: Behavior of Japan Users
Both users send tweets over different time intervals in the first 5 hours. Mobile users’ intervals decrease rapidly after 1 hour, however, the interval from 2 to 5 hours decreases linearly, and then after 5 hours both users send tweets almost in the same time interval distribution.

In Japan, the frequency of the users’ microblogging writing behavior is relatively unstable with great fluctuation in usage. Mobile phone users and web users’ usage in the daytime between 8:00 am to 15:00 pm are both unstable, and mobile phone users prefer to use microblogging after waking up, usually around 7:00, where a small peak can be seen. Web users prefer to write Tweets during work hours, normally after 8:30am. Mobile phone Twitter usage between 8:00 and 9:00 shows a passive state, were microblogging update frequency is very low. Web users at the same time also have a reduction in microblogging activity. The minimum value during this time period is close to 3%. The frequency of microblogging begins to increase after 12:00, and from 15:00 it begins to increase rapidly, peaking at 20:00, when the percentage of web users is 7.8% and the mobile phone users are at 6%.

In general, the probability distribution are almost exactly the same for mobile and web users. In this example however the Retweet use of the web users is unusual. For example from 10:00 to 20:00, we can see from the Retweets Figure that there are two consecutive peaks in common users’ activity. This feature in the user’s Retweet activity in Japan can be seen obviously from 10:00 to 20:00 where the common users’ activity also reaches two peak points. This feature can also be very clearly seen in the user’s Retweet activity in Japan, where in addition to these two peaks, web users’ Retweets distribution is not as high as seen in other users.

Because Japanese users represent 8% of all Twitter users, the Tweets activities in the above heat map clearly shows that the maximum Tweets activity rate per hour is 1,400 compared to 300 Tweets per hour of mobile users, however, the Tweets Type is almost the same as in the other users’ Types, such as Application for the web users, and in this heat map, only Application has such a high frequency, and the other Types seem to be inactive. The situation of mobile users is nearly the same when compared to web users; the users prefer sending Tweets over other activities.

3.5.2 Behavior of Users in UK

From Figure 3.17, it would appear that UK mobile phone users are not very active. Although the length of the microblogging from the mobile phone users is much like that of the other users, most of the Tweets are concentrated between 20 and 60 characters, and the peak is only 1.1% which was the lowest distribution of all types of users. The UK mobile phone users have a more evenly distributed Tweet length, although they prefer to send shorter microblogging messages. The time interval in the first six hours are almost the same for the two users types. In the first hour, the mobile tweets’ percentage exceeds 40% in comparison to 51% for web users. And in the second hour, both users percentages are practically the same. After 3 hours there are some difference between the two users.
3.5. COMPARISON OF USERS IN THE JAPAN, UK, AND USA BEHAVIOR

Figure 3.16: Japan User’s Tweets Type Distribution Heat Map Zoom In
CHAPTER 3. WEB AND MOBILE USERS BEHAVIOR COMPARISON

Figure 3.17: Behavior of UK Users
3.5. COMPARISON OF USERS IN THE JAPAN, UK, AND USA BEHAVIOR

Retweet activity in UK users has the highest peak of all the types, which appears at 14:00 and 18:00, reaching 3%. The two curves for UK user’s online activity look almost the same, lacking any particularly significant peaks or troughs. From 8:00 to 20:00, the activity of mobile users is between 5% and 6%, and web users’ activity is between 5.5% and 6.5%.

Figure 3.18: UK User’ Tweets Type Distribution Heat Map Zoom In

Figure 3.18 shows in UK the web users like to send Tweets on themes of Social and Blog, which are nearly in the same level for both Tweet types; they reach the peak of 900 Tweets per hour in the daytime. Perhaps at that time, the online media has begun to attract attention, but has not been fully networked on the Twitter. The heat map of the media is only shown in the visible level, which refers to 136 Tweets per hour. Compared to Shares, it is close to an even distribution, and the average number of Share Tweets is 365 per hour, and there is not much fluctuation between work time and evening time.
3.5.3 Behavior of Users in USA

As visible in Figure 3.19, from early morning to 8:00 o’clock in the USA both users’ online activities looks very similar. But in the morning at 9:00 the web users’ online activity drops suddenly and then rises immediately. Meanwhile mobile user’s use maintains a steady fluctuation.

The mobile users’ Interval distribution is almost overlapping at the beginning of the first hour with web users, then web users’ Retweets activity stay at a smooth rate, however, the mobile users are more active than web users after 4 hours.

Online Time and Retweets Distribution seem to be similar to other users, this explains how regardless of whether mobile and web users come from Japan, UK, or USA, they have similar behaviors, and their distributions’ trends have some predictable related characteristics, and through these characteristics it is possible to filter out some features of users’ behavior, referred to as Eigenvectors of prediction. But the difficulty in prediction is that some behaviors are very similar to each other, such as Interval, and they are nearly all satisfied by multiplicative inverses or reciprocal models, although there are differences between some discrete sporadic points. Thus identifying unique behaviors’ features is the most significant preparatory step for successful prediction.
3.5. COMPARISON OF USERS IN THE JAPAN, UK, AND USA BEHAVIOR

Figure 3.20: USA User’ Tweets Type Distribution Heat Map Zoom In
CHAPTER 3. WEB AND MOBILE USERS BEHAVIOR COMPARISON
Chapter 4

Data Fusion and Stratified Sampling

4.1 Data Fusion

In situations where events, activities, and situations must be predicted, data fusion is a relatively new engineering discipline that combines data from multiple but diverse sensors [22]. Data fusion is the process of combining the observation and integration of information into a coherent description of the world [15]. In the early 1990’s, Herman and Kanade [23] characterized the incremental combination of geometric information by combined passive stereo imagery from a sensor, another approach which used incremental construction of a mobile robot using a rotating ultrasonic sensor [24]. These works all presented data fusion as information combination from logical sensors.

Multisensor data fusion techniques are applied to a wide range of areas including in artificial intelligence, pattern recognition and statistical estimation, among others [25]. Data Fusion focuses on the current state of the network, based on past data [22]. The difference between mobile users and web users is often difficult to classify using one or two input variables, in particular using variables such as Tweet Interval, where the interval time between two Tweets is often nearly identical. In order to find a good method for observing such information, the feature will be aggregated in a combination matrix, in which the users’ properties can be best modelled.

Tweets were observed through metadata information contained in each row of data: Tweet ID, start symbol, Tweet length or forwarding Tweet ID, equipment, time, end symbols. Then each user’s file Profile contains the user’s basic information, such as name, screenName and creation date, through which the location for data aggregation can be extracted.

Mobile and web users behavior, such as Tweet Interval, online activity, etc., can be observed intuitively, but to observe the difference between two curves where the difference is less obvious, single data sets were used, such as Tweet length, or online activity to create the model, do training,
and predict the data, but the results were not satisfactory. So using this method we can use data aggregation to extract features and form better results. For example, to compare the two cases, the first case: US mobile users, 14:00, no Retweets, but wrote a 84-character Tweet, and a second case: a mobile user, who made a 84-character Tweet. If the two scenarios were modeled, in an attempt to predict the user’s device, mobile or web type, which established model will be more comprehensive and accurate? In both cases, it is clear the first case possesses more observed variables than the second, and we will compare the results in both cases later.

4.2 Stratified Sampling

Stratified sampling was first introduced by Neyman in 1934 [26], and showed constructed sample that could be representative of all the data, one way to avoid huge errors by random sampling is to improve randomness, not increased simple size [27]. In Figure 4.1 x-axis represents the time, y-axis represents the location, and the color shows active users number in different level, the active user distribution is similar in different location. Our dataset includes 250,000 users and more than 100,000,000 Tweets, it is not always practical when all data are used to build prediction model due to huge memory requirement and complex modeling time cost. One efficient way to reduce costs is to use sampling with repeating, the core method is stratified sampling, which means the dataset will be divided into homogeneous subgroups without overlapping, however, these subgroups can approximative represent the dataset.

The total number of Tweets is more than 100 million, called M, now that all the features are arranged in accordance with the characteristic expression of a matrix. To perform modeling, training and prediction in such a large number of matrixes is almost impossible. So the data must be sampled, to select the most representative sample to model. Big data is then broken down into smaller data. The assumption is that each file contains sets of data. In the previous chapter, the activities of users in different countries, in different time periods, with different behavior have
4.3. FEATURE DRAW

been observed, and it has been found that the area can be divided into four types: 1. UK 2. USA 3. Japan 4. Other. Time is also divided into 24 hours, so there are two axes, one for time, one for area, where all the data can be distributed into the two axis in the composition of the system; then we can put this coordinate system 24 * 4 into a small grid, where each small grid is one hour and an area, and each small grid has n data. The total of the entire small grid is equal to the number of all Tweets, a value here called N.

So the stratified sampling ratio should be:

$$r = \frac{n}{N} \times m$$  (4.1)

A total of 20 million sets of data are selected and divided by stratified sampling method into 1000 files, each file contains 20,000 sets of data. Then for each file modeling, training, and prediction will be performed.

4.3 Feature Draw

Now specific features must be found which have common characteristics to describe the specific users, to predict which device these users use.

- **Country**: 0 - USA, 1- UK, 2- Japan, 3 - Others.
- **Size**: The mobile users prefer short term Tweets compare to the web users’ longer length Tweets.
- **Time**: The time was computed as hour*3600 + minute*60 + second as one number, so that the “time” can be worked with accurately and efficiently.
- **Retweets**: For the normal web and mobile users, the Retweets distribution has almost no overlap.
- **Type**: The 84 chosen Tweet types are also an important predictor.
- **Device**: This class is for the predicted response. -1 represents mobile users, 1 represents web users.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>User Location</td>
</tr>
<tr>
<td>Size</td>
<td>0-140</td>
</tr>
<tr>
<td>Time</td>
<td>H<em>3600 + M</em>60 + S</td>
</tr>
<tr>
<td>Retweets</td>
<td>0 or -1</td>
</tr>
<tr>
<td>Country</td>
<td>0 - USA</td>
</tr>
<tr>
<td></td>
<td>1 - UK</td>
</tr>
<tr>
<td></td>
<td>2 - JAPAN</td>
</tr>
<tr>
<td></td>
<td>3 - OTHERS</td>
</tr>
<tr>
<td>Type</td>
<td>0 - 84</td>
</tr>
<tr>
<td>Device</td>
<td>-1 - MOBILE</td>
</tr>
<tr>
<td></td>
<td>1 - Web</td>
</tr>
</tbody>
</table>

Table 4.1: Feature Structure
Chapter 5

Prediction Method

In this chapter, we analyze the original dataset of data, using the stratified sampling method to identify different characteristics of this method of data collection, using a data aggregation method to combine data, then using SVM to perform analysis, training, and forecasting, so that predictions can be made about whether a specific user belongs to the web or mobile group.

5.1 General Modeling Framework

Figure 5.1: A General Framework for Modeling

Figure 5.1 shows the general framework for modeling. In this framework, independent elements are selected to form a common coordinate space. And form a process divided into three phases:
CHAPTER 5. PREDICTION METHOD

Predict, Match and Update [15].

- Model: the mathematical model is a description of the internal world by observation and then expression through mathematical concepts and language.

- Predict: In this phase, the current model is used to predict the external world.

- Match: The transformed observation with the predictions will be aggregated into a collection. The necessary condition is the expression of information from observation and the formation of a prediction which is qualitatively similar.

- Update: the observation integrates the prediction of the model to create an updated description of the environment composed of hypotheses.

Figure 5.2: A Framework for Modeling using Data Aggregation

Figure 5.2 shows the framework for our modeling. In this framework, the independent Eigenvector is selected and aggregated, the independent elements are location, online time, Tweet Type, Retweets, and Tweet Size, which are then fused into a single matrix Eigenmatrix. In order to avoid deviation due to a too large or too small number in the Eigenmatrix, the normalization of the Eigenvector is applied in common vocabulary through: \( \text{norm.Eigenvector} = \frac{\text{Vector}_{A}}{\text{Max.}(\text{Vector}_{A})} \), so that every Eigenvector is smaller than 1. Then the use of SVM or logistic regression is selected as a prediction model, through comparing their prediction outputs the input data can be modified to improve prediction accuracy.
5.2 Logistic Regression

Logistic regression is a classification model in machine learning, which was developed by statistician David Cox [18] in 1958. It is chiefly used for the analysis and prediction of a dichotomous outcome. This model is used for finding the relationship between a qualitative outcome variable and one or more predictor variables which have only two discrete possible types [28]. In this study the prediction outcomes are as ‘Mobile users’ and ‘Web users’.

5.2.1 Odds

‘Odds’ refers to the probability an event will happen compared to the probability that it will not happen. If \( p \) is a probability of an event happening, and \( 1 - p \) is the probability of an event not happening. Then the related:

\[
Odds = \frac{p}{1 - p}
\]  
(5.1)

5.2.2 Odds Ratio

The odds ratio (OR) is a value, which represents the measurement of two odds relative to different events. If there are A and B two related events, the corresponding odds of A happening relative to B happening is [29]

\[
Odds\ ratio = \frac{odds\{A\}}{odds\{B\}} = \frac{P_A/(1 - P_A)}{P_B/(1 - P_B)}
\]  
(5.2)

5.2.3 Logistic Model

The simple logistic model has the form:

\[
\ln\left(\frac{\pi}{1 - \pi}\right) = \log(\text{odds}) = \logit = \alpha + \beta \chi
\]  
(5.3)

Hence,

\[
\pi = \frac{e^{\alpha + \beta \chi}}{1 + e^{\alpha + \beta \chi}}
\]  
(5.4)

which is the probability of interested outcome, where

\[
\text{odds} = e^{\alpha + \beta \chi}
\]  
(5.5)
α is intercept parameter, β is a regression coefficient, χ is a predictor [28].

![Logistic curve, where $α=0, β=1$](image)

Figure 5.3: Logistic curve, where $α=0, β=1$

A continuous function can be used in the logistic regression to fit the data points, for the single data class, if the result takes the threshold value of 0.5 for tuned parameters for an accurate logistic curve for the data using the regression function, then it gives the result $R < 0.5$ for mobile users, and $R > 0.5$ for web users.

### 5.3 SVM

Support Vector Machine is a machine learning algorithm, and also a discriminative classifier [30], which is defined as a separating hyperplane. First the linear margin classifier will be discussed, which represents a method of determining optimal decision boundaries to separate the data. However, in real situations the data are non-linearly separable, so that we need kernel functions to transform SMVs efficiently into very high dimensional (even as infinite dimensional) feature spaces. Standard SVM training has $O(m^3)$ time and $O(m^2)$ [31] space complexities, where $m$ is the training set size.
5.3. SVM

5.3.1 Linearly-separable Data

Here a simple example is considered, a two class graph in Figure 5.4, where the data is linearly separable. There are many ways to form one straight line to separate the classes, and many different algorithms can be used to discover such a boundary, but which boundary is optimal?

5.3.2 Finding the Boundary

Let \( \{x_1, \ldots, x_n\} \) be data set and let \( y_i \in \{1, -1\} \)

We want to find the decision boundary, which can separate two classes. The equation of decision boundary is:

\[
W^T x_i + b = 0 \quad (5.6)
\]

The class, which above the decision boundary should be 1.

\[
i.e., W^T x_i + b > 0, \forall y_i = 1 \quad (5.7)
\]

The class, which below the decision boundary should be -1.

\[
i.e., W^T x_i + b > 0, \forall y_i = -1 \quad (5.8)
\]
or in one equation The class, which below the decision boundary should be -1.

\[ i.e., y_i(W^T x_i + b) \geq 0, \forall i \]  (5.9)

The primal problem of the SVM can be formulated:

\[ \text{Minimize} : \frac{1}{2}||w||^2 \]  (5.10)

Subject to:

\[ y_i(W^T x_i + b) \geq 0, \forall i \]  (5.11)

The margin should be maximized, and the decision boundary should be as far away from the data of both classes as possible. The points that lie on the separating line (hyperplane) are called support vectors, the solution to the problem is determined by these vectors.

5.3.3 Nonlinear Decision Boundary

In many cases, it is not possible to find a separating hyperplane in the data. In that case, we need to change the formalism. The key idea is to transform \( x_i \) to a higher dimension in order to separate the data.
First slack variables must be added into the standard formulations:

\[(W^T x_i + b) \geq 1 - \varepsilon_i, \quad y_i = 1, \quad \varepsilon_i \geq 0, \quad \forall i\]  
(5.12)

\[(W^T x_i + b) \leq -1 + \varepsilon_i, \quad y_i = -1, \quad \varepsilon_i \geq 0, \quad \forall i\]  
(5.13)

Now we need to choose the input space, so \(x_i\) is where the data are located.

And we want to have a feature space after transformation: \(\Phi(x_i)\). Where \(\phi(x_i)\) is a feature mapping [32].

### 5.3.4 Kernel Trick and Kernel Function

The transformation from input feature into the feature space is defined as well as the “Kernel Trick”, so that the non-linear separable feature can be linear separable classified again.

Let’s start with an example; if there non-linear mappings function:

\[\Phi : I = \mathbb{R}^2 \rightarrow F \rightarrow \mathbb{R}^3\]  
(5.14)

From the two-dimensional input space \(I\) into the three-dimensional feature space \(F\):

\[\Phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T\]  
(5.15)

If we account a separating hyperplane, then we get a \(\mathbb{R}^3\):

\[w^T \Phi(x) = w_1x_1^2 + w_2\sqrt{2}x_1x_2 + w_3x_2^2 = 0\]  
(5.16)
By classifying the data point in $I$ through a feature map into higher dimensional space $F$, the separating hyperplane can be easily found after transforming the non-linear separable data in low dimensional space into a linear separable higher dimensional space [33].

So the SVM optimization Problem is:

$$
\text{Max. } \sum_{i} a_i - \frac{1}{2} \sum_{jk} a_i a_j y_i y_j x_j^T x_k \to \forall a_i \geq 0
$$

(5.17)

$$
\text{Max. } \sum_{i} a_i - \frac{1}{2} \sum_{jk} a_i a_j y_i y_j \Phi(x_j)^T \Phi(x_k) \to \forall a_i \geq 0
$$

(5.18)

This $\Phi(x_i)$ is feature mapping which can be very high dimensional, so it can be highly time consuming to explicitly compute it. The Kernel Trick can be applied to linear models in which dual formulations appear as dot products such as SVM, PCA, etc [34].

The Kernel Trick can replace these dot products with an equivalent kernel function:

$$
k(x, x') = \Phi(x)\Phi(x')
$$

(5.19)

A Kernel Function is a function:

$$
\mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}
$$

(5.20)

The kernel functions enable the dot product to operate in feature high-dimensional space. If features are mapped to a high dimensional feature space, the linearly non-separable features can be linearly separable. But it is not necessary to calculate all the feature mappings, because computation of kernel is much easier than computation of feature mapping. Consequently, building a complex decision boundary and high dimensional based computation feature mapping with Kernel Function will be very efficient [35].

Define the Kernel Function:

$$
K(x_i x_j) = \Phi(x_j)^T \Phi(x_k)
$$

(5.21)

The Eigen functions determine the transformation, but the feature of Eigen functions can be difficult to compute explicitly. The time complexity taken to compute the objective is $O(n^2)$, which is based on the size of the dataset.

The inside of the dataset data distribution is unknown, but it is certain that mobile and web users are certainly not linearly separable. For that reason, a simple linear function is not suitable for this dataset when compared to polynomial kernel.
5.3. SVM

5.3.5 Gaussian Kernel

When using the Gaussian kernel function, the advantages of Gaussian kernel is that the data points can be separated in any distribution, which can then be switched into two-dimensional, three-dimensional, up to the infinite dimensions. Gaussian kernels are universal kernels, it contains not only a linear kernel, but also subsumes polynomial kernels as well.

The Gaussian kernel defined by:

\[
k(x, x') = \exp(-\gamma ||x - x'||^2)
\]  

where \(\gamma = \frac{1}{2\delta^2}\)

The \(\delta\) is the kernel width of Gaussian which controlled by a parameter \(\gamma > 0\). If we decrease value of parameter, the curvature of decision boundary is also decreased. If we increase the parameter, the curvature of decision boundary is also increased and this is leaded to accommodate the larger penalty for errors/margin errors [36].

![Figure 5.7: The effect of the inverse-width parameter of the Gaussian kernel (\(\gamma\)) for a fixed value of the soft-margin constant [36].](image)

Similar decision boundaries can be obtained using different combinations of SVM hyperparameters.
The Gaussian kernel is a decay function which calculates using a support vector, attained at the support vector and which decays consistently in all directions around the support vector, leading to hyper-spherical contours of the kernel function.
Chapter 6

User Type Prediction

The first two methods, which use single Eigenvector [37] to predict the users’ type, have each selected one from the five Eigenvectors, which means in each case 100 of the whole 500 groups, where each group includes 20,000 rows of single Eigenvectors are selected. The following two prediction methods allocate 20,000 rows each from the selected 1,000 groups. For the validation, 1,000 rows were randomly selected from the dataset without data fusion for single Logistic and SVM test, and the other 1,000 rows with five Eigenvectors are for Logistic and SVM test.

For the logistic regression and SVM, Matlab 2015b [38] was used on 2 x Intel Xeon® 2620, 12-core computer with 64GB of RAM, for the SVM is selected Gaussian Kernel, Box Constraint Level is 1, Kernel Scale Mode set to Auto.

The accuracy computing of logistic regression is as following code:

```
%%
load('test_data.mat');
%%
pathname = uigetdir;
filelist = dir(fullfile(pathname));
filelist = {filelist(~[filelist.isdir]).name};
for iii = 1:length(filelist)
    import_data = importJavaData(strcat(pathname,'\',filelist{1,iii}));

    variable_input = import_data(:,1:5);
    decision_output = import_data(:,6);

    temp_1 = find(decision_output == 1);
    temp_m1 = find(decision_output == -1);

    decision_output(temp_1) = 2;
    decision_output(temp_m1) = 1;
```
Logistic_Model = mnrfit(variable_input, decision_output);

prediction_input = test_data(:,1:5);
prediction_output = test_data(:,6);
Accu = mnrval(Logistic_Model,prediction_input);

abweichung = Accu(:,2) - Accu(:,1);
temp_web = find(abweichung > 0);
temp_mobile = find(abweichung <= 0);
abweichung(temp_web) = 1;
abweichung(temp_mobile) = -1;
Accu_real(iii) = 1 - length(find(abs(abweichung - prediction_output)~=0))/size(test_data,1);
disp(iii);
disp(Accu_real(iii));
end

The SVM parameter setting is as following code:

function [trainedClassifier, validationAccuracy] = SVM_Single_Parameter_trainClassifier(datasetTable)
    % Convert input to table
    datasetTable = table(datasetTable);
datasetTable.Properties.VariableNames = {'column'};
    % Split matrices in the input table into vectors
    datasetTable.column_1 = datasetTable.column(:,1);
datasetTable.column_2 = datasetTable.column(:,2);
datasetTable.column = [];
    % Extract predictors and response
    predictorNames = {'column_1'};
predictors = datasetTable(:,predictorNames);
predictors = table2array(varfun(@double, predictors));
response = datasetTable.column_2;
    % Train a classifier
    trainedClassifier = fitcsvm(predictors, response, 'KernelFunction', 'gaussian', '
        PolynomialOrder', [1], 'KernelScale', 'auto', 'BoxConstraint', 1, 'Standardize', 1, '
        PredictorNames', {'column_1'}, 'ResponseName', 'column_2', 'ClassNames', [-1 1]);
    % Perform cross-validation
    partitionedModel = crossval(trainedClassifier, 'KFold', 5);
    % Compute validation accuracy
    validationAccuracy = 1 - kfoldLoss(partitionedModel, 'LossFun', 'ClassifError');
First the easiest and the most efficient way to predict the users’ type will be to consider the input matrix as having only one single feature row.

6.1 Logistic Regression with Single Feature

Figure 6.1: Logistic Regression with Single Feature
The Tweet sizes for both users are similar in that from 1 to 20 both curves increase in the same trend. Later when the mobile users’ distribution has decreased, because they do not like to send longer tweets, the web users’ distribution stays on the same percentage almost until the end. The minimum value and the maximum value are 71.4% and 79.7%, the prediction median value of this feature is 74.95%, which is below the average value of 75.03%, which means these points are not normal Gaussian Distribution. The Probability-Frequency distribution shows that the probability is not evenly distributed, there are missing parts at the end of both functions, and the middle distribution is too great. Through this phenomenon we can assume that there are some deviations in the results.

The significant difference between mobile and web users’ Online Time is from 15:00 to 22:00, in this time period the web users send more Tweets than mobile users, in the other time period, both users’ frequencies remain almost at the same activity level. The minimum value is 67.4%, the maximum is 75.1%, and the median is 71.15%, which is the lowest result when using logistic regression. One reason shown in the Probability-Frequency Figure, for why the centre distribution does not follow the Gaussian Distribution curve, is that there should be more probability distribution located in this area.

There is nearly no overlap between the two users’ Retweets behavior in Figure 3.2, the mobile users’ distribution is located substantially on the same level. The web users’ behavior is similar to the Gaussian Distribution, in that the greatest quantity of Retweets’ is located in the middle range of the curve. However, the result is only 71%. One reason for this is that we have used the standardized curve, the mobile users’ Retweets behavior is on a small distribution range number. In fact, both curve trends are still similar, but just not the same percentage distribution. Overall, the Probability-Frequency distribution nearly fits the Gaussian Distribution, that means the final median can be represented in the final results.

Through location observation the 71.25% prediction result has been selected, because most users are from the USA, the UK, and Japan. The Intuitive observation probability distribution histogram is also very close to the Gaussian distribution, which The intervals are usually equal size, and must be adjacent \[39\].
Although the majority of both user groups are using the social realm of Twitter, we can use other elements to distinguish the users’ type, web users prefer using Twitter to share their applications, or mobile users like to use twitter client to send or share Tweets. The prediction result for this is 74% which is the second best result in using logistic regression.

<table>
<thead>
<tr>
<th>Eigenvector</th>
<th>Avg.Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets Size</td>
<td>74.95</td>
</tr>
<tr>
<td>Online Time</td>
<td>71.15</td>
</tr>
<tr>
<td>Retweets</td>
<td>71.00</td>
</tr>
<tr>
<td>Location</td>
<td>71.25</td>
</tr>
<tr>
<td>Tweets Type</td>
<td>74.00</td>
</tr>
<tr>
<td>Avg. Compute Time</td>
<td>40 Seconds</td>
</tr>
</tbody>
</table>

Table 6.2: Logistic Regression with Single Feature

Overall, the results are not too different from the five individual features when using logistic regression to predict the users’ type. The difference between the highest accuracy and the lowest accuracy is less than 4%, which is a very small range. Each probability distribution in the histogram appears quite dissimilar. However, the median line does not appear in the middle of the distribution, from this one can presume the results from using logistic regression are not ideal.

6.2 SVM with Single Feature
A large number of diverse data sets with a greater variability of features is a benefit for making predictions, here the Tweets Size own the large variable interval distribution, here, 0 to 140, and the mobile users prefer to send short tweets compared to the web users, consequently through this feature it can be observed that the maximum accuracy is 79.16%, the minimum accuracy is 77.68%, and the median value is 78.48%, which is better than using logistic regression. In general, the middle part should be relatively high probability located, compared to the Tweets Size histogram result, although there is a small part missing in the high frequency part.

In contrast, Online Time is almost uniformly distributed, with the prediction result: 65.9%, is not satisfactory compared to the 71.00% when using logistic regression. There are missing parts at the
end of the histogram and a greater probability than expected. Therefore, this logistic regression is more applicable to this feature than using SVM.

And the other two outcomes’ accuracy such as Retweets and Location are similar and acceptable, in the other words, the median value of the both features are the same: 70.23%, which is almost the same compared to the result in using logistic regression. Although the users’ behaviors in Retweets looks significantly different in the standardized percentage graph, it is mobile users’ Retweet values which were always at a lower percentage than the web users’, which was at a high, unstable percentage.

Overall, the predicted results of using the individual features are not particularly good. One advantage is that using single features to modeling is actually quite efficient; the average computing time is also in an optimal range, hence, if the feature set is large enough and suitably dissimilar, this method can be recommended.

Because the Tweets Type is specially distributed, the two class’ users concentrate in dissimilar functions, and the prediction result is the second best: 75.44% which is better than using logistic regression.

<table>
<thead>
<tr>
<th>Eigenvector</th>
<th>Avg. Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tweets Size</td>
<td>78.48%</td>
</tr>
<tr>
<td>Online Time</td>
<td>65.93%</td>
</tr>
<tr>
<td>Retweets</td>
<td>70.23%</td>
</tr>
<tr>
<td>Location</td>
<td>70.23%</td>
</tr>
<tr>
<td>Tweets Type</td>
<td>75.44%</td>
</tr>
<tr>
<td>Avg. Compute Time</td>
<td>2 Min.</td>
</tr>
</tbody>
</table>

Table 6.4: SVM with single Future

### 6.3 Logistic Regression with Data Fusion

Let’s look at the left side of Figure 6.3, a “white noise distribution” which can be represented as Gaussian distribution [40]. In our case, it is a predicted description between the maximum value and the minimum value, we have also chosen a yellow line as the median line, which refers to 73.22%. It can be seen very clearly that the distribution above and below the median value is not equal. The distribution above the median is very sparse. The difference between the maximum value and the minimum value is 1.96%. The distribution below the median is quite close to this value, while the difference of the absolute value between the minimum and the median values is
CHAPTER 6. USER TYPE PREDICTION

Figure 6.3: Logistic Regression with Data Fusion

0.93%. In other words, the result from using logistic regression with data fusion has not improved the accuracy when compared to the logistic regression with the single unique feature.

Right side of Figure 6.3 shows how the histogram graph fits in with this distribution, which is the result of using logistic regression with data fusion. The X axis represents the accuracy and the Y axis the frequency of accuracy. Here the average accuracy remains on 73.2%, the minimum accuracy is 72.18%, and the maximum accuracy is 75.07%. The distribution of the results can be understood as a Gaussian distribution, but unlike normal Gaussian distribution, these accuracies are not continuously smooth, in other words, the head and tail of the function are too small and too discrete, but the center distribution display is unexpectedly high, so this distribution is similar to the positive skew, where the left part is shorter and contains the greatest concentration of the distribution in this part. In our graph, the greatest distributions are located between 72% and 73.5%, the fit line shows that the greatest distribution in normal distribution graph should be allocated between 72% and 74%.

| Avg. Prediction Result | 73.22% |
| Avg. Compute Time      | 1 Min. |

Table 6.5: Logistic Regression with Data Fusion

6.4 SVM with Data Fusion

Left side of Figure 6.4 shows a white noise distribution, the median value is located almost in the middle of the figure; the average prediction value is 81.8%, which is highly similar to the median of the value of 81.7%. It means that the greatest probability is located around the median value.
For that reason, the final prediction result can be considered to be that the deviation exists in a very minimal range.

Right side of Figure 6.4 presents the result of using Gaussian kernel histogram graph to fit in with the distribution, it is obvious that the prediction result is extremely similar to the Gaussian Distribution, which has a minimum value is 80.08% and the maximum value is 83.44%, and the other results are very smoothly gathered in the center, with barely predicted results are outside of the curve.

The result of using data fusion is more accurate than using logistic regression or SVM with single unique feature, and is also more accurate than using logistic regression with data fusion. This indicates that accuracy is improved when using SVM with data fusion for the prediction.

<table>
<thead>
<tr>
<th>Avg. Prediction Result</th>
<th>81.75%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Compute Time</td>
<td>4 Min.</td>
</tr>
</tbody>
</table>

Table 6.6: SVM with Data Fusion
Chapter 7

Conclusion and Future Work

7.1 User Behaviors

This section lists several behaviors of both mobile phone and computer users. Most behavioural studies show the way that the related users’ behaviors have changed in different time periods. The emphasis of analyzing a graph of such behavior is to define and standardize all the users’ of that time zone in order to pay it proper attention and locate deviations that exist. The difficulty is in defining the class of the users’ device when no such lists exist in the dataset and the 7,000 values must be manually filtered into two classes: mobile and web. Therefore, the final result is that the geographical time difference and device have been exactly classified.

7.1.1 Tweet Size Distribution

The difference between the two users’ Tweets Size shows obvious differences when analyzed. Because web users send Tweets without geographic restriction and data volume limit, the web users’ Tweets Size has almost occupied the whole length range, which distribute a nearly smooth linear line between 20 to 130 characters. On the other hand, mobile users focus more on writing their Tweets efficiently. The majority of users choose to send short Tweets of around 10 to 60 characters, thus, their Tweets Size distribution is more of a Positive Skew, with a lot of the distribution concentrated on the left side of the figure.

7.1.2 Tweet Interval

From different Tweets Interval figures it is possible to find out that the majority of users’ behaviors are very similar, the time interval between first and the second Tweets is more or less two hours. One reason is that these users are doing multiple social activities during their discrete online time,
such as in the early morning, 14:00 to 16:00, or in the evening, which are the maximum or the minimum percentage of these users. After 4 hours there is almost no activity and that means that both curves are infinitely close to the x axis, in other words, close to 0.

7.1.3 Online Activity

The most significant difference between the two classes of users in Online Activity refers to two time periods: 14:00 to 16:00, and 20:00 to 23:00. During these two intervals the web users' activity increases rapidly and then reach their peak. On the other hand, the mobile users' curve is similar to the web users at the beginning of 10:00 o'clock, and then the Online Activity remains at a stable level, although there are some changes during and after work time.

7.1.4 Retweets

Retweet activity for both users are also similar, although both curves are not overlapping in Figure 3.2. One reason for this is that the distribution occupies a wide range of the web percentage when compared to the mobile users. However, if the mobile users are individually observed and zoomed in on without comparison to web users, the curve is distributed as a Negative Skew, where the majority of probability is focusing on the right side, much like web users.

7.1.5 Tweet Type Distribution

Since Twitter is a social based application, the greatest section of the Tweets Type is the social function which is: sending Tweets, sharing Tweets, or Retweets and so on. Certainly through its basic functions and the explosion of information on Twitter, the user can obtain some interesting information, such as Data Analyze, Twitter Management, and Twitter client of mobile users; or Blog, Application, and Media of web users. In general, the mobile users usually focus on how to connect themselves best with others; the web users are concentrated on their own interest areas.

7.2 Prediction

The difficulty in predicting the users' device lies in selecting the best sampling data since it is impossible to use the full dataset to create a prediction model due to the limit of physical memory and the high computing complexity. This section lists the result of two prediction methods, which have used two different sampling methods.
7.2. PREDICTION

7.2.1 Logistic Regression

By comparing the results of the previous chapter it is not difficult to see that the two results from using individual feature data fusion are closely similar. However, some individual results are more accurate in some special features rather than in data fusion. The logistic regression model is based on a set of independent features, which is a feature matrix, in order to predict the outcomes. This is also a weakness of this model, if the feature matrix includes some no prognostic values, no relative values, or some wrong values, the model will fail. It means that the logistic regression model is only suitable for identified relevantly independent features. There are some low weight features such as Online Time, Retweets, and Location, when these features are collected into a data fusion matrix to create logistic regression model, the prediction results will be affected by these low weight features and not predict accurately. This is a disadvantage, because in this research field the multiple observations involve the same independent features, which determine the single unique Tweet. If observations are not independent compared to other data points, meaning the single data point becomes related to another data point, then the logistic regression model might calculate false observations.

In this prediction situation, Logistic regression is used to predict through which device a user generated Tweet is sent, hence the outcome is binary. However, it is impossible to use logistic regression to predict continuous outcomes. For example, logistic regression could not be used to determine at what time the users sent tweets or which length Tweet users would send at 15:00. Because these predictions are distributed continuously, it is meaningless to convert it into discrete categories like “In the morning or in the afternoon”, or even into “no characters or full characters”. This is a significant disadvantage of predicting with continuous outcomes.

When logistic regression uses data fusion, the result has not been improved upon compared to the result of some individual independent features, but in terms of the time needed to create the model, use training data, and compute the final results it is actually quite efficient. In general, if features are discretely distributed and are not related to each other or overlapping, the prediction results will not be on a continuous scale, and the logistic regression model is an alternative time efficient option for the prediction.

7.2.2 SVM

The final result formed using SVM with data fusion is intuitively better than logistic regression, SVM can create an accurate model and produce robust classification, whether the input features are linearly or non-linearly separable. After using data fusion the accuracy is obviously improved to 81.75%, although the prediction results of individual features remain almost on a consistent level: about 70%.

Unlike logistic regression, which can only find out one decision boundary, one of the advantages
of using the SVM is that it can also identify the non-linear decision boundaries. Thus the SVM is regarded as an effective tool for more complex information predictions.

Another advantage is that the SVM can randomly classify the dataset through different kernel functions. There are some SVMs which can be regarded as an alternative option for separating the dataset, for example Linear SVM, Quadratic SVM and Cubic SVM. However, the selected Gaussian Kernel Function for predicting the users’ device is more operational, the \( \gamma \) value can be manually defined when the dataset distributes randomly and discretely and/or the one prediction target is surrounded by another prediction target; in that situation, the Linear Regression is no longer efficient, but the Gaussian Kernel can precisely classify their boundary when the right \( \gamma \) value is selected.

Otherwise, if the \( \gamma \) value is too small, the boundary, which represented as a straight line like Linear Regression, will not correctly separate the dataset, and if the \( \gamma \) becomes a huge number, the separated class will be shrunk by the decision of the boundaries, in an extreme case, each point is defined as a class which is also as a decision boundary, which is known as overfitting. This will of course harm the predictions.

Another consideration is the complexity, when SVM is \( O(n^2) \) which is then compared with the logistic regression’s \( O(n) \), if the dataset includes a huge quantity of metadata, for example, one billion lines’ of information, it is almost unnecessary to compute the full data bank for modeling because it is so time-consuming. So in this case, an effective sample for modeling, training, and prediction should be considered, which requires an effective sampling method, sampling size, and sampling data structure.

### 7.3 Future Work

Future work on social network should attempt to find out the correlation between independent users to precisely separate the network into related aggregation groups. By identifying these small aggregation groups or specific groups of people further work can attempt to compare the individual’s behavior. There are further potential areas of research focusing on subtle behavior differences, for which these results will provide a robust diversification of sample data for further modeling and validation. The sampling of metadata should be attempted using more methods to ensure that the results become more comparable and accurate. The prediction for the non-linear dataset should be an accuracy comparison of computing the same data feature with different parameters in different Kernel Functions, although such permutations are a relatively large magnitude.
List of Tables

3.1 User Data File Structure ................................................. 9
3.2 Profile Structure ......................................................... 11
3.3 Tweet Structure .......................................................... 11
4.1 Feature Structure ......................................................... 36
6.1 Eigenvector ................................................................. 50
6.2 Logistic Regression with Single Feature .......................... 51
6.3 Eigenvector ................................................................. 52
6.4 SVM with single Future .................................................. 53
6.5 Logistic Regression with Data Fusion ............................... 54
6.6 SVM with Data Fusion .................................................... 55
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>User location distribution</td>
<td>10</td>
</tr>
<tr>
<td>3.2</td>
<td>Normal Users’ Behavior</td>
<td>12</td>
</tr>
<tr>
<td>3.3</td>
<td>User Activity in Christmas Day and Easter Sunday</td>
<td>14</td>
</tr>
<tr>
<td>3.4</td>
<td>Mobile User Tweets’ Type Distribution</td>
<td>15</td>
</tr>
<tr>
<td>3.5</td>
<td>Mobile User Tweets’ Type Distribution</td>
<td>16</td>
</tr>
<tr>
<td>3.6</td>
<td>Heat Map Zoom In of User Tweets Type Distribution</td>
<td>16</td>
</tr>
<tr>
<td>3.7</td>
<td>Users’ Behavior Who Have the Most Followers</td>
<td>17</td>
</tr>
<tr>
<td>3.8</td>
<td>Most followed Mobile User Tweets Type Distribution</td>
<td>18</td>
</tr>
<tr>
<td>3.9</td>
<td>Most followed Web User Tweets Type Distribution</td>
<td>19</td>
</tr>
<tr>
<td>3.10</td>
<td>User’ Tweets Type Distribution Heat Map Zoom In</td>
<td>20</td>
</tr>
<tr>
<td>3.11</td>
<td>Behavior of Users Who Tweets the Most</td>
<td>21</td>
</tr>
<tr>
<td>3.12</td>
<td>Mobile User Tweets Type Distribution</td>
<td>22</td>
</tr>
<tr>
<td>3.13</td>
<td>Web User Tweets Type Distribution</td>
<td>23</td>
</tr>
<tr>
<td>3.14</td>
<td>User’ Tweets Type Distribution Heat Map Zoom In</td>
<td>24</td>
</tr>
<tr>
<td>3.15</td>
<td>Behavior of Japan Users</td>
<td>25</td>
</tr>
<tr>
<td>3.16</td>
<td>Japan User’ Tweets Type Distribution Heat Map Zoom In</td>
<td>27</td>
</tr>
<tr>
<td>3.17</td>
<td>Behavior of UK Users</td>
<td>28</td>
</tr>
<tr>
<td>3.18</td>
<td>UK User’ Tweets Type Distribution Heat Map Zoom In</td>
<td>29</td>
</tr>
<tr>
<td>3.19</td>
<td>Behavior of USA Users</td>
<td>30</td>
</tr>
<tr>
<td>3.20</td>
<td>USA User’ Tweets Type Distribution Heat Map Zoom In</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>Time Location Distribution</td>
<td>34</td>
</tr>
<tr>
<td>5.2</td>
<td>A Framework for Modeling using Data Aggregation</td>
<td>38</td>
</tr>
<tr>
<td>5.3</td>
<td>Logistic curve, where $\alpha=0, \beta=1$</td>
<td>40</td>
</tr>
<tr>
<td>5.4</td>
<td>Linearly-separable Data</td>
<td>41</td>
</tr>
<tr>
<td>5.5</td>
<td>Finding the Boundary</td>
<td>42</td>
</tr>
<tr>
<td>5.6</td>
<td>Feature map</td>
<td>43</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.7</td>
<td>The effect of the inverse-width parameter of the Gaussian kernel ($\gamma$) for a fixed value of the soft-margin constant [36].</td>
<td>45</td>
</tr>
<tr>
<td>6.1</td>
<td>Logistic Regression with Single Feature</td>
<td>49</td>
</tr>
<tr>
<td>6.2</td>
<td>SVM with Single Feature</td>
<td>52</td>
</tr>
<tr>
<td>6.3</td>
<td>Logistic Regression with Data Fusion</td>
<td>54</td>
</tr>
<tr>
<td>6.4</td>
<td>Using SVM with Data Fusion</td>
<td>55</td>
</tr>
</tbody>
</table>
Bibliography


