Bebavior Analysis Based SMS Spammer Detection in Mobile Communication Networks

Zhang Bin, Zhao Gang, Feng Yunbo, Zhang Xiaolu[1], Jiang Weiqiang, Dai Jing, Gao Jiafeng[2]
China Mobile Communications Corporation
[2]China Mobile (Suzhou) Software Technology Co., Ltd.
[1]zhangxiaolu@chinamobile.com

***Abstract*—In a communication network, automatic short message service (SMS) spammer detection is a big challenge for a telecommunication operator nowadays, especially with the development of the rich communication services (RCS). Three main problems exist in the areas of research and real practice. They are (1) the whole-volume content based SMS spam detection techniques cannot be easily used on the side of network due to the issue of user privacy; (2) traditional ways to filter the spam according to the combination of key words and sending frequency can be easily bypassed by adding the interference words; (3) Most of them result in a great deal of manual review after the automatic filtering due to a low precision rate. To make up the aforementioned gaps, we study the user behavior characteristics. A two-dimensional visualized result indicates that any combination of two user behavior attributes cannot distinguish the abnormal users from the whole set by splitting the 2-dimensional space. Thus, the integration of multiple user behavior attributes is exploited to train the classifier in a labeled set by machine learning algorithms, respectively, including decision tree, random forest, supported vector machine (SVM), logistic regression, and self-organized feature mapping (SOM). The performance comparison indicates that random forest is a good choice to balance the tradeoff of the precision rate and the recall rate, and in an acceptable time. The experimental result shows the proposed method without the knowledge of SMS content has a significant improvement in terms of precision rate and recall rate compared with the traditional method using the combination of key words and sending frequency used in most of existing networks.**

Keywords—Behavior Analysis; SMS spammer; machine learning; communication network

I. INTRODUCTION

The SMS Spam is a form of spam directed at the text messaging or other communications services of mobile phones. It is even more annoying than email for the users because some of them may need to pay for the spam messages received. The severity of SMS spam varies from different regions. In some parts of Asia, up to 30% of messages were spam. According to "2015 1st Half Year China Mobile Internet Security Report" by Baidu [1], the amount of SMS spam reached 19 billion by the end of June of 2015, and 7 spam messages were received by per users on average.

Depending on the supporting information, the spam detection methods can be classified into 2 categories: (1) content-based and (2) behavior-based ones. Content-based filtering [2] is most used on the side of user terminal and network. Each message is searched for the features of spam, e.g. the key words of "Free" and "Fa Piao (invoice in Chinese character)", or special symbols like "BUY!!!".Before being used in SMS spammer detection, Contend-based methods are widely used in the unwanted email filtering ([3] and [4]). The first mathematical tool applied in email spam detection is the algorithm of Bayes by Sahami et.al in 1996 [5]. Due to the good performance of Naive Bayes Classifier (NBC), it was commonly used in email filtering systems and then expanded in the area of SMS spam detection [6]. In [6], the authors test a number of messages representation techniques and machine learning algorithms and show that Bayesian filtering techniques can be effectively transferred from email to SMS spam detection. However, the whole-volume content based SMS spam detection techniques cannot be easily used due to the issue of user privacy and the huge storage of the content for a telecom operator.

Behavior based analysis ([7] and [8]) is popular after the content-based one. In [9], behavioral characteristics of email spammer can significantly improve the efficiency of spam governance. The behavioral characteristics may include the statistics of spam messages from different spammers, the spam arrival patterns across the IP address space, the number of mail servers in different networks, and the active duration of spammers. Behavioral characteristics of spammers bring in many anti-spam mechanisms for email, e.g. [10] and [11]. Social network analysis to detect the spammer is one of the sub-branches of behavior based analysis [12]. Usually, the node of a spammer in a SMS network has a low in-degree and high out-degree. The special characteristics of social network are considered in our proposed method.

Depending on the position where to detect the SMS spam, the methods are generally divided into two categories: (1) the ones to detect spam at destination point; (2) the one to detect spam or spammer on the side of network. The spam filtering App developers such as 360 Safeguard takes more care of the former. They are unaware of spammer behavioral characteristics and can only use the content-based technique to avoid spam at the mobile terminal. A network operator, who focuses on the latter one, has both the knowledge of content and user behavior. The most commonly used method to detect the spammer is to set a frequency threshold in a given period for a certain key word or a logic combination of several key words. The disadvantage is that the aforementioned can be easily bypassed by adding the interference words.

The remainder of the paper is structured as follows. In Section II we describe analysis methodology and the data set of the SMS call detail record (CDR). Subsequently, we study

978-1-5090-1192-6/16 $31.00 © 2016 IEEE
DOI 10.1109/DSC.2016.48

the behavioral characteristics of spammers and carry out the attributes and model selection in Sections III and V, respectively. Experiment results in a real telecom network are discussed in the Section VI. The paper is concluded in Section VII.

II. PRELIMINARIES

A. Analysis Methodology

The flow chart of SMS spam detection is seen in Fig. 1. SMS CDR without the information of SMS content is the input of the system. The accumulation of CDR in a given period for analysis is used for statistical properties calculation in block 2 for each SMS sender, with the user behavior characteristics as the output of block 2. The input of classifier trained using machine learning algorithms in block 4 is divided into two sets: training set and test set, according to whether the label of spammer or non-spammer is given. The training set is used to train the classifier using machine learning algorithm at the beginning of the given period (training process). Then, the classifier with updated arguments obtained in the training process takes charge of SMS spammer detection in the test process.

Fig.1. The flow chart of SMS spam detection

SMS CDR: SMS SCR records the sending and receiving information for each SMS, including the calling number, called number, transmission time, receiving time and SMS text length.

User Behavior Characteristics: for each calling number, we use statistical properties to characterize the user behavior according to the information of SMS CDR in a given period for analysis. The list of statistical properties is chosen according to the following principles: (1) to cover as many characteristics as what can be imagined; (2) to include as few correlated ones as possible. According to the principle 1, 46 properties are proposed first. Following the principle 2, the number of properties is reduced to 33. They are listed in the Table I. To be mentioned, the values of x_{32} and x_{33} are obtained from social network analysis. The time period for statistical analysis can be varied according to practical situation. In our study, 1 minute, 5 minutes, 1 hour and 1 day are tested, respectively. In the following all experiment, 1 day is chosen as the time period for statistical analysis to avoid a large deal of vacancy in the list of statistical properties.

Machine learning algorithms: we consider 5 commonly used machine learning algorithms to train the classifier of SMS sender (spammer or non-spammer). They are self-organized feature mapping (SOM), binary decision tree, support vector machine (SVM), logistic regression and random forest. The theory and manual of the algorithms can be seen in [13].

TABLE I. THE LIST OF 33 STATISTICAL PROPERTIES

Var.	Property Name	Description
x_1	Calling freq.	The times of calling for a calling No.
x_2	Called No. number	The number of called No. for a calling No.
x_3	Regions number in SO	The number of regions called in a same operator with the calling No.
x_4	Called No. number in SR & SO	the number of called No. in a same region and a same operator with the calling No.
x_5	Called No. number in DR & SO	The number of called No. in different regions and a same operator with the calling No.
x_6	Top 1 called No. freq.	The times of calling to the 1st most called No.
x_7	Top 2 called No. freq.	The times of calling to the 2nd most called No.
x_8	Top 3 called No. freq.	The times of calling to the 3rd most called No.
x_9	Called No. dispersion	Called No. Number/calling freq.
x_{10}	Called No. number with EQ	The number of called No. number with equal difference feature
x_{11}	Called No. ratio with EQ	Called No. number with EQ/(Called No. number with EQ-1)
x_{12}	Similar called No. max freq.	The maximum times of called No. with the same top 7 numbers
x_{13}	Similar called No. max freq. ratio	Similar called No. max freq./Called No. Number
x_{14}	Called No. number in SR & SO v.s. SR & DO	Called No. Number in SR & SO/Called No. Number in SR & DO
x_{15}	Sorted called No. difference's s.d.	Standard deviation for the differences of sorted called numbers
x_{16}	Sorted called No.'s different digit avg. number	Average number of sorted called No.'s different digit in the same position
x_{17}	Top 1 called No. freq. ratio	Top 1 called No. freq./Calling freq.
x_{18}	Top 2 called No. freq. ratio	Top 2 called No. freq./Calling freq.
x_{19}	Top 3 called No. freq. ratio	Top 3 called No. freq./Calling freq.
x_{20}	Calling freq. in rush hours	Calling freq. in 9-11am and 3-5pm
x_{21}	Calling freq. in idle hours	Calling freq. in 9-11am and 0-8am
x_{22}	Max hour calling freq.	Maximum hour calling freq. in a given period
x_{23}	Calling freq. ratio in rush hours	Calling freq. in rush hours/Calling freq.
x_{24}	Calling freq. ratio in idle hours	Calling freq. in idle hours/Calling freq.
x_{25}	Max hour calling freq. ratio	Max hour calling freq./Calling freq.
x_{26}	Avg. calling interval	Average calling interval
x_{27}	Calling interval s.d.	Standard deviation of calling interval
x_{28}	Max SMS text length	Maximum SMS text length
x_{29}	Min SMS text length	Minimum SMS text length
x_{30}	Avg. SMS text length	Average SMS text length
x_{31}	SMS text length s.d.	Standard deviation of SMS text length
x_{32}	Linked number ratio in social network	For example, a calling No. contacts 100 called No., and 4 called No. have contact. The ratio is 4/100
x_{33}	Link ratio in social network	For example, a calling No. A contacts 100 called No., and 3 links have contact. The ratio is 3/100

B. Data Source

The data for analysis was the whole set of SMS CDR collected in a regional telecom network between April 7, 2015 to April 13, 2015. The difficulty point lies in how to label the sample of the training set as spammer or non-spammer. The spammer label information comes from the database of inner blacklist, user complaint, and artificial review result, and the 3rd part blacklist; the non-spammer one comes from the list of high value VIP customers and their frequent contacts. Among all SMS senders in that period, 322,217 calling No. has been labeled with 317,552 labeled non-spammers and 4,665 labeled spammer. The other 3,717,529 SMS senders have CDR but no label information.

III. BEHAVIOR CHARACTERISTICS

In this section, the study on the behavior characteristics of spammers is presented. In particular, we compare 3 representative statistics of the behavior properties of spammer and non-spammer. Some of the findings can be used in anti-spam system design directly.

A. The called No. number

It is intuitive that the number of called No. for a spammer is greater than a non-spammer. The analysis result verifies the hypotheses with p-value of 0. The average value of call No. number is 2.67 for a non-spammer and 40.96 for a spammer. Fig. 2 shows the probability density function (PDF) of the called No. number for spammer and non-spammer.

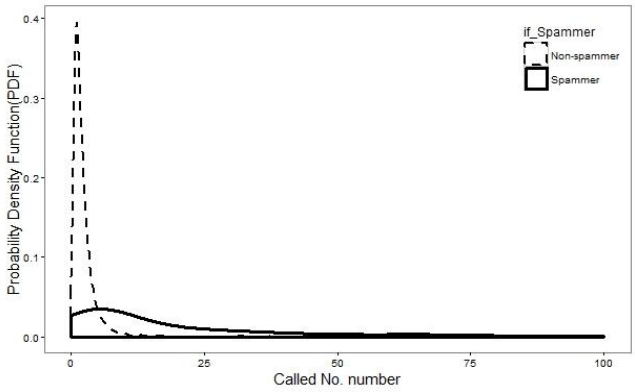

Fig. 2 The distribution of the called No. number of spammer and non-spammer

The difference of the called No. number between spammer and non-spammer is statistically significant. For 93% of non-spammers, the called No. number is less than 5. For 70% of spammers, the called No. number is more than 5.

B. The top 1 called No. freq. ratio

For a non-spammer, it is possible to send more SMS to the closed contract, but spammer sends SMS dispersedly to the victim. The analysis result verifies the hypotheses that the top 1 called No. freq. ratio of non-spammer is significantly more than spammer with p-value of 0. The average values of the top 1 called No. freq. ratio of non-spammer and spammer are 0.75 and 0.24 respectively. Fig. 3 shows the distribution of the top 1 called No. freq. ratio for spammer and non-spammer.

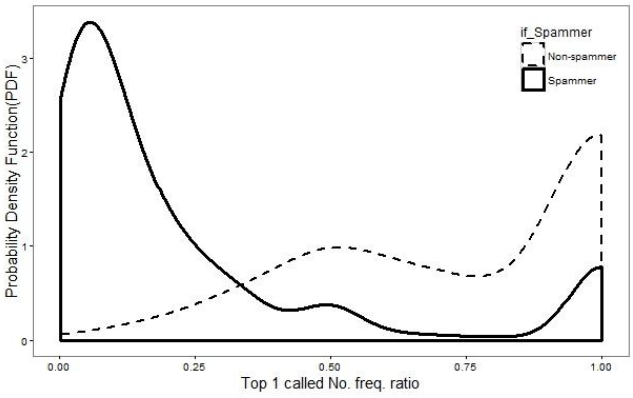

Fig. 3 The distribution of the top 1 called No. freq. ratio for spammer and non-spammer

In Fig. 3, we can see that the called No. number of non-spammer is mostly distributed above 0.5, and the one of spammer is below 0.5.

C. Avg. SMS text length

SMS text messages are limited to 160 seven-bit characters. Each SMS length cannot go beyond 140 Chinese characters. When the text length is more than the limitation, the message will be divided in to multiple ones automatically. To fully utilize the limitation, spammers design the spam content with a long text length but within the limitation. The hypotheses that the avg. SMS text length for a spammer is significantly more than non-spammer is proved by the statistical study with the p-value of 0. The avg. SMS text length for a spammer is 102.53, greater than the value for a non-spammer by 186%. Fig. 4 shows the distribution of the avg. SMS text length for spammer and non-spammer. The data analysis indicates that the avg. SMS text length for 76.53% non-spammers is less than 50, and the one for 95.20% spammers is more than 50.

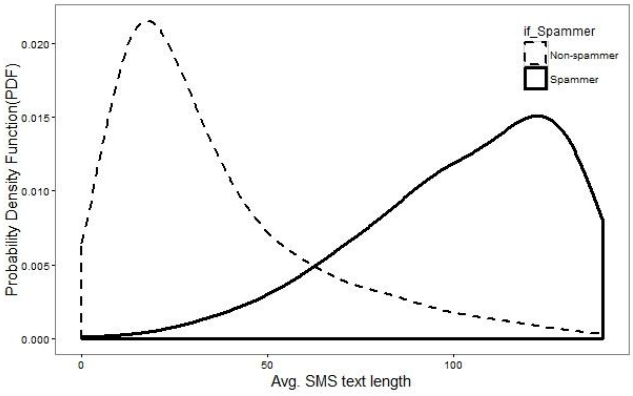

Fig. 4 The distribution of the avg. SMS text length for spammer and non-spammer

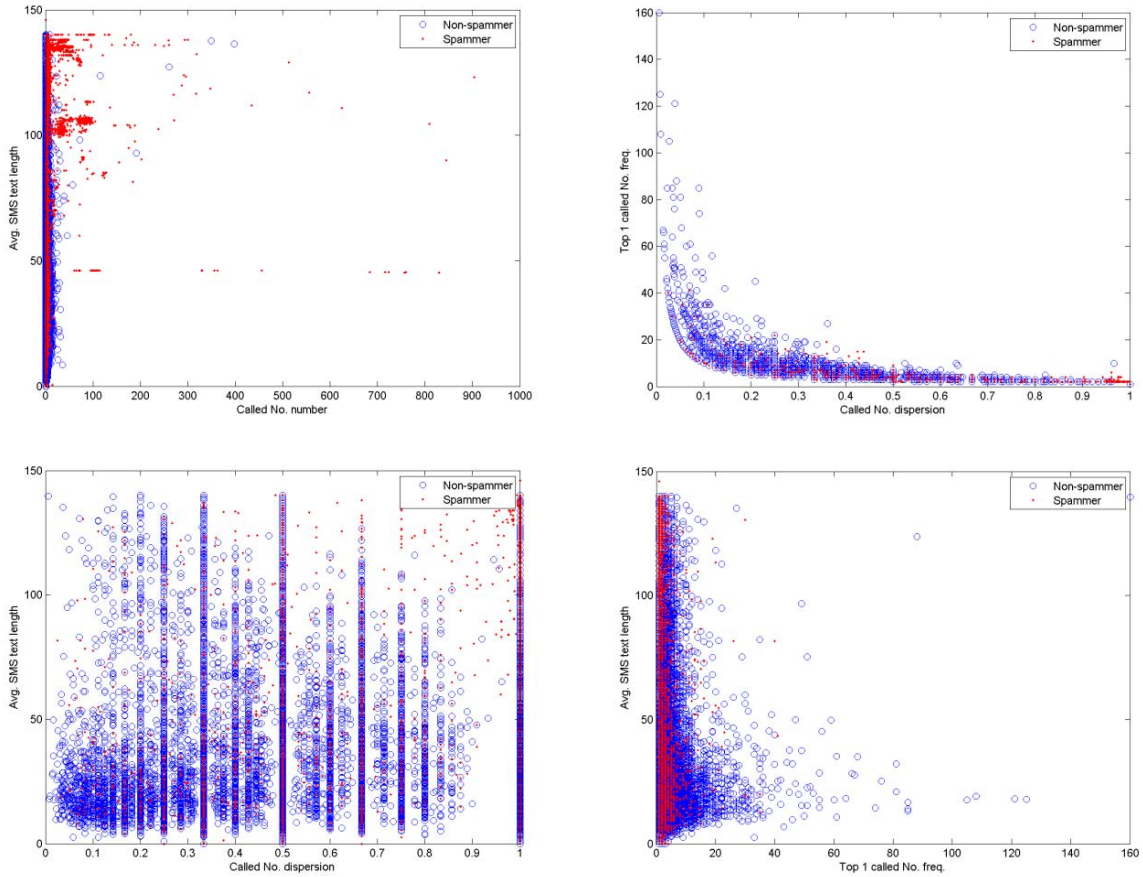

Fig. 5 Four of 528 scatter diagrams plotted for any two of 33 behavior characteristics

Three representative behavior characteristics are compared between the spammer and non-spammer. All the differences are statistical significant. A problem comes: is it possible to distinguish the spammer from non-spammer by splitting 2 dimensional spaces of behavior characteristics. To test it, the scatter diagrams are plotted for any two dimensions of 33 ones. The total number of combination is 33*32/2=528. In all 528 scatter diagrams, 4 of them are shown in Fig. 5. The diagrams show that it is hard to visually distinguish the spammers from non-spammers by splitting 2 dimensional spaces of behavior characteristics, although they have totally different possibility distribution. For that, we resort to designing a classifier using machine learning algorithms.

IV. FEATURE AND MODEL SELECTION

A. Feature selection

In the machine learning's application, irrelevant or redundant features may result in: (1) long time to train the model; (2) the curse of dimensionality [15] that complex model leads to be hard to apply in practice. To improve the precision rate and the recall rate, and reduce the training/test time, relevant features are selected from the set of 33 behavior characteristics. The principle to select features is (1) selected features have low correlation with each other; (2) selected features are relevant to the variable of if-spammer. Fig. 6 shows the correlation matrix of 33 behavior characteristics.

The axis of symmetry displays the variables of behavior characteristics. The area of sector represents the absolute value of the correlation coefficient with the red color indicating positive and blue negative.

According to the correlation matrix, the 33 variables can be divided into several groups. Within each group the variables are more correlated, and two variables not in the same group are less correlated. In each group, the variable that is most relevant to the variable of if-spammer is chosen as the selected feature. The final 10 selected features are:

1. x_2 Called No. number (CNN)
2. x_9 Called No. dispersion
3. x_{12} Similar called No. max freq.
4. x_{14} CNN in SR & SO v.s. SR & DO
5. x_{17} Top 1 called No. freq. ratio
6. x_{23} Calling freq. ratio in rush hours
7. x_{24} Calling freq. ratio in idle hours
8. x_{27} Calling interval s.d.
9. x_{30} Avg. SMS text length
10. x_{33} link ratio in social network

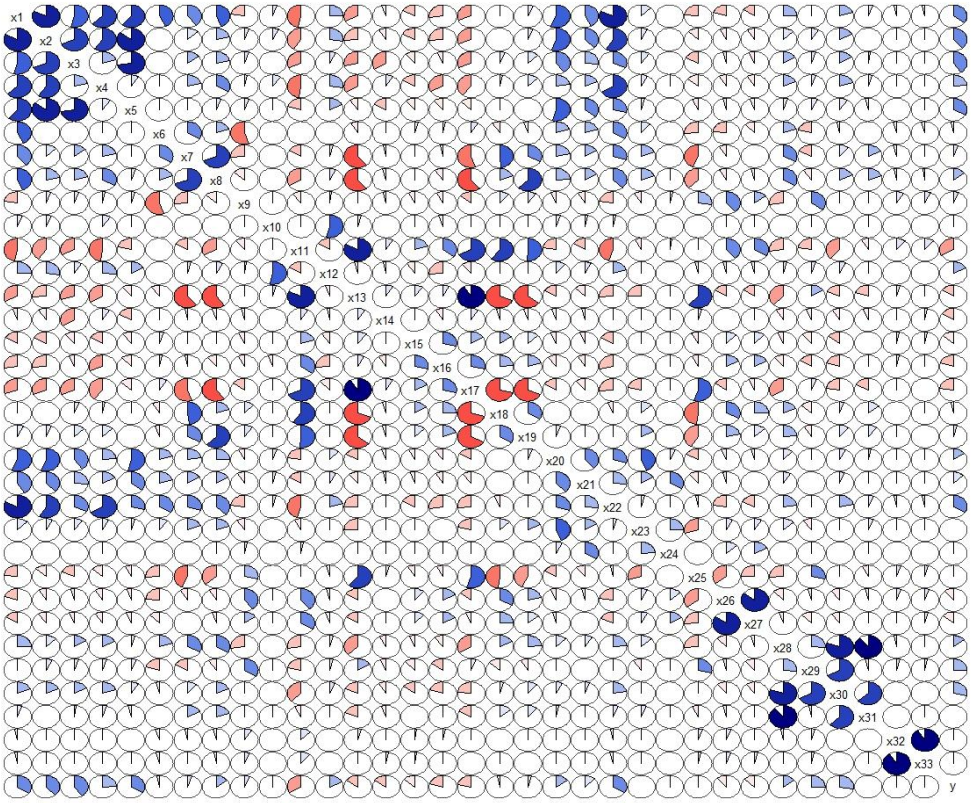

Fig. 6 The correlation matrix of 33 behavior characteristics

TABLE II. THE PERFORMANCE COMPARISON OF ALGORITHMS

	Random Forest	SOM	Decision Tree	SVM-Linear	SVM-Gaussian	Logistic Reg.
Precision Rate	95.52%	92.85%	90.14%	86.08%	96.77%	92.75%
Recall Rate	86.26%	77.73%	88.11%	87.76%	74.38%	87.44%
F-measure	90.65%	84.62%	89.11%	86.91%	84.11%	90.02%
Training time (s)	27	28	7	967	262	6
Test time (s)	3	4	1	2	22	1

B. Model Selection

As mentioned in Section II, multiple machine learning algorithms can be exploited to train the classifier. One of ways to optimize the classifier is to design a linear combination of multiple classifiers obtained from different algorithms with a minimum mean square error (MSE). However, it is impractical when the proposed classifier is used in real-time system where the decision should be made within an hour. Thus, choosing a best machine learning algorithm before the classifier comes online is reasonable. Using the data of 322,217 labeled calling No., we compare the performance of 6 algorithms in terms of precision rate, recall rate, F-measure (key index), training time and test time. Table II indicates that random forest is a good choice to be deployed online.

V. EXPERIMENT RESULTS IN A REAL TELECOM NETWORK

The experiment is conducted following the process shown in Fig. 1. The 10 behavior characteristics selected in Section V are the output of statistical properties calculation and the input of the classifier trained by machine learning algorithm. According to the test result in Section V and practical constraints, the random forest is selected as the algorithm of automatically learning. The whole set of 322,217 labeled users is used to train the classifier and the whole set of 3,717,529 SMS senders with CDR but without label information is used to test the performance of our proposed SMS spammer detection method in a real telecom network.

In the algorithm of random forest, two parameters, the depth of the tree, h, and splitting number, s, is to be determined. An ergodic search is conduct for all integers of h in [3, 7] and s in [15, 35]. Taking the F-measure as the objective, the combination of (5, 24) is chosen as the optimal parameter pair. The value of Mean Decrease Gini (MDG) can be used to evaluate the contribution of the different features. The experiment result shows that the top 5 important features are shown in TABLE III.

TABLE III. THE TOP 5 IMPORTANT FEATURES IN RANDOM FOREST

Importance	Variables	Property Name
1	x_{17}	Top 1 called No. freq. ratio 1
2	x_{30}	Avg. SMS text length 2
3	x_{2}	Called No. number 3
4	x_{23}	Calling freq. ratio in rush hours 4
5	x_{14}	Called No. number in SR & SO v.s. SR & DO 3 5

The importance of the features can be treated as the significant reference of the real system design. The number of the suspected spammers as the output of the classifier is 13,632. Under the manual review, in 13,632 suspected ones, 8,482 are the real spammers. The precision rate is 62.22%, 3 times higher than the results running in real network using the combination of key words and sending frequency on average. Compared with the results running in real network, the recall rate is increased by 401%. The reason why we cannot calculate the value of recall rate but a increasing rate is that it is impossible to give labels to a set of 3,717,529 SMS senders in practice within a limited time and labor load.

VI. CONCLUSION

In the paper, the behavior analysis based spammer detection is studied on the side of network. Firstly, we compare the behavior characteristics between spammer and non-spammer. It is shown that the differences in probability distribution are statistical significant. We found that (1) the average value of call No. number is 40.96 for a spammer, 15 times higher than the one for a non-spammer; (2) the top 1 called No. freq. ratio of non-spammer is mostly distributed above 0.5, and the one of spammer is below 0.25; 3. the avg. SMS text length for a spammer is 102.53, greater than the value for a non-spammer by 186%. A further study indicates that it is hard to visually distinguish the spammers from non-spammers by splitting 2 dimensional spaces of behavior characteristics, although they have totally different possibility distribution.

After comparing the performance of different machine learning algorithms, a random forest algorithm is proposed to automatically training the classifier with selected features. The experiment result shows that the proposed method has a significant improvement in terms of precision rate and recall rate compared with the traditional method using the combination of key words and sending frequency. Since the SMS content is not used as the input of classifier, we can guarantee the issue of user privacy. Another advantage is that the adaptively learning prevents that the proposed policies can be easily bypassed.

In the future work, it is expected to design a second layer classifier to improve the precision rate with sacrificing the recall rate properly. It may be realized by taking the SMS text content or content characteristics of suspected spammers as the input of the classifier. The 2nd layer classifier may reduce the workload of manual review.

REFERENCES

[1] Baidu, 2015 1st Half Year Mobile Internet Security Report in China, http://anquan.baidu.com

[2] Sebastiani, F., 2002. “Machine learning in automated text categorization,” ACM Computing Surveys, 34(1): 1-47.

[3] R.J. Hall. “How to avoid unwanted email,” Communications of the ACM, Mar. 1998.

[4] G. Tang, J. Pei, and W-S. Luk. "Email Mining: Tasks, Common Techniques, and Tools,” Knowledge and Information Systems: An International Journal, Vol 41(1): 1-31, Oct. 2014

[5] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, “A Bayesian Approach to Filtering Junk Email,” AAAI Technical Report WS-98-05, AAAI Workshop on Learning for Text Categorization, 1998.

[6] V.D. Odón and G. C. Bringas, Content Based SMS Spam Filtering, In Proc. of the 2006 ACM symposium on Document engineering, 107-114

[7] B. Shie, P. S. Yu, V. S. Tseng, “Mining interesting user behavior patterns in mobile commerce environments,” Appl. Intell. 38(3): 418-435, 2013

[8] L. Cao, P.S. Yu, H. Motoda, and G. Williams, “Special issue on behavior computing,” Knowledge and Information Systems, vol. 37, no. 2, pp. 245-249

[9] Z. Duan, K. Gopalan and X. Yuan, “Behavioral characteristics of spammers and their network reachability properties,” in proc. of ICC, 2007

[10] M. Delany, “Domain-based email authentication using public-keys advertised in the DNS,” Internet Draft, RFC 4870, 2007

[11] M. Lentczner and M. W. Wong, “Sender policy framework (spf): Authorizing use of domains in MAIL FROM,” RFC 4408, 2006

[12] F. Li, H. Mo-Han and G. Pawel, “The Community Be-havior of Spammers,” Communications and Network, 3(3):153-160, 2011

[13] J. Han, M. Kamber, and J. Pei. Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann, 2011, ISBN: 978-0-123-81479-1.

[14] G. James. D. Witten, T. Hastie; R. Tibshirani, An Introduction to Statistical Learning, Springer, p. 204, 2013

[15] A. Zimek, E. Schubert, H.-P. Kriegel, "A survey on unsupervised outlier detection in high-dimensional numerical data," Statistical Analysis and Data Mining 5 (5): 363–387, 2012