International Journal of Security and Its Applications
Vol.9, No.1 (2015), pp.329-336
http://dx.doi.org/10.14257/ijsia.2015.9.1.31

SMS Spam Filterinig Using Keyword Frequency Ratio

Sin-Eon Kim[1], Jung-Tae Jo[2] and Sang-Hyun Choi[3,*]

[1]*Department of Information Security Management*
[2]*Department of Business Data Conversions*
[3]*Professor, Department of Management Information System, BK21+ BSO Team*
Chungbuk National University
52 Naesudong-ro, Heungdeok-gu, Chungbuk 361-763 Korea

Abstract

As the amount of cellphone text message use has increased, spam text messages also have increased. Presently in mobile devices, spam filtering methods are in a very basic level such as simple character string comparison or specific number blocking. Typical filtering methods such as bayesian classifier, logistic regression and decision tree for detecting spam messages take quite a long time. In order to perform spam filtering with these methods, high performance computer resources and lots of SMS samples are required. In addition, if servers come to store normal messages, the problem of personal information infringement could arise. For mobile devices to independently perform spam filtering, there are many limitations in the aspects of storage space, memory, and CPU processing capability. Thus, this study tries to propose light and quick algorithm through which SMS filtering can be performed within mobile devices independently.

Keywords: *Mobile phone spam, SMS spam, spam filtering, Data Mining*

1. Introduction

Spam is unsolicited and unwanted messages sent electronically. Email spam is sent/received over the Internet while SMS spam is typically transmitted over a mobile network. Traditional email spammers are moving to the mobile networks as the return from the email channel is diminishing due to effective filtering, industry collaboration and user awareness [3]. The Short Messaging Service (SMS) mobile communication system is attractive for criminal gangs for a number of reasons [3]. It is becoming cost effective to target SMS because of the availability of unlimited pre-pay SMS packages in countries such as India, Pakistan, China, and increasingly the US. In addition SMS can result in higher response rates than email spam as SMS is a trusted service with subscribers comfortable with using it for confidential information exchange [3] (SJ Delany 2012).According to the GSMA it is inevitable that mobile network operators across the globe will see a rise in the volume and sophistication of SMS attacks in 2011 [3, 7] (GSMA, 2011b).

SMS spam is an emerging problem in the Middle East and Asia, with SMS spam contributing to 20–30% of all SMS traffic in China and India [3, 7] (GSMA, 2011b). As an example of this Chinese mobile subscribers received 200 billion spam messages in one week in 2008.1 While it is estimated that in North America the current level of mobile spam is currently only 0.1% of all messages per person per day [3, 6] (GSMA, 2011a), 44% of mobile device owners surveyed in the US reported receiving SMS spam [3] (SJ Delany 2012).

In this paper, efficient spam filtering, techniques to remove unnecessary data are needed. These data reducing techniques include data filtering, feature selection, data clustering, etc. The main idea is to select important features using relative magnitude of

*Corresponding author, chois@cbnu.ac.kr, Tel: +82-43-261-3742

ISSN: 1738-9976 IJSIA

feature values. We compare the performance of our method with standard feature selection methods; Naive Bayes, J-48 Decision Trees, Logistic. We propose a new feature selection methodthe average ratio of each class relative to total data. We compare between proposed method and other methods.

2. Related Work

The researches include statistic-based methods, such as bayesian based classifiers, logistic regression and decision tree method. There are still few studies about SMS spam filtering methods available in the research journals while researches about email spam classifiers are continuously increasing. We present the most relevant works related to this topic.

Gómez Hidalgo *et al.,* (2006) evaluated several Bayesian based classifiers to detect mobile phone spam. In this work, the authors proposed the first two well-known SMS spam datasets: the Spanish (199 spam and 1,157 ham) and English (82 spam and 1,119 ham) test databases. They have tested on them a number of messages representation techniques and machine learning algorithms, in terms of effectiveness. The results indicate that Bayesian filtering techniques can be effectively employed to classify SMS spam [5].

Cormack *et al.* (2007) have claimed that email filtering techniques require some adaptation to reach good levels of performance on SMS spam, especially regarding message representation. Thus, to support their assumption, they have performed experiments on SMS filtering using top performing email spam filters (e.g. Bogofilter, Dynamic Markov Compression, Logistic Regression, SVM, and OSBF) on mobile spam messages using a suitable feature representation. However, after analyzing the results, it was concluded that the differences among all the evaluated filters were not clear, so more experiments with a larger dataset would be required [1].

Delany *et al.* (2012) have reviewed recent developments in SMS spam filtering and also discussed important issues with data collection and availability for furthering research, beyond being analyzed a large corpus of SMS spam. They have built a new dataset with ham messages extracted from GrumbleText and WhoCallsMe websites and spam messages from the SMS Spam Collection. They analyzed different types of spam using content based clustering and identified ten clearly-defined clusters. According to the authors, such result may reflect the extent of near-repetition in data due to the similarity between different spam attacks and the breadth of obfuscation used by spammers [2].

2.1. SMS Spam Collection v.1 Data Set

The SMS Spam Collection v.1 is a set of SMS tagged messages that have been collected for SMS spam research. It contains one set of SMS messages in English of 5,574 messages, tagged according being ham or spam. The data is contain one message per line. Each line is consist of two columns: one with label (ham or spam) and other with the raw text.

Table 1. Type of Features

Message	Amount	%
Hams	4,827	86.60
Spams	747	13.40
Total	5,574	100%

As shown in Table 1, the data set has 86.6% of Ham message and 13.4% of Spam message. Table 2 shows some examples about ham and spam messages.

Table 2. Examples of Messages in the SMS Spam Collection

ham	Even my brother is not like to speak with me. They treate me ……
ham	Is that seriously how you spell his name?
spam	WINNER!! As a valued network customer you have been selected to .....
ham	I've been searching for the right words to thank you for this ……
ham	Is that seriously how you spell his name?
spam	FreeMsg Hey there darling it's been 3 week's now and no word back!.....

2.2. Data Mining Algorithm

2.2.1. Navie Bayesian: The naive Bayesian classifier provides a simple approach, with clear semantics, to representing, using, and learning probabilitistic knowledge. The method is designed for use in supervised induction tasks, in which the performance goal is to accurately predict the class of test instances and in which the training instances includes class information [10] (John and Pat 1995).

One can view such a classifier as a specialized form of Bayesian network, termed naive because it relies on two important simplifying assumptions. In particular, it assumes that the predictive attributes are conditionally independent given the class, and it posits that no hidden or latent attributes influence the prediction process. Thus, When depicted graphically, a naive Bayesian classifier has the form shown in Figure 1, in which all arcs are directed from the class attribute to the observable attribures [10] (Buntine 1994).

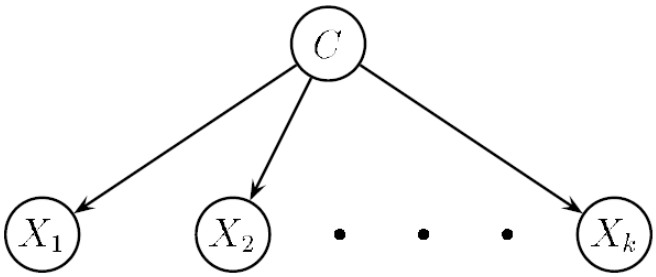

Figure 1. A Naive Bayesian Classifier Depicted as a Bayesian Network in which the Predictive Attributes (X_1,X_2,…X_k) are Conditionally Independent Given the Class Attribute (C) [10]; (John and Pat 1995)

2.2.2. Decision Tree: Decision tree algorithms begin with a set of cases, or examples, and create a tree data structure that can be used to classify new cases. Each case is described by a set of attributes (or features) which can have numeric or symbolic values. Associated with each training case is a label representing the name of a class. Each internal node of a decision tree contains a test, the result of which is used to decide what branch to follow from that node. For example, a test might ask "is x > 4 for attribute x?" If the test is true, then the case will proceed down the left branch, and if not then it will follow the right branch. The leaf nodes contain class labels instead of tests. In classification mode, when a test case (which has no label) reaches a leaf node, C4.5 classifies it using the label stored there [12] (Ross Quinlan 1993).

Decision trees have been used as classifiers for numerous real-world domains, some of which are mentioned and used as examples by Quinlan; *e.g.*, labor negotiations, hypothyroid diagnosis, soybean disease diagnosis, and credit approval. For many ofthese domains, the trees produced by C4.5 are both small and accurate, resulting in fast, reliable classifiers. These properties make decision trees a valuable and popular tool for classification [12].

2.2.3. Logistic Regression: Logistic regression is a widely used statistical modeling technique in which the probability of an outcome is related to a series of potential predictor variables by an equation of the form

$$\log [p/(1-p)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i (1)$$

Where , p is the probability of the outcome of interest, β_0 is an intercept term, $\beta_1, \ldots, \beta_i$, are the β coefficients associated with each variable, X_1 , … , X_i are the values of the potential predictor variables,and i is a unique subscript denoting each variable [8]. The usual assumption is that these predictor variables are related in a linear manner to the log odds { log $[p/(1-p)]$} of the outcome of interest. Variables are usually selected for inclusion in these models through some form of backward or forward stepwise regression technique, although stepwise variable selection techniques may be prone to problems [4]. Full model fits without deletion of insignificant variables may be preferred under certain circumstances [15]. Logistic regression models use as their convergence criterion the maximization of a likelihood function. β coefficients can easily he converted into the corresponding odds ratios by raising e to the coefficient if variables are represented by a single linear term or as dummy variables, and one can easily interpret the magnitude of importance of a predictor. In addition, widely accepted criteria such as the Hosmer-Lemeshow statistic have been developed for assessing the "goodness of fit" of these models [8]. Thus, logistic regression has become the technique of choice for statistical modeling in which the outcome of interest isdichotomous [16] (Tu Jack V 1996).

3. Experimental Study

We explained above that SMS spam data is rapidly increasing. In order to detect spam messages, filtering algorithms or feature selection methods have to be more efficiently run. The above three methods use a complex calculation to do this. For this reason, these methods are inefficient for dealing with large scale data. In this paper, we propose a simple and efficient feature selection method.

For the purpose of analyzing the SMS messages, we break each message into a set of unit keywords or words by using the function 'string to word vector' in WEKA. For example, a message 'WINNER!! As a valued network customer you have been selected to ...' can be break into { a, as, value, network, customer, WINNER, ... }. We performed the above preprocessing procedures for total 5,574 messages. Finally we got the message-keyword profile matrix as shown in Table 3. In Table 3, W_{24} shows that the 4th keyword 'date' is referred in the 2th message. The variable W_{ij} represents whetherthe jth keyword is referred in the ith message or not.

Table 3. Example of Message-keyword Profile Matrix

Msg No.	are	chat	cost	date	do	to	class
message 1	0	0	1	0	1	0	spam
message 2	1	0	0	1	0	0	ham
message 3	0	1	0	0	0	1	ham
message 4	1	0	0	0	1	0	spam
message 5	0	0	1	0	0	0	spam

3.1. Experimental Setup

We used WEKA 3.7 a machine learning tool, to evaluate the performance of feature selection methods such as Naive Bayes, J-48 Decision Trees, and Logistic, and to evaluate the classification performance on each of these feature sets. We chose algorithm with full training set and 10-fold cross validation for the testing purposes. In 10-fold cross-

validation, the available data is randomly divided into 10 disjoint subsets of approximately equal size. One of the subsets is then used as the test set and the remaining nine sets are used for building the classifier. The test set is then used to estimate the accuracy, and the accuracy estimate is the mean of the estimates for each of the classifiers. Cross-validation method has been tested extensively and has generally been found to work well when sufficient data is available. Finally, we used accuracy measures of TP and FP rate and efficiency measure of cpu time in the unit of second.

3.2. Proposed Method

This study proposes a FR (Frequency Ratio) measure for evaluating lightness and quickness of filtering methods so that SMS filtering can be performed independently within mobile devices.

First, each Class (Spam and Ham) is divided, and appearance frequencies of words on SMS messages are evaluated. Then the appearance frequencies of each word are aggregated and then divided by the number of messages to calculate an average. The formula is as below.

$$\overline{W}_j^s = \sum_{i \in spam} W_{ij} / k = W_j / k \qquad (2)$$

$$\overline{W}_j^h = \sum_{i \in ham} W_{ij} / k = W_j / k \qquad (3)$$

Here, i and j represent row and column respectively, and total messages is k.

A FR measure by using calculated $\overline{W}_j^s$ and $\overline{W}_j^h$ values is calculated as follow.

$$\text{FR(j)} = \overline{W}_j^s / \overline{W}_j^h \qquad (4)$$

FR(j) represents the relative ratio of average frequency of *j*th keyword in spam messages to that in ham messages. As the value of FR(j) is larger, the words are more frequently referred in spam messages.

3.3. SMS Spam Collection Data Set

The next table is the result of executing each algorithm with the whole SMS Spam Data set on Weka. Naive Bayes shows 1.23 seconds of CPU Time and 96.92% of Accuracy, J-48 algorithm 61.43 and 96.33%, and Logistic algorithm 28.34 and 96.92%.

Table 4. The Result of Algorithm with Total Data

algorithm	Correctly Classified Instances	TP rate		FP rate		CPUTime (seconds)
		spam	ham	spam	ham	
Naive	96.92%	0.879	0.983	0.017	0.121	1.23
J-48	96.33%	0.799	0.989	0.011	0.201	61.43
Logistic	94.70%	0.905	0.954	0.046	0.095	28.34

3.4. Selected Variables

The following table is the result of executing each algorithm with the FR on Weka. Naive Bayes shows 0.01 seconds of CPU Time and 94.70% of Accuracy, J-48 algorithm 0.02 and 94.82%, and Logistic algorithm 0.1 and 94.71%.

Table 5. The Result of Algorithm with Reduced Data using FR

algorithm	Correctly Classified Instances	TP rate		FP rate		CPU Time (seconds)
		spam	ham	spam	ham	
Naive	94.70%	0.992	0.658	0.342	0.008	0.01
J-48	94.82%	0.993	0.342	0.342	0.007	0.02
Logistic	94.71%	0.991	0.661	0.339	0.009	0.1

As above, algorithms with using the FR feature selection technique can drastically reduce CPU Time with maintaining the accuracy similar.

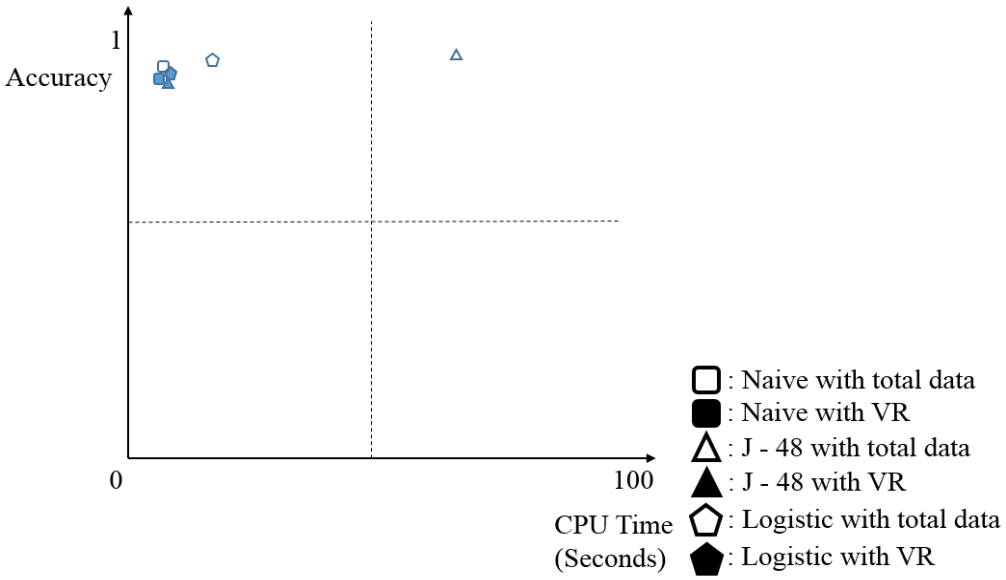

Figure 2. The Result of Algorithms

As shown in the figure, as a result of executing algorithms by using the FR attribute selection technique, CPU Time varied much. Thus, it is expected to fit for executing algorithms independently in the mobile environment that has many limitations in the aspects of storage space, memory, and CPU processing capability.

4. Discussion

The FR (Frequency Ratio) attribute selection technique proposed in this study has an advantage that it has a simple calculation formula compared to other techniques so that it can create indexes quickly. It could be proved that it has a similar capability to those of others even though it uses a simple calculation formula.

In the future, researches should make a program with the method proposed in this study and prove that it is an efficient technique by conducting a comparative analysis on calculated times taken when it is performed within actual mobile phones independently. Because spam messages continuously increase, data should be added constantly for a precise analysis. Additionally, the proposed method should not be limited in the spam filtering but applied to various fields to extract useful information so that researches on data reducing techniques for an efficient analysis in the massive data environment can be conducted.

Acknowledgements

This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2014-H0301-14-1022) supervised by the NIPA (National IT Industry Promotion Agency).

This research was supported by the MSIP (The Ministry of Science, ICT and Future Planning), Korea, under the "SW master's course of a hiring contract" support program (NIPA-2013-HB301-13-1008) supervised by the NIPA (National IT Industry Promotion Agency).

References

[1] G. V. Cormack, J. M. G. Hidalgo and E. P Sánz, "Feature engineering for mobile (SMS) spam filtering", In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, **(2007)**, pp. 871-872.
[2] S. J. Delany, M. Buckley and D. Greene, "SMS spam filtering: methods and data", Expert Systems with Applications, vol. 39, no. 10, **(2012)**, pp. 9899-9908.
[3] S. J. Delany, M. Buckley, and D. Greene, "SMS spam filtering: methods and data", Expert Systems with Applications, vol. 39, no. 10, **(2012),** pp. 9899-9908.
[4] D. S. Keselman, HJ, "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables", Br. J. Math Stat. Psycho., vol. 45, (**1992**), pp. 265-282.
[5] J. M. Gómez Hidalgo, G. C. Bringas, E. P. Sánz, and F. C. García, "Content based SMS spam filtering. In Proceedings of the 2006 ACM symposium on Document engineering, **(2006)**, pp. 107-114.
[6] GSMA, Operator FAQs, GSMA spam reporting service, **(2011a).**
[7] GSMA, "SMS spam and mobile messaging attacks – Introduction", Trends and examples, GSMA spam reporting service, **(2011b).**
[8] Hosmer Jr, David W., and S. Lemeshow, "Applied logistic regression", John Wiley & Sons, **(2004).**
[9] http://www.cs.waikato.ac.nz/~ml/weka/
[10] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers", Proceedings of the Eleventh conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. **(1995).**
[11] L. H. Setiono, R. Motoda and H. Zhao Z., "Feature Selection: An Ever Evolving Frontier in Data Mining", **(2010)** JMLR: Workshop and Conference Proceedings, pp. 4-13.
[12] R. Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, **(1993)**, San Mateo, CA.
[13] S. Mukherjeea and N. Sharmaa, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction", **(2012),** Procedia Technology, pp. 119-128.
[14] SMS Spam Collection v.1 **(2012)** http://archive.ics.uci.edu/ml/index.html
[15] D. J. Spiegelhalter, "Probabilistic prediction in patient management and clinical trials", Stat Med., vol. 5, (**1986**), pp. 421-433.
[16] J. V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes", Journal of clinical epidemiology, vol. 49, no. 11, **(1996)**, pp. 1225-1231.

Authors

Kim, Sin-Eon
Master student / Dept. of Information Security Management
Chungbuk National University, KOREA
Email: trebrones@gmail.com

Jo, Jung-Tae
Master student / Dept. of Business Data Conversions
Chungbuk National University, KOREA
Email: shllj007@gmail.com

Choi, Sang-Hyun
Professor /Dept. of Management Information System, BK21+ BSO Team
Chungbuk National University, KOREA
Email: chois@cbnu.ac.kr

