Advanced Science and Technology Letters
Vol.76 (CA 2014), pp.29-33
http://dx.doi.org/10.14257/astl.2014.76.08

A Spam Message Filtering Method: focus on run time

Sin-Eon Kim [1], Jung-Tae Jo [2], 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
trebrones@gmail.com, shlla007@gmail.com, chois@cbnu.ac.kr

Abstract. In this paper, we tries to propose light and quick algorithm through with SMS filtering can be performed within mobile devices independently. After introducing this algorithm, it can be the solution for limitations of memory and resources that had not been solved.

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

1 Introduction

As SMS spam messages have drastically increased, typical filtering methods are not effective to be processed within mobile phones anymore. For 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 feature values. We compare the performance of our method with standard feature selection methods; Naive Bayes, J-48 Decision Trees, Logistic. In this paper, we propose a new feature selection method the 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

ISSN: 2287-1233 ASTL

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[1].

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[8].

2.2 Data Mining algorithm

Typical methods to detect spam messages include bayesian classifiers, logistic regression, decision tree and so on. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms[7]. A decision tree is a flowchart-like structure in which internal node represents test on an attribute, each branch represents outcome of test and each leaf node represents class label (decision taken after computing all attributes). A path from root to leaf represents classification rules[2]. Logistic regression is a type of probabilistic statistical classification model. It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable based on one or more predictor variables features[3].

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 is inefficient for dealing with large scale data. In this paper, we propose a simple and efficient feature selection method.

3.1 Proposed Method

This study proposes a VR (Value 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.

$$\overline{W}_j^s = \sum_{i \in spam} w_{ij} \ / \mathrm{k} = w_{ij} \ / \mathrm{k} \qquad (1)$$

$$\overline{W}_j^s = \sum_{i \in ham} w_{ij} \ / \mathrm{k} = w_{ij} \ / \mathrm{k} \qquad (2)$$

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

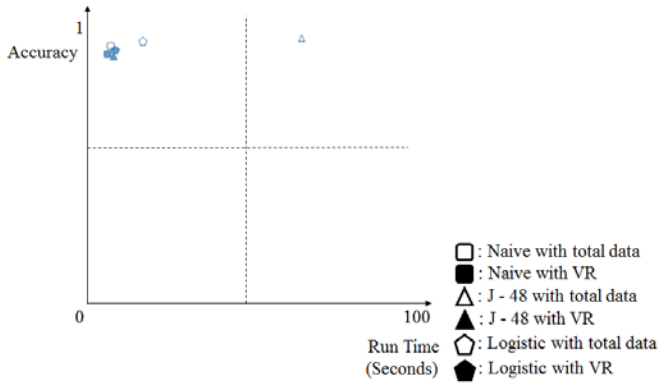
The result of calculating a VA by using calculated $\overline{W}_j^s$ and $\overline{W}_j^h$ value is as below.

$$\mathrm{VR}(j) = \overline{W}_j^s \ / \ \overline{W}_j^h \qquad (3)$$

VR(j) represent the relative ratio of average frequency of jth keyword in spam messages to that in ham messages. As the value of VR(j) is larger, the words are more frequently refered in spam messages.

As shown in the figure, as a result of executing algorithms by using the VR attribute selection technique, run 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 processing capability.

Figure 2. The result of algorithms



4 Future work

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. Gómez Hidalgo, J. M., Bringas, G. C., Sánz, E. P., & García, F. C. (2006). Content based SMS spam filtering. In Proceedings of the 2006 ACM symposium on Document engineering,107-114.
2. http://en.wikipedia.org/wiki/Decision_tree
3. http://en.wikipedia.org/wiki/Logistic_regression
4. http://www.cs.waikato.ac.nz/~ml/weka/

5. Liu H. Setiono R. Motoda H. Zhao Z. (2010). Feature Selection: An Ever Evolving Frontier in Data
6. Mining, JMLR: Workshop and Conference Proceedings, 4-13
7. Saurabh Mukherjeea. Neelam Sharmaa. (2012) Intrusion Detection using Naive Bayes Classifier with Feature Reduction, Procedia Technology, 119-128.
8. SMS Spam Collection v.1 (2012) http://archive.ics.uci.edu/ml/index.html