World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 1, No. 3, 96-99, 2011

A Multi-Phase Feature Selection Approach for the Detection of SPAM

Ahmed Khalid *
Department of computer science
Sudan University of Science and Technology
Khartoum, Sudan
Asalih2@hotmail.com

Izzeldin M. Osman
Department of computer science
Sudan University of Science and Technology
Khartoum, Sudan
izzeldin@acm.org

Abstract- In the past few years the Naïve Bayesian (NB) classifier has been trained automatically to detect spam (unsolicited bulk e-mail). The paper introduces a simple feature selection algorithm to construct a feature vector on which the classifier will be built. We conduct an experiment on SpamAssassin public email corpus to measure the performance of the NB classifier built on the feature vector constructed by the introduced algorithm against the feature vector constructed by the Mutual Information algorithm which is widely used in the literature. The effect of the stop-list and the phrases-list on the classifier performance was also investigated. The results of the experiment show that the introduced algorithm outperforms the Mutual Information algorithm.

Keywords- component; detection; feature selection; Naïve Bayesian classifiers.

I. Introduction

Recently, electronic email has become one of the most effective methods of communication. Spam continues to plague computer users. Sophos research revealed that 92.3 percent of all email during the first quarter of 2008 was spam [9]. As a result, many users of the email must now waste a lot of time dealing with such unwanted messages. Moreover, spam has considerable effect the systems, networks and users. It wastes recourses such as storage space, bandwidth and users time.

Recently automated anti-spam filters have become a familiar method in spam detection [11][1][2]. Most of these filters use the Naïve Bayesian classifier [3][5][6][7]. While such filters are quite effective, we believe that their performance can be improved by more judicious feature selection algorithms and feature weighting.

Most of the previous anti-spam filters built on the Naïve Bayesian classifiers use the Mutual Information algorithm [4] to select their feature vector. To improve the performance of the Naïve Bayesian classifier we introduce a new Multi-Phase Feature Selection Algorithm and we also test a new feature weighting function.

To measure the effectiveness of the introduced approaches experiments were conducted on SpamAssassin email corpus [12. In these experiments we seek to measure the performance of the Naïve Bayesian classifier using the features selected by the Mutual Information algorithm versus the features selected by the Multi-Phase Feature Selection Algorithm. Our investigation also examines the effect of the stop-lists and a phrases list.

The remaining sections of this paper are organized as follows: section 2 discusses the Naïve Bayesian classifier and the Mutual Information algorithm. Section 3 introduces the Multi-Phase Feature Selection Algorithm and presents the feature weighting function. Section 4 presents the SpamAssassin email corpus, the experiment and the discussions of the results. Finally section 5 presents the conclusions.

II. The Naïve Bayesian Classifier

The Naïve Bayesian classifier [6] assumes that each document (in our case each message) is represented by a vector $\vec{x} = < x_1, x_2, x_3, ..., x_n >$ where $\mathbf{x_1, x_2, ..., x_n}$ are the value of the attributes $\mathbf{X_1, X_2, ..., X_N}$ and a class variable C. From the Bayes theorem and the theorem of the total

*Currently at the University of Science and Technology, Omdurman, Sudan probability, it follows that the probability that a message with a vector $\vec{\mathbf{x}} = < \mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, ..., \mathbf{x_n} >$ belongs to a class c is:

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c).P(\vec{X} = \vec{x} \mid C = c)}{\sum_{k \in \{spam,legitimate\}} P(C = k).P(\vec{X} = \vec{x} \mid C = k)} \tag{2.1}$$

The critical quantity in Equation (2.1) is $P(\vec{X} = \vec{x} \mid C = c)$, which is impossible to estimate without simplifying assumptions. The oldest and most restrictive form of these assumptions is embodied in the Naïve Bayesian classifier [6] which assumes that each feature $\mathbf{X_i}$ is conditionally independent of every other feature, given the class variable C. Formally, this yields

$$P(\vec{X} = \vec{x} \mid C = c) = \prod_i P(X_i = x_i \mid C = c) \tag{2.2}$$

This allows us to compute

$$P(C = c \mid \vec{X} = \vec{x}) = \frac{P(C = c).\prod_i^n P(X_i = x_i \mid C = c)}{\sum_{k \in \{spam,legitimate\}} P(C = k).\prod_i^n P(X_i = x_i \mid C = k)} \tag{2.3}$$

$\mathbf{P(X_i \mid C)}$ and $\mathbf{P(C)}$ are easy to estimate from the frequencies of the training corpus. Recently a large number of studies have found that Naïve Bayesian classifier is very effective in spam detections [1] [11][3]. Also! Classifying a legitimate message as spam is generally more severe an error than classifying a spam message as legitimate. Following Androutsopoulos et al. [2], we use $L \rightarrow S$ (legitimate to spam) and $S \rightarrow L$ (spam to legitimate) to denote the two error types, respectively, where $L \rightarrow S$ is λ times more costly than $S \rightarrow L$. We classify a message as spam if the following classification criterion is met:

$$\frac{P(C = spam \mid \vec{X} = \vec{x})}{P(C = legitimate \mid \vec{X} = \vec{x})} > \lambda \tag{2.4}$$

In our case, $P(C = spam \mid \vec{X} = \vec{x})$ =1- $P(C = legitimate \mid \vec{X} = \vec{x})$ and the criterion above is equivalent to :

$P(C = spam \mid \vec{X} = \vec{x}) > t$, with

$$t = \frac{\lambda}{1+\lambda} \quad , \quad \lambda = \frac{t}{1-t} \tag{2.5}$$

Where t is the threshold. As in Sahami et al. [11] experiments we set the threshold t to 0.999, which corresponds to λ =999. This means that mistakenly blocking legitimate message was taken to be as bad as letting 999 spam messages pass the filter [10].

Assuming that $n_{L \rightarrow S}$ and $n_{S \rightarrow L}$ are the numbers of $L \rightarrow S$ and $S \rightarrow L$ errors, and that $n_{L \rightarrow L}$ and $n_{S \rightarrow S}$ count the correctly treated legitimate and spam messages respectively, spam recall (SR) and spam precision (SP) are defined as follows:

$$SR = \frac{n_{S \rightarrow S}}{n_{S \rightarrow S} + n_{S \rightarrow L}} \tag{2.6}$$

$$SP = \frac{n_{S \rightarrow S}}{n_{S \rightarrow S} + n_{L \rightarrow S}} \tag{2.7}$$

Most of the published anti-spam filters built on the Naïve Bayesian classifiers, use the Mutual Information algorithm to select their feature vector[11][1]. The Mutual information $MI(X_i, C)$ of each attribute X_i with the class variable C is computed as follows:

$$MI(X, C) = \sum_{c \in \{spam,legitimate\}} P(X = x, C = c).\log \frac{P(X = x, C = c)}{P(X = x).P(C = c)} \quad [11]$$

The probabilities $P(X, C)$, $P(X)$ and $P(C)$ are estimated from a training corpus as frequency ratios. The features with highest MI are selected as the feature vector from which the classifier is built.

III. The Multi-Phase Feature Selection Algorithm

The Multi-Phase feature selection algorithm assumes there is a training set of messages from two classes c1 and c2 (in our case spam and legitimate). It constructs two types of features, common and rare features.

$\mathbf{x_i}$ Belongs to common features if $\mathbf{x_i} \in \mathbf{c_1} \cap \mathbf{c_2}$,
Whereas $\mathbf{x_i}$ belongs to rare features if
($x_i \in c_1 \cap x_i \notin c_2$) or ($x_i \in c_2 \cap x_i \notin c_1$).
The common features are constructed from $\mathbf{V_{c1}}$ **and** $\mathbf{V_{c2}}$ where:

$$V_{c1} = < x_1, x_2, ..., x_n > \quad where \quad p(x_i, c_1) > t \\ and \quad p(x_i, c_2) < t, \quad i = 1....n, \quad t < 0.05$$

$$V_{c2} = < x_1, x_2, ..., x_n > \quad where \quad p(x_i, c_1) < t \\ and \quad p(x_i, c_2) > t, \quad i = 1....n, \quad t < 0.05$$

For each $\mathbf{x_i}$ in the rare features we compute

$$f(x_i, c\ ) = \frac{n + p(x_i, c\ )}{n + \max(p(x_i, c))} \qquad (3.1)$$

Where $f(x_i, c)$ is degree of belief about whether, when we see the word x_i again, it will be in class c. In our experiment we set $f(x_i, c_1) = 1 - f(x_i, c_2)$.

$\mathbf{x_i}$ s with the highest $\mathbf{f(x_i, c_1)}$ and lowest $\mathbf{f(x_i, c_2)}$ are selected and added to the common features to construct the feature vector from which the classifier is built

IV. An Experiment with SpamAssassin corpus:

To measure the efficiency of the Multi-Phase feature selection algorithm we conducted an experiment using the Naïve Bayesian classifier on SpamAssassin public email corpus [12]. In this experiment we seek to measure the performance of the naïve Bayesian classifier using the features selected by the Multi-Phase Algorithm against the features selected by the Mutual Information algorithm.

The corpus consists 1800 messages (1200 are spam and 600 are legitimate). This corpus is split into 1000 messages as a training set (700 of which are spam) and 800 messages as a testing set (500 of which are spam). Our experiment includes a word stemming that returns each word to its base form (e.g. "attaching" becomes "attach"), each capital letter is converted to its corresponding small letter and the following characters { ; , = | .< > : + - _ ( ) & ^ % # [ ] 1 2 3 4 5 6 7 8 9 0} are removed from the beginning and the end of the word. We also test the existence and the absence of 150 stop words, like (the, to, from …) and a phrases-list.

We select the best 500 features by each algorithm as the feature set from which to build a classifier. As in Sahami et al. [11] experiments, we set the threshold t to 0.999, which corresponds to λ =999. This means that mistakenly blocking a legitimate message was taken to be as bad as letting 999 spam messages pass the filter.

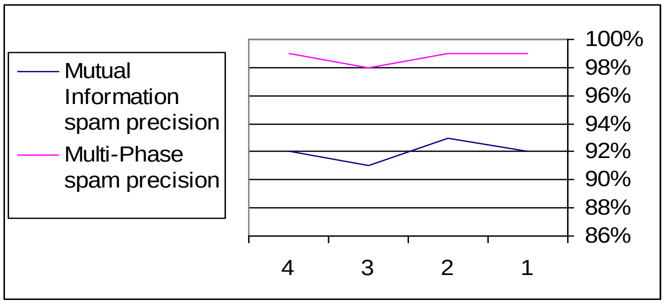

Figure1: Spam precision for the NB classifier using Multi-Phase and Mutual information algorithms

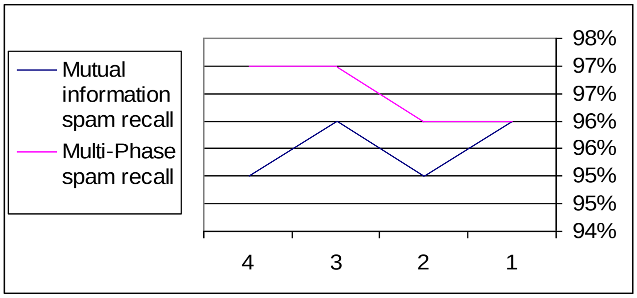

Figure2: Spam recall for the NB classifier using Multi-Phase and Mutual information algorithms

TABLE I. Results on SpamAssassin corpus using 500 attributes selected by the two algorithms (1800 total messages, 60% spam)

		Multi-phase algorithm			
	λ	***Legitimate***		***spam***	
		recall	precision	recall	precision
features only	999	94%	98%	96%	99%
features + stop words	999	99%	99%	96%	99%
features + phrases	999	97%	97%	97%	98%
features + stop words +phrases	999	98%	98%	97%	99%

TABLE II. Results on SpamAssassin corpus using 500 attributes selected by the two algorithms (1800 total messages, 60% spam) for Mutual Information algorithm

		Mutual Information algorithm			
Feature Regime	λ	***Legitimate***		***spam***	
		recall	precision	recall	precision
features only	999	86.0%	93%	96%	92%
features + stop words	999	88.0%	91%	95%	93%
features + phrases	999	85%	93%	96%	91%
features + stop words +phrases	999	87%	93%	95%	92%

Figure 1 and 2 shows that the NB classifier achieved impressive spam recall and precision using the two algorithms. The results as shown in Table 1 and 2 shows that the Multi-Phase algorithm outperforms the Mutual Information algorithm in spam precision, spam recall, legitimate precision and legitimate recall. Our experimental results confirm Androutsopoulos, et al [1] experiment conclusion that the addition of the stop-list does not seem to have any noticeable effect on the classifiers performance. This is because the two algorithms rarely pick words that are so common as those of the stop-list.

V. CONCLUSIONS

Since the spammers always search for the ability to circumvent known anti-spam filters, the objective of this paper is to improve the performance of the Naïve-Bayesian classifier [5] through introducing a features selection approach. It is clear form the experiment results the introduced algorithm outperforms the Mutual Information algorithm. The results also confirm that there is no effect for adding stop words.

REFERENCES

[1] Androutsopoulos,I. Koutsias, J. Chandrinos, K.V. Spyropoulos, C. D. "An Experimental Comparison of Naïve Bayesian and Key-Based Ant-Spam Filtering with Personal Email Messages" in Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval July 24,2000, pp 160-167.

[2] Androutsopoulos, I. Koutsias, J. Chandrinos, K.V. Paliouras,G. and Spyropoulos, C.D. "An Evaluation of Naïve Bayesian Ant-Spam Filtering". 11th European Conference on Machine Learning,2000, pp 9-17

[3] Aris Kosmopoulos, Georgis Paliouras, Ion Androutsoploulos, “adaptive Spam Filtering Using Only Naïve Bayes Text Classifiers”, CEAS 2008- Fifth Conference on Email and Anti-Spam, August 21-22, 2008,pp

[4] Cover, T. M. and Thomas, J. A. " Elements of Information Theory ", Wiley, 1991,

[5] Duda R.O. and Hart, P.E. "Bayes Decision Theory" Chapter 2 in pattern classification and scene analysis. John Willey, 1973

[6] Elkan, Charles "Naïve Bayesian Learning" Technical Report No. CS97-557, Department of Computer Science and Engineering, University of California, San Diego, September 1997.

[7] Good, I.J. "The Estimation of Probabilities, An Essay on Modern Bayesian Methods", M.I.T. Press, 1965.

[8] Goodman, Joshua, Cormack Gordon V. and Heckerman David "Spam and the Ongoing Battle for the Inbox", CACM Februarys 2007, vol. 50, No. 2, pp 24-33

[9] www.net.security.org/secworld.php?id=6056 ., visited on 2nd March 2009

[10] MessageLabs,2005,http://www.messagelabs.co.uk/published content/publish/threat-watch-dotcom-en/threat-statistics/spam-intercepts/DA-114633.chp.html/

[11] Sahami, M., Dumais, S., Heckerman, D. and Horvitz, E. "A Bayesian Approach to Filtering Junk E-Mail" in Learning for Text Categorization workshop. AAAI Technical Report WS-98-05, 1998, pp 55-62.

[12] www.spamassassin.apacheorg/publiccorpus, visited on may,2008