A Two-Phase Model for Retweet Number Prediction

Gang Liu[1], Chuan Shi[1,*], Qing Chen[2], Bin Wu[1], and Jiayin Qi[1]

[1] Beijing University of Posts and Telecommunications, Beijing, China
{liugang519,shichuan,wubin}@bupt.edu.cn, qijiayin@139.com
[2] China Mobile Communications Corporation
chenqing@chinamobile.com

Abstract. With the surge of social media, micro-blog has become a popular information share tool, in which retweeting is a basic way to share and spread information. It is important to predict the retweet number for influence measure and precision market. Contemporary methods usually consider it as a classification or regression problem directly, which can be regarded as one-phase models. However, they cannot accurately predict the number of retweet. In this paper, we propose a two-phase model to predict how many times a tweet can be retweeted in Sina Weibo. That is, the model first classifies tweets into several categories, and then does regression on each category. Extensive experiments on real Sina Weibo dataset show that our model is a general framework to achieve better performances than traditional one-phase prediction model without complex feature extraction.

Keywords: Social media, Sina Weibo, Retweet, Classification, Regression.

1 Introduction

Recently, there is a surge of social media. Many social network services have emerged, among which micro-blog service is a platform of sharing, spreading and acquiring message based on users' relationship. People can post messages of up to 140 characters through Web or smart phones to share information timely. Micro-blog services gain worldwide popularity. As the most popular micro-blog service, Twitter [1] had about 500 million registered users in 2012 and these users post about 340 million messages every day. In China, Sina Weibo [2] had 503 million registered users before March 2013.

In micro-blog network, retweet is the main way to spread messages. When a user posts a message, this message will be pushed to the user's followers. When followers see this message, he/she can choose to retweet the message, so the message will be pushed to his/her followers. By retweet, messages can be continued to spread in the micro-blog network. Therefore, the times of retweet (i.e., retweet number) can be as an important indicator of the message's influence. Predicting the retweet number of a message in micro-blog network (i.e., tweet) has practical significance in evaluating

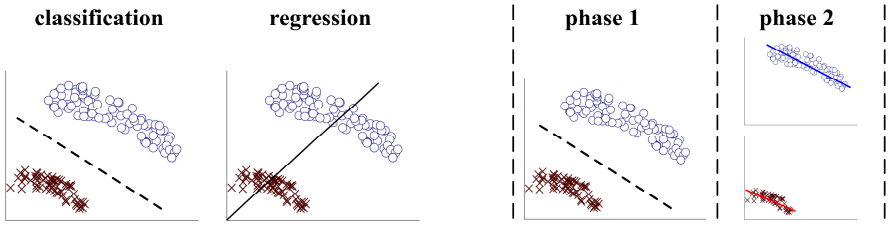
* Corresponding author.

F. Li et al. (Eds.): WAIM 2014, LNCS 8485, pp. 781–792, 2014.

the influence and the value of a tweet. What's more, it contributes to controlling the spread of illegal information like rumors.

There has been some studies [3,4,5,6,7,8,9,10] on the retweet prediction in micro-blog network. Many of them consider the problem as a two-classification problem, which predicts whether a message will be retweeted or not. Some studies [9] also treat it as a multi-classification problem. However, it is difficult to determine the threshold of multiple classifiers. There are a few works predicting the retweet number directly [10]. All these work can be considered as one-phase model as shown in Fig. 1(a). Due to the complexity of retweet behavior, it is hard to accurately predict the retweet number with these one-phase models. Moreover, most of work focuses on English micro-blog services like Twitter and few studies are on Chinese micro-blog services.

In this paper, we first analyze the characteristics of retweet in real Sina weibo dataset and point out it is not rational to directly do regression on training data due to the power law distribution of retweet number. Then we propose a two-phase model to predict retweet number. Fig. 1(b) illustrates the basic idea of our model. In the first phase, the model classifies tweets into one of multiple categories, where the classification threshold can be automatically determined by the 80/20 rule. In the second phase, the model does regression on each category to predict the retweet number. The two-phase model has the following advantages: (1) It is a general framework, which can employ any classifier or regression model in it; (2) It is a simple but effective method without complex feature extraction. Extensive experiments on Sina Weibo data validate the above benefits through achieving better performances than one-phase models under different model settings.



(a) One-phase model (b) Two-phase model

Fig. 1. Different methods on retweeting behavior

This paper is organized as follows. In Section 2, we describe related work about micro-blog network and retweet behavior analysis. In Section 3, we describe our dataset and analyze the characteristics of retweet behavior. Section 4 illustrates our two-phase model for retweet number prediction. Section 5 provides numbers of experiments to validate our model. Finally, we summarize our findings and conclude with Section 6.

2 Related Work

Since micro-blog has become a popular social network service, a lot of studies have been done to explore its traits. Some studies are about micro-blog network structure and user characteristics. Java et al. [11] introduce the basic functions and features of Twitter and give a preliminary analysis of its social networking features, the results indicate that Twitter showed some power law distribution and small world properties.

As retweet behavior is the key mechanism for information diffusion in micro-blog network, many studies focus on retweet behavior. Boyd et al. [12] took a detailed analysis of retweet in Twitter to explore how and why people retweet a tweet. Suh et al. [13] examined a number of features that might affect retweet, and they found that the number of followers and followees seem to affect retweet behavior, while the number of past tweets does not affect a tweet to be retweeted. Zhang Yang et al. [8] analyzed the importance of different features and investigate the feasibility of applying classification method and proposed a feature-weighted model. Their model can predict a major fraction of tweets(nearly 86%). Bandari et al. [9] proposed a model to predict popularity of new articles. They classified articles to three classes based on their retweet number 1-20, 20-100 and 100-2400. The model can predict ranges of popularity on Twitter with an overall 84% accuracy. Most of studies above focus on Twitter and few studies are on direct retweet number.

There are also a few work on retweet number prediction of Chinese micro-blog. Li Ying-le et al. [10] proposed a prediction model based on SVM algorithm with five features: user influence, user activity, interest similarity, the importance of micro-blog content and users' closeness. The experiment with Sina Weibo data shows a good result that the predict accuracy is up to 86.63%. However, the features they extracted are very complex and expensive and thus the model is not suitable for large-scale data.

3 Data and Features

This section describes the dataset and features extracted from tweets. Then we analyze the characteristic of retweet behavior.

3.1 Dataset Preprocess

We use Sina Weibo API to collect tweets for three months from April 2013 to July 2013 and finally get 54M tweets and 142K different users. In real micro-blog network, retweet number of a tweet will change with time. For example, when a tweet is just posted, the retweet number may be 0. After an hour, the retweet number may increase. Because we want to build a prediction model to predict the final retweet number of a tweet, we filtered data and got 49M tweets which exist in micro-blog network more than 30 days. Since retweet number of a tweet will trend to be stable with the time passed by, we consider that the retweet number will be stable after 30 days.

3.2 Feature Description

We extracted 28 features from the tweet and the tweet's creator. Most of them can be directly crawled from Sina Weibo API, and these features have been proved to be effective in some papers [8],[13]. Moreover, most of the features are basic information from dataset and they don't require complex computation. Table 1 and Table 2 show the details. In Table 1, we also compute 4 features which describe the tweet's creator's influence.

Table 1. Feature about the tweet creator

Feature	Explanation	Feature	Explanation
GD	Gender of the tweet's creator	VR	The tweet's creator is a verified user
NL	The length of the nickname of the tweet's creator	VT	The verified type of the tweet's creator
FON	The number of the followers who follow tweet's creator	ED	The number of days since the tweet's creator registered
FRN	The number of the friends who are followed by tweet's creator	MSPD	SN/ED
BFN	The number of the friends who and tweet's creator follow each other	MFPD	FON/ED
FAN	The number of the favorites which the tweet's creator has	MFPS	FON/SN
SN	The number of tweets of tweet's creator post	MAFPS	(FON-FRN)/SN

Table 2. Feature about the tweet

Feature	Explanation	Feature	Explanation
HI	The tweet has hashtag in text	TL	The length of the text
HC	The number of the hashtag in text	TM	The month of the tweet which was created
AI	The tweet has @ in text	TD	The day of the tweet which was created
AC	The number of @ in text	TH	The hour of the tweet which was created
HI	The tweet has link in text	TW	The week of the tweet which was created
HC	The number of link in text	HOI	The tweet was created on holiday
PI	The tweet has pictures	EH	The hours since the tweet was created

3.3 Characteristic of Retweet Behavior

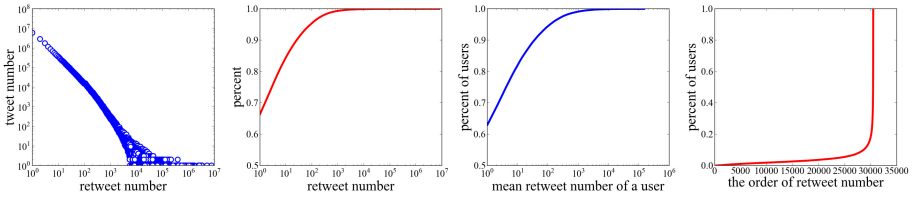
We take a basic statistics analysis on retweet behavior of users.

For retweet number, we count tweets of different retweet number and calculate its cumulative distribution. Fig. 2(a) shows the log distribution of total tweets over all data, demonstrating a long tail shape. We can see the cumulative distribution of different retweet number in Fig. 2(b). Apparently, the retweet number is extremely unbalanced.

For example, 66.4% of all tweets' retweet number is less than 1 and 90% of that is no more than 30. The tweets that retweet number more than 1000 is less than 1%.

For users, we compute the mean retweet number of each user (i.e., MRN). Fig. 2(c) shows the cumulative distribution of different mean retweet number of users. From Fig. 2(c), we can see that 63.1% of all users 's mean retweet number is less than 1 and users with retweet number less than 30 is 90%. While, less than 0.8% users' mean retweet number is more than 1000. Fig. 2(d) is also a cumulative distribution. We sort the mean retweet number according to descending order. From Fig. 2(d), the top 10% tweets come from 0.78% of users. 7.8% of all users post top 90% tweets. Obviously, only a small part of users can post tweets with high retweet number.

In all, we can get two observations: (1) In the micro-blog network, the retweet number and number of tweets comply with the power-law distribution, and only a small part of tweets' retweet number is high. (2) For users in micro-blog network, only a small percentage of them have the potential to post a tweet with high retweet number.



(a) Distribution of retweet number (b) CDF of retweet number (c) CDF of MRN (d) CDF of the top order retweet number

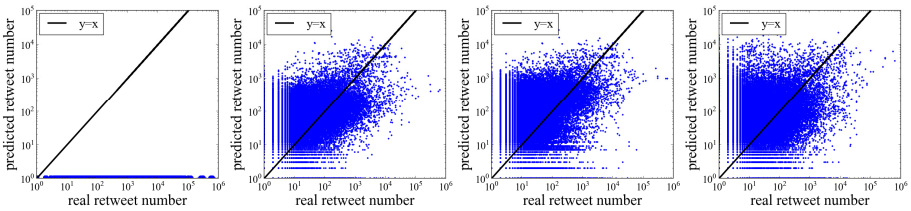
Fig. 2. The characteristic of retweet behavior

4 Two-Phase Prediction Model

In this Section, we first discuss the disadvantage of one-phase models to predict retweet number, and then describe our two-phase retweet number prediction model.

4.1 One-Phase Prediction Model

In order to predict the retweeting number, a direct solution is to do regression on training data including features and its retweet number. This solution is called one-phase prediction model in this paper. Here, we select four one-phase regression models: LeastMedSq[14], LinearRegression, M5P[15] and MultilayerPerceptron to predict retweet number. We random select 1 million tweets from the whole dataset, because our dataset is too large to train a model in time. 80% of the data is training set and the rest is test set. We draw the prediction result of four one-phase models in Fig. 3, which shows the relation of prediction values and real values.



(a) LeastMedSq (b) LinearRegression (c) M5P (d) MultilayerPerceptron

Fig. 3. One-phase prediction model scatter diagram

In Fig. 3(a), we can see LeastMedSq model predicts all tweet with a retweet number 0 or 1. Because most of tweets in training set have retweet number less than 1. Apparently, this model cannot solve the prediction problem. In Fig. 3(b)(c)(d), the prediction results are chaos. Even for most of tweets with retweet number 0 or 1, these models predict high retweet numbers. These three models also cannot predict retweet number precisely. From Fig. 3, we can conclude that traditional one-phase regression model cannot effectively predict retweet number.

4.2 Two-Phase Retweet Number Prediction Model

Basic Idea

Since the traditional one-phase regression model cannot solve the prediction problem, we propose a two-phase prediction model. The different between one-phase prediction model and two-phase prediction mode is shown in Fig. 1.

From Section 3.3, we know that the retweet number is extremely unbalanced. This characteristic may lead some models to predict small values, which can get less error. And comparing with predicting whether a tweet can be retweeted, predicting the retweet number is more complex. There are many features affecting the retweet number of a tweet. For example, the tweet is about special events, like earthquake or the tweet is created at a special time, like someone's birthday. What's more, the user may pay money to some people to get a high retweet number, like marketing and so on. Because of the unbalances distribution and the essential complexity, the one-phase can hardly predict very well.

To reduce the influence of the two problems above, we propose two-phase model. We know that only a small number of user can post tweets with high retweet number from Section 3.3. If we can classify tweets into some classes based on retweet number, the influence of the two problems will be reduced in each class. So the regression models which build in each class will get a better prediction. In our two-phase model, we build a multi-classification model on dataset in the first-phase. And in the second-phase, we build a regression model for each class. With "Divide and Conquer" strategy, our two-phase model can solve the unbalanced and complex problem in a way.

Table 3. Notations

Symbol	Description	Symbol	Description
T	The training set	L	The set of threshold values
x_i	the feature vector of the ith tweet in T	N	The set of class number which includes 1,2,3…,$\|L\|$+1
X	The set of feature vector in T	c_i	The class number of the ith tweet in T
r_i	The retweet number of the ith tweet in T	C	The set of class number in T
R	The set of retweet number in T	$\|S\|$	The size of set S

Algorithm 1. Two-phase retweet number prediction model training algorithm

Input: X, R, L, N
Output: Two-phase retweet prediction model

```
for i = 1 to |T| do
        compute the class number c_i of the ith tweet based on r_i and thresholds in L
end for
train classification model CLF based on X and C
classify X and R into |N| different parts based on different class number in C
for i = 1 to |N| do
        train regression model REG-i in part-i
end for
return (CLF,REG-1,REG-2,…,REG-|N|)
```

Algorithm 2. Two-phase retweet number prediction model predicting algorithm

Input: models CLF,REG-1,REG-2,…,REG-|N|, the feature of a tweet x
Output: the prediction retweet number r_p

```
use classification model CLF to classify x, return the class number c
use the regression model REG-c to predict the r_p based on x
return r_p
```

Classification Phase and Regression Phase

In first-phase of our model is the classification part. This part is important to our model because if the classification model doesn't have a good precision, our model cannot work very well. But there are two problems here: (1) How to select thresholds for classifying? (2) How many classes should we make?

For the first problem, we should not ignore the unbalanced distribution of the dataset. Considering the proportion of the tweets with different retweet number, we recommend to use the 80/20 Rule to choose thresholds. For example, if we want to classify tweets into 2 classes, the threshold will be retweet number of the tweet at the 80% position with a sorted datasets on retweet number. If we want to classify tweets into 3 classes, the positions will be 80% and 96%. Obviously, there are many other methods

to select thresholds to classify tweets. In Section 5, we discuss other methods with experiments. For the second problem, in Section 5, we also discuss the different influence on our model with different number of classes.

From Section 4.1, we have known training regression models directly on the whole dataset cannot get a good model. To prove the effectiveness of our two-phase model, we will still choose the same 4 regressions in Section 4.1 on experiments.

Algorithm Framework
Before describing the training algorithm, some notations are introduced in Table 3. In our two-phase model framework, if we classify tweets into n classes, we should train (n+1) models. The training algorithm and predicting algorithm are described in Algorithm 1 and 2.

5 Experiments

In this section, we conduct a series of experiments on Sina Weibo dataset. Since our dataset is too large, we random select 1 million tweets on experiments. First, we conduct experiments to select classification model. Then, we compare the prediction results of one-phase model and two-phase model. Last, we discuss some parameters of our model on experiments.

5.1 Effectiveness Experiments

Classification Model Comparison
Firstly, we compare the performances of different classifiers. We use the 80/20 Rule to classify tweets into three classes, the thresholds in our dataset is 5 and 118. So, the range of retweet number in three classes are 0-5, 6-118, 119-MAX. We choose 4 classification models, RandomForest (RF), Logistic (LO), DecisionTree (DT) and NativeBayes (NB) to experiment. We random select 60%, 70%, 80% and 90% of the dataset as training set and the rest as test set.

Table 4. The results of different classification models

M	60%			70%			80%			90%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
RF	**85.5%**	**86.8%**	**86.1%**	**85.7%**	**87.0%**	**86.3%**	**86.0%**	**87.2%**	**86.6%**	**86.0%**	**87.2%**	**86.6%**
LO	77.5%	81.5%	79.4%	77.6%	81.5%	79.5%	77.7%	81.6%	79.6%	77.8%	81.7%	79.7%
DT	84.0%	85.1%	84.5%	84.3%	85.3%	84.8%	84.6%	85.6%	85.1%	84.7%	85.7%	85.2%
NB	75.5%	79.7%	77.5%	75.6%	79.7%	77.6%	75.6%	79.8%	77.6%	75.7%	80.0%	77.8%

In Table 4, we compute precision (P), recall (R) and F1 value to evaluate each classification model. The result of RandomForest model is the best. So, we choose RandomForest model as our first phase model in the following experiments.

Two-phase Model vs One-phase Model

In this part, we compare results of one-phase model and two-phase model. We still use the same regression models in Section 4: LeastMedSq (LMS), LinearRegression (LR), M5P and MultilayerPerceptron (MP). We use MAE and RAE to evaluate the results. They are defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \tag{1}$$

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n}|p_i - r_i|}{\sum_{i=1}^{n}|r_i - r_m|} \tag{2}$$

where p_i is the prediction retweet number of the ith tweet in test set and r_i is the real retweet number. r_m is the mean retweet number in training set.

In order to fully test this model, we still random select 60%, 70%, 80% and 90% of the dataset as training set and the rest as test set to do experiments. From Table 5, we can see that two-phase model get a better prediction for each regression model. In the four two-phase models, the combination of RandomForest model and LeastMedSq model gets the best results.

Table 5. Comparation of one-phase and two-pahse models

Method	**60%**		**70%**		**80%**		**90%**	
	MAE	RAE	MAE	RAE	MAE	RAE	MAE	RAE
RF+LR	66.77	69.65%	65.76	69.29%	69.44	70.08%	73.89	71.28%
LR	104.15	108.64%	102.63	108.14%	101.89	102.83%	106.97	103.19%
RF+MP	65.81	68.65%	63.01	66.40%	62.37	62.95%	76.85	74.13%
MP	161.46	168.43%	132.02	139.11%	123.59	124.73%	388.33	374.61%
RF+LMS	**47.89**	**49.96%**	**47.54**	**50.09%**	**53.39**	**53.88%**	**58.22**	**56.16%**
LMS	51.59	53.81%	51.35	54.11%	57.42	57.95%	62.27	60.08%
RF+M5P	67.03	69.92%	63.88	67.31%	60.6	61.16%	71.66	69.13%
M5P	89.66	93.52%	84.13	88.65%	92.41	93.26%	106.97	103.19%

Most of time, we want to predict the approximate range of the retweet number of a tweet, rather than a specific value. So, we define the range of a specific number. Supposing a, b and n are positive integers, a<b and $10^a < n < 10^b$, the range of n is:

$$\mathrm{range(n)} = \left[\mathrm{n} - \frac{10^b - 10^a}{m}, \mathrm{n} + \frac{10^b - 10^a}{m}\right] \tag{3}$$

where m is a parameter to control the radius.

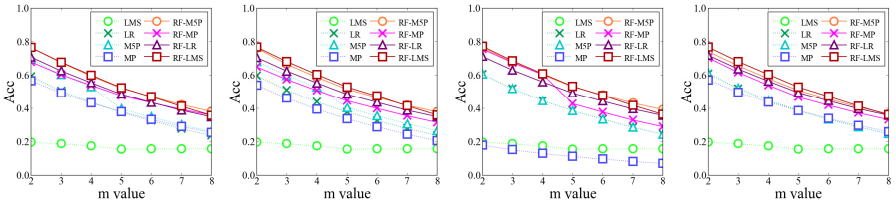
Supposing n_p is the prediction retweet number and n_r is the real retweet number, when n_p satisfies

$$n_p \in [n_r - \frac{10^{\lfloor log_{10}(n_r) \rfloor} - 10^{\lfloor log_{10}(n_r) \rfloor}}{m}, n_{r+} \frac{10^{\lfloor log_{10}(n_r) \rfloor} - 10^{\lfloor log_{10}(n_r) \rfloor}}{m}] \quad (4)$$

the prediction is right, otherwise is wrong. Then prediction accuracy is defined as follows.

$$\text{Acc} = \frac{the\ number\ of\ right\ predictons}{the\ number\ of\ all\ predictions} \quad (5)$$

Because of the unbalanced distribution of the retweet number, most of the retweet number is 0 and 1. If a model predicts all tweets with a small number, it will get high prediction accuracy. To avoid this phenomenon, we random select an extremely tough test set in the rest of the 1 million training set. The test set has 100k tweets with retweet number 0-100, 100k tweets with retweet number 101-1000 and 100k tweets with retweet number more than 1001. In the following experiments, we use this set as test set. We still random select 60%, 70%, 80% and 90% of the 1 million tweets to be training set. From Fig. 4, we can see that two-phase models have a better performance than one-phase models.



(a) 60% training set (b) 70% training set (c) 80% training set (d) 90%training set

Fig. 4. Comparison of one-phase model and two-phase model

5.2 Parameters Study

In previous experiments, we classify tweets into three classes based on the 80/20 Rule. Here, we test the influence of different class number on our model. We classify tweets into 2-6 classes with thresholds 5, 118, 766, 4143 and 16866. The details are in Table 6. In Fig. 5, with the increased number of classes, the model gets better prediction. But when the number is more than 4, the precision become stable.

In previous experiments, we extract feature from the tweet's creator and the tweet. To find out which kind of feature has more contribution to prediction, we use each of the two kind of feature alone to train the model. From Fig. 6, we can see that the tweet's creator(user) has more influence on prediction. This result tells us that when we want to evaluate the influence of a tweet, we should pay more attention to who posts the tweet rather than the tweet itself.

Table 6. Thresholds of different number of classes

	5	118	766	4143	16866
2-class	√	×	×	×	×
3-class	√	√	×	×	×
4-class	√	√	√	×	×
5-class	√	√	√	√	×
6-class	√	√	√	√	√

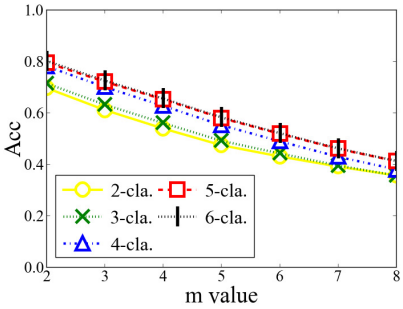


Fig. 5. Effect of different classes

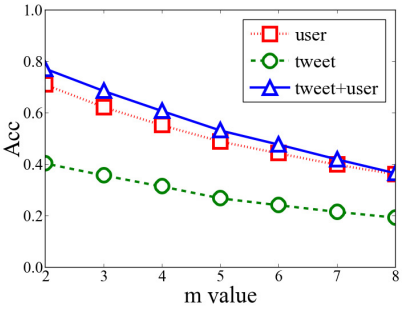
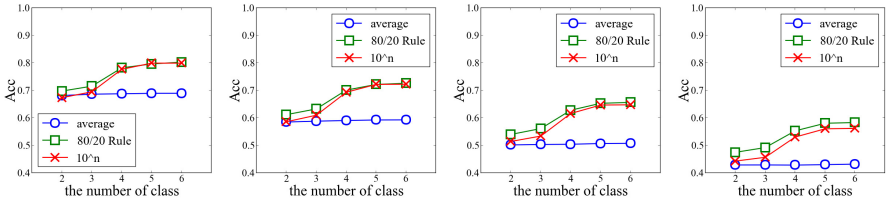


Fig. 6. Effect of different features



(a) m=2 (b) m=3 (c) m=4 (d) m=5

Fig. 7. Effect of different methods of classification

In Section 4, we say that there are other methods besides the 80/20 Rule to classify tweets. Here we discuss other methods based on the retweet number rather than proportion. Obviously, choosing 10^n as thresholds is an easy way. If we want 2 classes, the threshold is 10; 3 classes, the thresholds are 10, 100 and so on. Another method can be average allocation on retweet number. First, we choose the maximum retweet number as MAX_R. Then if we want 2 classes, the threshold is MAX_R/2; 3 classes, the thresholds are MAX_R/3, 2MAX_R/3 and so on. But in dataset of tweets, there are always some tweets that their retweet number like explosion. To avoid tweets like this, we filter retweet number that has less than 10 tweets heuristically, and then choose the MAX_R. We conduct lots of experiments to compare the three methods. From Fig. 7, we can see that no matter under what circumstances, the method of the 80/20 Rule gets the best prediction.

6 Conclusion

Retweet behavior is the key mechanism for information diffusion in micro-blog network. Retweet number, which denotes how many times that a tweet can be retweeted, is good measurement of both influence in diffusion and value in market of a tweet. To predict retweet number of a tweet in Sina Weibo, we build a two-phase retweet number prediction model. Experiments conducted on real dataset in Sina Weibo show that our two-phase retweet number model has better performance than traditional one-phase prediction model. In the experiments, we also find that the features of tweet's creator have more influence than the feature of tweet itself on retweet number.

Acknowledgments. This work is supported by the National Basic Research Program of China (2013CB329603). It is also supported by the National Natural Science Foundation of China (No. 61375058, 61074128, 71231002) and Ministry of Education of China and China Mobile Research Fund (MCM20123021).

References

1. Twitter, http://en.wikipedia.org/wiki/Twitter
2. Sina Weibo, http://en.wikipedia.org/wiki/Sina_Weibo
3. Petrovic, S., Osborne, M., Lavrenko, V.: RT to Win! Predicting Message Propagation in Twitter. In: ICWSM (2011)
4. Ma, H., Qian, W., Xia, F., et al.: Towards modeling popularity of microblogs. J. Frontiers of Computer Science 7(2), 171–184 (2013)
5. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: Proc. of CIKM, pp. 1633–1636 (2010)
6. Yang, J., Counts, S.: Predicting the speed, scale, and range of information diffusion in twitter. In: ICWSM (2010)
7. Peng, H., Zhu, J., Piao, D., Yan, R., Zhang, Y.: Retweet Modeling Using Conditional Random Fields. In: ICDM Workshops, pp. 336–343 (2011)
8. Zhang, Y., Lu, R., Yang, Q.: Predicting Retweeting in Microblogs. Journal of Chinese Information Processing 26(4), 109–114 (2012)
9. Bandari, R., Asur, S., Huberman, B.: The pulse of news in social media: forecasting popularity. In: ICWSM (2012)
10. Li, Y., Yu, H., Liu, L.: Predict algorithm of micro-blog retweet scale based on SVM. Application Research of Computers 30(9), 2594–2597 (2013)
11. Java, A., Song, X., Finin, T., et al.: Why we twitter: understanding microblogging usage and communities. In: WebKDD, pp. 56–65 (2007)
12. Boyd, D., Golder, S., Lotan, G.: Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In: 43rd Hawaii International Conf. on System Sciences, pp. 1–10 (2010)
13. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In: SocialCom/PASSAT, pp. 177–184 (2010)
14. Rousseeuw, P.J., Leroy, A.M.: Robust regression and outlier detection (1987)
15. Wang, Y., Witten, I.H.: Induction of model trees for predicting continuous classes. In: ECML, pp. 128–137 (1997)