A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification

Article *in* Research Journal of Applied Sciences, Engineering and Technology · January 2014

DOI: 10.19026/rjaset.7.299

Assoc Prof. Dr. Othman Ibrahim
Universiti Teknologi Malaysia
134 PUBLICATIONS **2,083** CITATIONS

Ahmed Hamza Osman
King Abdulaziz University
32 PUBLICATIONS **344** CITATIONS

Research Journal of Applied Sciences, Engineering and Technology 7(3): 625-638, 2014
DOI:10.19026/rjaset.7.299
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 21, 2013 Accepted: July 09, 2013 Published: January 20, 2014

Research Article

A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification

Nadir Omer Fadl Elssied, Othman Ibrahim and Ahmed Hamza Osman
Faculty of Computing, University Technology Malaysia, 81310, Skudai, Johor Bahru, Malaysia

Abstract: Spam is commonly defined as unwanted e-mails and it became a global threat against e-mail users. Although, Support Vector Machine (SVM) has been commonly used in e-mail spam classification, yet the problem of high data dimensionality of the feature space due to the massive number of e-mail dataset and features still exist. To improve the limitation of SVM, reduce the computational complexity (efficiency) and enhancing the classification accuracy (effectiveness). In this study, feature selection based on one-way ANOVA F-test statistics scheme was applied to determine the most important features contributing to e-mail spam classification. This feature selection based on one-way ANOVA F-test is used to reduce the high data dimensionality of the feature space before the classification process. The experiment of the proposed scheme was carried out using spam base well-known benchmarking dataset to evaluate the feasibility of the proposed method. The comparison is achieved for different datasets, categorization algorithm and success measures. In addition, experimental results on spam base English datasets showed that the enhanced SVM (FSSVM) significantly outperforms SVM and many other recent spam classification methods for English dataset in terms of computational complexity and dimension reduction.

Keywords: Feature selection, machine learning, one-way ANOVA F-test, spam detection, SVM

INTRODUCTION

With the rapid development of the internet, electronic mail (e-mail) has grown to be an important information source for various purposes (Wang and Cloete, 2005; Wang *et al.*, 2006). Automatic e-mail classification becomes very critical for dealing with massive data and the major problem of e-mail classification is the high dimensionality of feature space (Shang *et al.*, 2007). A long with the growing use of e-mail lead to occurrence many problems caused by unsolicited bulk e-mail messages, referred to as Spam (Guzella and Caminhas, 2009). Spam is undesirable message and unwanted commercial e-mail appearing in e-mail and become a global threat against e-mail users (Nazirova and Alguliyev, 2012; Yang *et al.*, 2006). Recently, the number of undesirable messages coming to e-mail has strongly increased (Alguliyev and Nazirova, 2012). According to Symantec Intelligence Report in September 2012 the percentage of spam in e-mail traffic was increased by 2.7 percentage points from August and averaged 75%, in addition to Kaspersky Lab annual report the total amount of spam in mail traffic was 78.5% (Bulletin, 2012; Wood, 2012). There are several critical problems connected with increasing volumes of spam mail such as filling users' mailboxes, wasting storage space, network bandwidth and consuming users' time to delete all spam messages (Lai and Wu, 2007). In spam messages a big number of messages is include a link to websites for advertised and spammer are focused to send out millions of copies of the same e-mails and try to hide form spam filters (Sirisanyalak and Sornil, 2007). As problem sizes continue to scale up with the explosive growth of using e-mail, essential research is required to enhance the classification efficiency and effectiveness (accuracy) (Forman, 2003; Sanasam *et al.*, 2010). E-mail classification is supervised learning problem and classification is very important methods to cancel this problem of spam. Recent research shows that spam classification is usually processed by statistical theory and Machine Learning (ML) algorithms, to differentiate between non-spam and spam e-mail (Fagbola *et al.*, 2012; Guzella and Caminhas, 2009; Saad *et al.*, 2012). ML methods are able to extract the knowledge from a group of emails supplied and using the gained information in the categorization of newly received e-mail (Saad *et al.*, 2012). The aim of ML is to improve the performance of the computer program through experience so as to make best decisions and solve problems in an intelligent way by using illustration data (Salehi and Selamat, 2011).

In the field of ML, Feature Selection (FS) is an importance topic to select a subset of features among the full features and then reduce the high data dimensionality, lead to show the best performance in classification accuracy (Yun *et al.*, 2007). Many practical applications of e-mail classification involve a

Corresponding Author: Nadir Omer Fadl Elssied, Faculty of Computing, UTM, 81310, Skudai, Johor Bahru, Malaysia

massive number of data and/or a large number of features (Unler *et al*., 2011). Usually, the performance of feature selection algorithms has been measured by comparing the performance of classification algorithms before and after feature selection. They are several types of feature methods such as feature extraction, feature selection and feature weighting. Furthermore, feature extraction is to extracts the asset of new features from the original feature into a distinct feature space or transforming arbitrary data such as text or images, into numerical features usable for machine learning (Uğuz, 2011). Application of a Feature Selection (FS) algorithm to classify the dataset that can improve the performance of some classifiers. FS is commonly used to reduce the dimensionality of the feature space and to enhance the efficiency and effectiveness of classification (Chen *et al*., 2009). It is the identification of the subset or subsets of features that best satisfy some specified criteria. The importance it comes from its ability of improving learning performance (Tu *et al*., 2007). This means that, through feature selection, which can reduce the cost of learning by reducing the number of features for learning, removes irrelevant, noisy and redundant data and provide a better learning accuracy compared to using full feature sub set (Tu *et al*., 2007). FS has two common feature techniques namely feature synthesis and feature subset selection. Furthermore, feature selection has many evaluation function (measurements function) such as Information Gain (IG), Term Frequency (TF), Pearson Chi-square (X^2), odd ratio, Gini index (Gini coefficient), Mutual Information (MI), expected cross entropy, Term Strength (TS) and Document Frequency (DF). According to Yang and Pedersen work, the IG and X^2 to be most effective than others and X^2 is better in accuracy measure (effectiveness) (Méndez *et al*., 2006; Yang and Pedersen, 1997). In their work Mendez *et al*. told MI method is clearly disadvantaged when it is applied to e-mail classification. Because spammers obfuscate terms introducing noise into messages.

Currently, SVM is one of the most popular algorithms for e-mail classification and obtain a maximum margin of a hyperplane in order to enhance the obtained support vector machine (Sun *et al*., 2010). However, in many cases it takes long processing time and provides less accuracy rate for classification due to huge data (Fagbola *et al*., 2012; Morariu *et al*., 2006). The result of the study by Priyanka *et al*. (Chhabra *et al*., 2010) on SVM for massive data classification showed that SVM takes time consuming when the size of data is massive. A major problem in SVM, it weak to classify highly dimensional dataset with the large number of features (Chen *et al*., 2009). This study, propose a novel spam detection scheme by using a combination of Feature Selection (FS) based on one-way ANOVA F-test statistics and Support Vector Machine (SVM) to reduce the high data dimensional of the feature space, enhancing the computational time complexity of SVM and to improve the accuracy of e-mail spam detection scheme. The novelty of this study is used one-way ANOVA F-test statistics as prediction techniques to measure similarity for relevant feature and to reduce the high data dimensionality of feature space by identifying the important features and it primary goals are improving the computational time complexity or classification accuracy or both. A comparative study has been carried out between spam detection using SVM and using the combination of FS and SVM. Experimental results show that this algorithm has high classification accuracy and time efficiency. The remaining of this study is organized as follows. Section two discusses the related work of spam detection. Section three describes the proposed method. Section four presents the dataset. Finally, the study describes the experiment and discussions.

LITERATURE REVIEW

Over the past years, there are many algorithms to select the important features, removing the irrelevant and redundant features and classify spam mails such as Support Vector Machine (SVM), Particle Swarm Optimization (PSO), Naïve Bayesian (NB) and Feature Selection algorithm (FS) (Chhabra *et al*., 2010; Golovko *et al*., 2010; Ma *et al*., 2009; Mohammad and Zitar, 2011; Salcedo-Campos *et al*., 2012; Wu *et al*., 2008). Most of researchers using feature selection algorithm for removing the irrelevant or redundant features to avoid processing overhead and reduce the high data dimensional before classification phase. In his study Wang (Wang *et al*., 2005) integrating feature selection based on Genetic Algorithm (GA) and SVM based on Structural Risk Minimization (SRM) for detecting e-mail as spam or non-spam. GA approach is adopted to select features that are most suitable to SVM classifier. Spam base benchmark dataset is carried out for this study. The experimental results based SVM and 57 features for training were 94.38% of accuracy and for testing was 87.73% while the result based on GA-SVM and 52 selected features for training was 94.43% of accuracy and for testing was 87.89%. The result is achieved better classification results comparing with original SVM classifier. Also, in their work (Wang *et al*., 2006) suggested feature selection based on Information Gain (IG) and SVM based on Radial Basis Function (RBF) to reduce the high dimensionality of the instance space and to solve the problem of spam message that enters e-mail box every day. Ling spam dataset is implemented in this study and it composed of 2412 non spam and 481 spam messages. The experimental result for testing was 94.75% of accuracy and show that the proposed method outperforms other conventional spam classification methods such as Naïve Bayes (NB) and Neural Network (NN). In their study

(Lee *et al*., 2010) proposed spam detection model using Random Forest (RF) based on feature selection and parameters optimization simultaneously to reduce processing overheads with guaranteeing high detection rates. They used spam base benchmark dataset for experiments and the result are summarized as optimizing the parameters of RF, identifying main features as a numerical value and detecting spam with low processing overheads and high detection rates and the result of accuracy was 95.4. In their work (Chen *et al*., 2009) developed feature section based on two evaluation metrics namely Multi class Odd Ratio (MOR) and Class Discriminating Measure (CDM) with Naïve Bayes to improve spam classification. He applied two dataset for experiment the first one is composed of 9779 messages are divided to 7053 for training and 2726 for testing and the second is contains 2816 messages divided into 1882 for training and 934 for testing. The result for the first dataset was 70.77% of accuracy using CDM while 84.95% of accuracy using MOR. The second dataset result is 72.59% of accuracy using CDM and 70.77 of accuracy using MOR. In their work (Fagbola *et al*., 2012) used GA to select a set of good finite feature subset and used SVM as classifier to solve the inefficient in SVM about high dimensionality when it deal with massive email dataset because it consumes a lot of computational time. He applied the spam assassin dataset for experiment to improve the classification accuracy and computational time. The result of accuracy was 93.5% and the computational time was 119.562 Sec while using GA-SVM and the result of accuracy was 90% and the computational time are 149.984 sec while using SVM only. Then, in his study Uysal (Uysal and Gunal, 2012) presented a novel probabilistic feature selection method namely Distinguishing Feature Selector (DFS) and Support Vector Machine (SVM) for text classification to solve the problem of high dimensionality due to processing time and accuracy considerations. The result was 71% of accuracy while the time cost was 79.9 sec. Experimental results explicitly indicate that DFS outperforms recent spam detection in terms of classification accuracy, dimension reduction rate and processing time. in their study (Liang *et al*., 2008) presented a new Feature Selection algorithm based on a Distance Discriminant namely (FSDD) and Support Vector Machine (SVM) to solve the problem of computational complexity due to processing time and accuracy considerations. The result was 85% of accuracy while the time cost was 62.0 sec. Experimental results explicitly indicate that FSDD outperforms recent spam detection in terms of classification accuracy, dimension reduction rate and processing time. In his work (Zhu, 2008) developed a novel approach based on rough set and SVM to improve the classification level. Rough set is mathematical tools for dealing with uncertainty and it has ability to compute the reduction of high dimension. In his study rough set is used as feature selection to reduce the number of feature or attribute and SVM as a classifier. The experimental result using spam base standard dataset was 93.6 of accuracy when using SVM before using rough set for feature reduction and 94.6 of accuracy when using SVM and rough set as feature reduction. In spite of numerous methods in the related work, feature selection is still ongoing area. Researchers are still looking for new techniques to select distinctive attributes or features so that the categorization accuracy can be enhanced and the processing time can be reduce as well. Finally, from all above studies still there are shortcoming regarding to the problem of enhance SVM and improve both of efficiency and effectiveness of e-mail classification. These study implement combination of Feature Subset selection (FS) based on one-way ANOVA F-test statistics and SVM based on normalized poly kernel approach to enhance SVM in term of increase the percentage of spam detection and reduce the data dimensionality.

PROPOSED METHODOLOGY

In this Section, we present our proposed work to meet the objective of this scheme. Mainly one-way ANOVA F-test as a feature selection and SVM based on normalized poly kernel as classifier is considered to classify the spam base dataset are shown in Fig. 1. The input dataset is partitioned into 70% for training dataset and 30% for testing dataset after feature selection. A classifier will be induced from the training data and applied into the testing data to obtain a performance percentage. In the training phase, we adopt SVM to build the classification scheme and with testing phase e-mail classification are carried out by the proposed scheme and classify each e-mail can obtain.

Feature selection method based on one-way ANOVA F-test statistics is carried out to skip unimportant attributes from the dataset. The proposed scheme of this study decreased 57 features of spam base dataset to 52 features in order to discard unrelated feature to avoid a high data dimensionality and low classification accuracy using one-way ANOVA F-test statistics as a feature selector. Furthermore, the aim of feature selection is to find the best subset consisting of m features chosen from the total n features. After skipping unimportant attributes the dataset is transformed to the classifier system, which then divided into two parts namely training and testing data. In this study, we use Support Vector Machine (SVM) as the classifiers. The SVM based on normalized poly kernel, selects key data points as its support vectors and uses these support vectors for prediction. A comparative analysis is carried out on the performances of classifiers before and after feature selection. After that applied

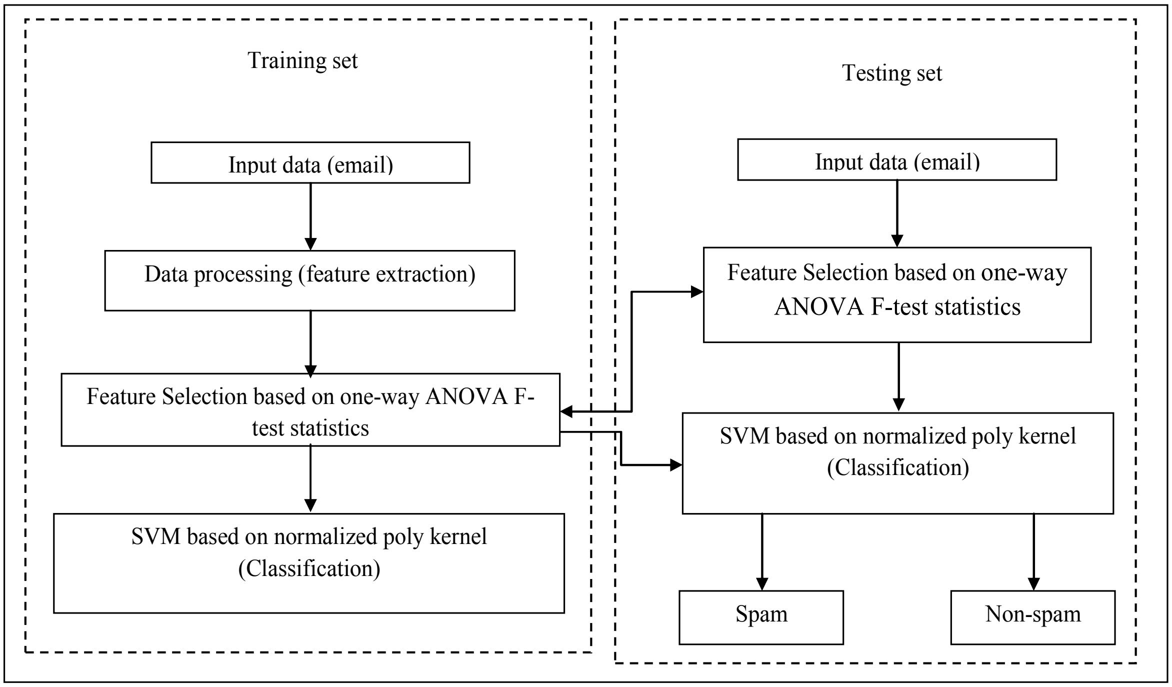
Training set
Input data (email)
Data processing (feature extraction)
Feature Selection based on one-way ANOVA F-test statistics
SVM based on normalized poly kernel (Classification)
Testing set
Input data (email)
Feature Selection based on one-way ANOVA F-test statistics
SVM based on normalized poly kernel (Classification)
Spam
Non-spam

Fig. 1: The methodology of proposed method

Table 1: The advantage and disadvantage of feature synthesis and feature subset selection

Feature type	Advantage	Disadvantage
Feature synthesis selection	Change the original features to new feature set which is low dimensional and ideally consist of uncorrelated feature from the original feature set.	Numerical coefficients of the new feature are not informative about the importance of individual features in the original set.
Feature subset selection	Reduce the computational effort by selected features related to the performance of classification system.	The prediction accuracy of the selected subset depends on the size of the subset as well as the feature selected.

SVM based on normalized poly kernel approach to the classifier and the steps of proposed method Fig. 1 demonstrates feature selection and SVM phases and activities of the scheme.

Feature Selection (FS): Feature selection is also referred to as variable subset selection, variable selection, feature reduction or attribute selection. Feature Selection (FS) is the technique that can be used to remove features in the training documents so as are statistically uncorrelated with class labels and it reduce the set of features to be used in classification (Parimala and Nallaswamy, 2011). FS is helps to reduce the dimensionality and to improve the efficiency and accuracy of e-mail classification (Chen *et al*., 2009). It identifies the fields that are most important in predicting a certain outcome. Thus, feature selection is several times applied to reduce the amount of training data needed to obtain a desired level of performance and improve the classification accuracy (Forman, 2003). Moreover, feature selection is frequently used as a preprocessing step to machine learning based on statistical theory (Sanasam *et al*., 2010). Furthermore, it a process of choosing a subset of original features so that the feature space is optimally reduced according to a certain evaluation criterion (Parimala and Nallaswamy, 2011). In addition, feature selection plays an important role in spam e-mail classification to speed up the computation as well as to enhance both efficiency and effectiveness (Alper and Gunal, 2012). The advantage of feature selection are using fewer predictors is less expensive and computing time is improved. The predictors which contribute less in prediction can be skipped from the dataset. Ultimately, it ends up with a quicker, more efficient model that uses fewer predictors, executes more quickly and may be easier to understand. To select the importance feature there are many feature ranking or feature selection techniques have been proposed such as likelihood ration, Cramer's v, information gain, Pearson‘s Chi-square, mutual information, term frequency and deviation from Poisson (Sanasam *et al*., 2010).

Feature selection has two commonly techniques namely feature synthesis and feature subset selection, each one of them has advantages and limitation or drawback see Table 1. Also, the data can be either labeled or not, leading to the development of supervised and unsupervised feature selection methods. Supervised

Table 2: The different between wrapper and filter

Wrapper	Filter
- It is high accuracy, but it takes much time if the number of feature is high	- It time efficiency is high, but it takes low accuracy
- It not suitable for text classification	- Widely used for text classification
- To evaluate the feature it use train classification for each feature subset	- It uses metrics or measures such as IG, TF and chi-square to evaluate the feature
- It is used to select the desired number of features	- It uses to reduce the number of features

feature selection determines relevant features by their relations with the corresponding class labels and discards irrelevant and redundant features. Unsupervised feature selection explores data variance to choose relevant features. For supervised feature selection, the existing feature subset selection can be classification into two methods namely wrapper and filter algorithm (Liang *et al*., 2008). The main difference between the two methods is that filter algorithms select the feature subset before the application of any classification, using statistical properties or probabilistic characteristics of feature in filter to remove the less important feature from the subset based on their comparative computational efficiency, filter methods define the relevant feature without prior classification of data and are independent of the learning machines. The wrapper method approach using machine learning and selecting the feature subset according to the accuracy on the training data and testing the classification (based on the prediction performance) using the test data, wrapper methods generally is not suitable for text classification because it is time consuming the compare with filter, filter has much less time consuming and widely used for text classification (Unler and Murat, 2010; Unler *et al*., 2011). See Table 2 illustrates the different between wrapper and filter.

In this study, Feature selection algorithm was applied to recognize important features showing strong correlation with e-mail spam classification. The algorithm considered one attribute at a time to see how well each predictor alone (feature) predicted the target variable (output). The importance value of each variable was then calculated as (1-p) where p was the p value of the appropriate test of association between the candidate predictor and the target variable. The association test for categorized output variables was different from the test for continuous ones. When target value was continuous, p values based on the F statistic were used. If some predictors were continuous and some categorical in the dataset, the criterion for continuous predictors was still based on the p value from a transformation and that for categorical predictors from the F statistic. Predictors were ranked according to sorting by p value in ascending order. If ties occurred, the rules for breaking ties were followed among all categorical and continuous predictors separately and then these two groups (categorical predictor group and continuous predictor group) were sorted by the data file order of their first predictors. The predictors were then labeled as ‘important’, ‘marginal’ and ‘unimportant’ with values above 0.998, between 0.998 and 0.997 and below 0.997, respectively. The important of a feature indicates that the feature is always necessary for an optimal subset; it cannot be removed without affecting the original conditional class distribution, unimportant indicates that the feature is not necessary and the performances are compared against the performances of the classifiers before feature selection:

p-value based on the F-statistic: The goal was to perform a one-way ANOVA F-test for each continuous predictor, which tests whether or not all the different classes of Y have the same mean as X. The following notation applies:

N_j = The number of cases with Y = j
$\bar{x}_j$ = The sample mean of predictor X for target class Y = j
s_j^2 = The sample variance of predictor X for target class Y = j:

$$s_j^2 = \sum_{i=1}^{N_j}(x_{ij} - \bar{x}_j)^2/(N_j - 1)$$

$\bar{\bar{x}}$: The grand mean of predictor X:

$$\bar{\bar{x}} = \sum_{j=1}^{J} N_j\bar{x}_j/N$$

The above notations are based on non-missing pairs of (X, Y).

Then, the p value based on the F statistic is calculated by p value = Prob {F (J-1, N-J) >F}:
where,

$$F = \frac{\sum_{j=1}^{J} N_j(\bar{x}_j - \bar{\bar{x}})^2/(J-1)}{\sum_{j=1}^{J}(N_j - 1)s_j^2/(N-1)}$$

F (J-1, N-1) is a random, variable that follows an F distribution with degree of freedom J-1 and N-J. If the denominator for a predictor is zero, set the p-value = 0 for the predictor. Predictor is ranked by sorting according to the p value in ascending order. If ties are occur, sort by F in descending order and if it still ties sort by N in descending order. Classification of features indicated that out of 57 features, 52 features were the most important features related to the email spam detection.

Support vector machines: Support Vector Machines (SVMs) are relatively new methods that have quickly gained popularity because of the appropriate results that have been achieved in a wide variety of machine

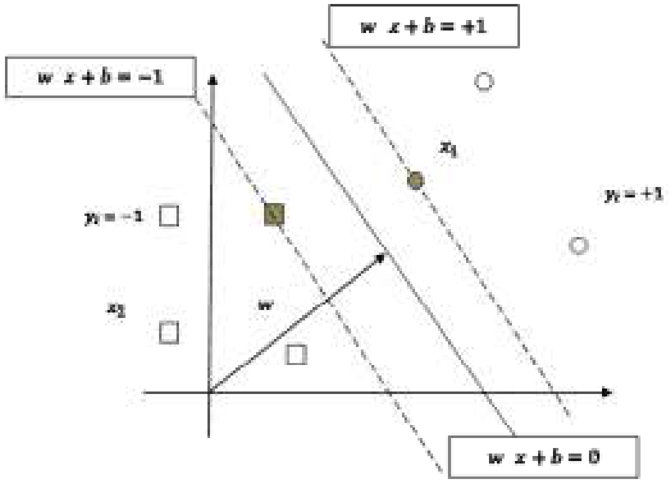

Fig. 2: Classification of spam using SVM

learning problems and because they have solid theoretical underpinnings in statistical learning theory (Long *et al*., 2011; Salcedo-Campos *et al*., 2012; Ying *et al*., 2010). SVM is a Binary classification technique based on statistical learning theory that was applied with great success in many challenging non-linear classification problems and on large datasets (Morariu *et al*., 2006; Noble, 2006). Binary classification has many advantages such as provide a limited problems space that is easier to analyze mathematically and it is convenient to use for classification data (Ji and Dasgupata, 2004). It can be used to solve Linearly separable as well as non-linear separable problems (Chhabra *et al*., 2010; Fagbola *et al*., 2012).

The SVM is a supervised learning method that generates input-output mapping functions from a set of labeled training data (Wang, 2005). Before the discovery of SVMs, machine learning was not very successful in learning and generalization tasks, with many problems being impossible to solve (Youn and McLeod, 2007). There are many kernel-based functions such as linear kernel function, the normalized poly kernel, polynomial kernel function, Radial Basis Function (RBF) or Gaussian Kernel and Hyperbolic Tangent (Sigmoid) Kernel sigmoid function can be implemented in SVM (Chhabra *et al*., 2010). In this study for classification, normalized poly kernel are used to transform input data to a high-dimensional feature space in which the input data become more separable compared to the original input space. Maximum-margin hyperplanes are created and SVM algorithm divide the n-dimensional space representation of the data into two regions using a hyperplane (Youn and McLeod, 2007). Hyperplane is a concept in geometry and in n-dimensional space the hyperplane is a space with the dimension n-1. The produced model depends only on a subset of the training data near the class boundaries. SVM has many advantages such as obtaining the best result when deal with the binary representation, able to dealing with low number of features (Xiao-Li *et al*., 2009). In addition, SVM using statistical learning method and lead to good performance without the need to incorporate prior information, very effective in text classification field because it has the ability to handle high dimensional data by using kernels and it can also use large input data and feature set. Furthermore, it is easy to test the influence of the number of features on classification accuracy, SVM more robust to different dataset and pre-processing procedure and much more for efficient for training and able to deal with supervised and unsupervised learning (Guzella and Caminhas, 2009; Jin and Ming, 2011; Marsono, 2007). It has high accuracy for binary classification, but has the level of misclassification such as non-spam message is very high that mean get low precision (Xiao-Li *et al*., 2009). SVM has a number of disadvantages such as require longer learning time, time and memory consuming when the size of data is huge and training time can be very large if there are the large number of training example (Chhabra *et al*., 2010; Fagbola *et al*., 2012).

SVM classifiers utilize the hyperplane to separate classes. Every hyperplane is characterized by its direction (w), (b) is the exact position in space or a threshold, (x_i) is the input vector of dimension N or text content and indicates the class. Thus, a set of labeled training examples:

$$(x_1, y_1), (x_2, y_2), \dots . (x_k, y_k)$$

$X \in R^d$ where d the dimensionality of the vector is; $y_i \in \{-1, +1\}$; $i = 1, 2, \dots , k$

We consider decision function of the form $f(x, w, b) = \mathrm{sgn}\ ((w.x_i) + b),\ w \in R^d,\ b \in R$.

Then, the region between the hyperplane if and if, which separates two classes, is called as the margins Fig. 2 illustrates the classification of spam using SVM.

Width of the margin is equal to $\frac{1}{2}\|w\|$ and get the maximum possible margin is the underlying idea of SVM algorithm. Maximization of the margin requires minimization of:

$$f(w, b) = \frac{1}{2}\|w\|^2$$

This is subject to:

$$wx_i + b \gg 1 \text{ , if } y_i = 1$$

$$wx_i + b \ll 1 \text{ , if } Y_i = -1$$

K is a user defined constant and ε is the margin error. Margin error occurs if data belonging to one class is on the wrong side of the hyperplane. Minimizing the cost is therefore a trade-off issue between a large margin and a small number of margin errors. Solution of this optimization problem is obtained as:

$$w = \sum_{i=1}^{N} \lambda_i \gamma_i \chi_i$$

This is the weighted average of the training features. Here, λ_i is a Lagrange multiplier of the optimization task and γ_i is a class label. Values of $\lambda's_i$ are non zero for all the points lying inside the margin and on the correct side of the classifier.

Experiment dataset: There are various benchmark datasets available for researchers related to e-mail classification (Chhabra *et al*., 2010). There has been significant effort to generate public benchmark datasets for the anti-spam classification. One of the main concerns is how to protect the privacy of the users whose non-spam messages are included in the datasets. The first approach is to use non-spam messages collected from mailing lists with public archives. There are many examples of dataset such as:

- Ling-Spam
- The Spam Assassin
- The Spam base

To develop and test email spam detection, a dataset containing both non-spam and spam email is required. Furthermore, certain metrics are also required to evaluate the performance of the system. Both problems are discussed below to clarify the experimental framework considered in the study (Salcedo-Campos *et al*., 2012). Machine learning repository has been used for the datasets, the center of machine learning and intelligent system for classifying e-mails as spam and non-spam. The spam base dataset collection is composed of 4601 emails, the dataset are divided to 1813 emails (39.4%) marked as spam while the non-spam dataset has 2788 emails (60.6%) was proposed by Mark Hopkins *et al*. (Year). Analyze the spam and the non-spam ratio as demonstrated in Fig. 3. '0' is representing as non-spam while '1' is representing as spam. Generally the dataset was divided into training and testing, training refers to building a new scheme by using historical data and training is done on a large proportion of the data available while testing refers to trying out the scheme on new and testing is done on the small proportion of the total data. This dataset was divided into two classes training and testing dataset which were divided in the ratio of 70 and 30%, respectively. Use the combination of Feature Selection (FS) based on one-way ANOVA F-test and Support Vector Machine (SVM) for trained classifiers by using the spam vector and the non-spam vector to detect the testing sample.

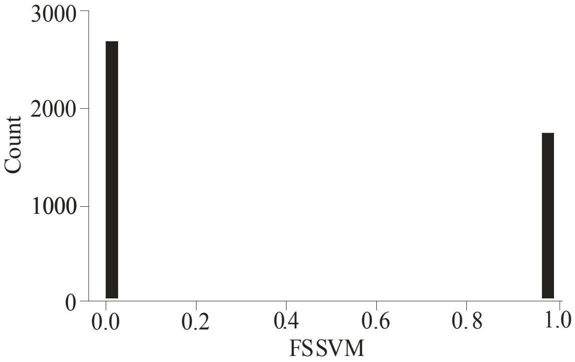

Fig. 3: Dataset analysis

A performance index was used for feature selection and SVM to verify the effectiveness and efficiency of the proposed approach. The parameter for Feature Selection (FS) and support vector machine that are to be used in this experiment is considered as a constant change optimization process carried out by SVM algorithm. Each partition use 70% as the training data and 30% as the testing data using Feature Selection (FS) and SVM as the classification algorithm. This study used Feature Selection (FS) to skip unimportant attributes from the dataset, then the output of feature selection used as input of classification using SVM.

EXPERIMENTAL AND DISCUSSION

This section explains the effectiveness (accuracy) and efficiency (time-cost) results, compares this with others for training and testing results used in the experiments. The used of statistical testing for the significance of the study and the evaluation metrics for the accuracy and error rate (misclassification). The programming tool used to implement the algorithms is MATLAB. This is because MATLAB is a very powerful computing system for handling calculations involved in scientific and engineering problems. With

Table 3: Illustrates how the false positive and false negatives are calculated

	Non-spam	Spam
Non-spam	True Negative (TN)	False Positive (FP)
Spam	False Negative (FN)	True Positive (TP)

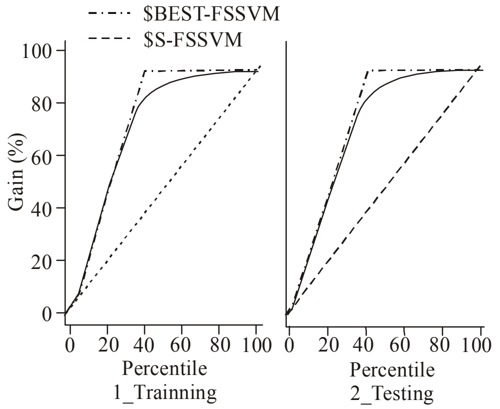

Fig. 4: Training and testing result for SVM before using feature

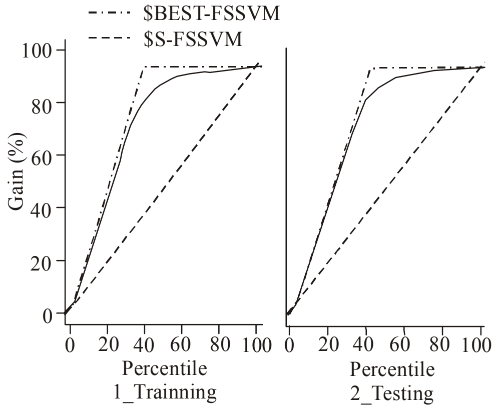

Fig. 5: Training and testing result for SVM after feature selection (include best line)

MATLAB, computational and graphical tools to solve relatively complex science and engineering problems can be designed, developed and implemented. Specifically, MATLAB 2007b was used for the development. Table 3 illustrates how the false positive and false negatives are calculated. The first row displays the total non-spam. That is divided to True Negative (TN) and false positive. The second row is the same as a first row. This row contains the total spam in data sets. Some of them are dedicated wrongly as non-spam and the rest of them correctly dedicated as spam. We provide two general testing parameters or function measurement that are commonly used in spam detection as following see Eq. (1) to (5):

$$\text{Accuracy rate} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} * 100 \qquad (1)$$

$$\text{Error rate} = \frac{\text{FP+FN}}{\text{TP+TN+FP+FN}} * 100 \qquad (2)$$

$$\text{FP} = \frac{\text{FP}}{\text{FP+TP}} * 100 \qquad (3)$$

$$\text{FN} = \frac{\text{FN}}{\text{FN+TN}} * 100 \qquad (4)$$

$$\text{F} - \text{Measure} = \frac{2{*}\text{Precision}{*}\text{Recall}}{\text{Precision+Recall}} \qquad (5)$$

EXPERIMENTAL RESULTS AND ANALYSIS

These sections discuss the result. Figure 4 illustrates the gain charts for both training and testing results of the SVM based on the normalized poly kernel before using feature selection. Gains chart with a baseline, best line is ($Best-FSSVM) and the result of SVM before improvement is ($S-FSSVM). Table 4 and 5 above demonstrate the results of the experiment for training, testing, false positive, time-cost and others using SVM based on normalized poly kernel. The correct classification using SVM based on normalized poly kernel for training 93.33% and testing 93.55% of the accuracy while the wrong classification is 6.67% for training, 6.45% for testing of accuracy, the time-cost result is 169.42 sec, false positive rate is 0.042 and false negative rate is 0.10. Figure 5 illustrates the gain charts for both training and testing results of the SVM based on the normalized poly kernel and feature selection based on one-way ANOVA F-test. Gains chart with a baseline, best line is ($Best-FSSVM) and the result of SVM after improvement is ($S-FSSVM).

Table 6 and 7 above illustrates the training, testing, false negative, false positive, time-cost and others results using feature selection based on one-way ANOVA F-test and SVM based on the normalized poly kernel. To verify the best low time, low misclassification (false positive and false negative) and accuracy it must make result equivalent to the blue line. The result after using hybrid of feature selection and SVM is verifying the best line that is equivalent to the blue line. Classification using hybrid of feature selection and SVM for training 93.45% and testing 93.55% accuracy while the wrong classification is 6.55% for training and 6.45% for testing, time-cost is 63.09 sec, false positive rate is 0.055 and false negative rate is 0.09.

The final results after using feature selection and SVM are giving the best time-cost and reduce the percentage of false negative than others. Furthermore, increase the percentage of false positive. Gains chart with baseline and the result of SVM after enhancement ($S-FSSVM).

Table 8 and Fig. 6 to 8 above illustrate the comparison of the accuracy, false negative, false positive and time-cost results between SVM and hybrid

Table 4: Analysis the training and testing result using SVM before using feature selection

Classification	Training	Training (%)	Testing	Testing (%)	Time-cost
Correct	3006	93.33	1291	93.55	169.42 sec
Wrong	215	6.67	89	6.45	
Total	3221		1380		

Table 5: Detailed of accuracy depend on type of class using SVM

False negative	False positive	Precision	Recall	F-measure	Class
0.10	0.042	0.93	0.90	0.93	1
		0.94	0.96	0.93	0

Table 6: Analysis of training and testing result after using SVM and feature selection

Classification	Training	Training (%)	Testing	Testing (%)	Time-cost
Correct	3010	93.45	1291	93.55	63.09 sec
Wrong	211	6.55	89	6.45	
Total	3221		1380		

Table 7: Detailed of accuracy by class after using SVM and feature selection

False negative	False positive	Precision	Recall	F-measure	Class
0.09	0.055	0.91	0.91	0.91	1
		0.94	0.95	0.94	0

Table 8: Summary of accuracy, false positive, false negative and time cost

Classifier	Accuracy	False negative	False positive	Time cost
SVM	93.55	0.10	0.042	169.42 sec
SVM and FS	93.55	0.09	0.055	63.09 sec

Table 9: Comparisons between different methods

Paper authors	Method	Accuracy (%)	Time-cost/sec	False positive
Chen *et al*. (2009)	Naïve Bayes and CDM	70.77	-	0.50
Wang *et al*. (2005)	GA and SVM	87.89	-	1.03
Fagbola *et al*. (2012)	GA-SVM	93.50	119.562	-
Fagbola *et al*. (2012)	SVM	90.00	149.984	-
Uysal and Gunal (2012)	DFS-SVM	71.00	79.900	-
Before enhanced	SVM before	93.55	169.420	0.06
After enhanced	FS-SVM	93.55	63.090	0.04

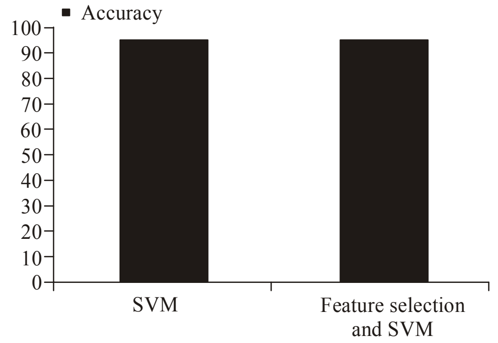

Fig. 6: Accuracy comparisons between SVM and hybrid of SVM and feature selection

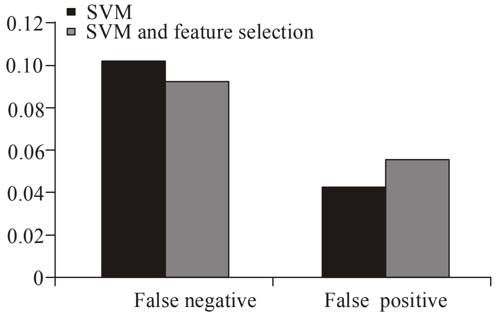

Fig. 7: False positive and false negative comparisons between SVM and hybrid SVM

of (SVM and feature selection). The result after using hybrid of feature selection and SVM is enhancement the percentage of false negative, improvement the time-cost and less false positive than using SVM only.

Compression with others methods: This section demonstrates the results of comparison between several different methods using enhanced spam detection. When comparing the result of different approaches with our result we find our result is better than the different approaches. Table 9 shows the summarized results obtained after comparing the result with others method using different methods. Figure 9 presents the accuracy comparisons between different methods.

From Table 9, Fig. 9 and different methods were used to enhance spam detection. Chen *et al*. (2009) used Naïve Bayes and Class Discriminating Measure (CDM) in their study the result of testing was 70.77% while using naïve Bayes and Multi-class Odd Ratio (MOR) the result of testing was 72.59.

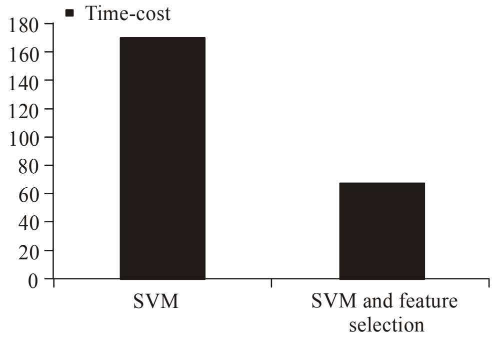

Fig. 8: Time-cost comparisons between SVM and hybrid between SVM and FS

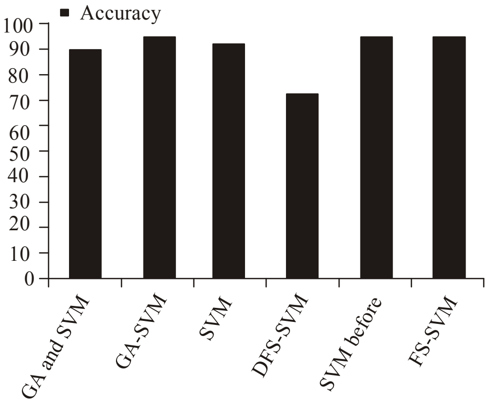

Fig. 9: Column accuracy comparisons between different methods

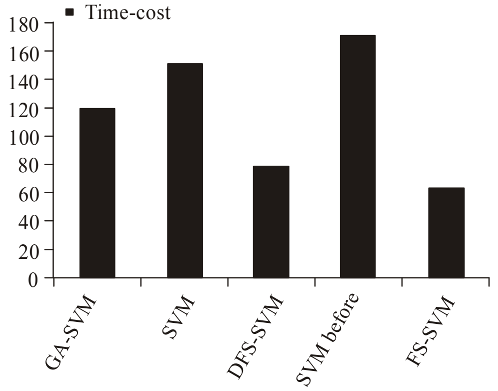

Fig. 10: Column time cost comparisons different methods

Uysal and Gunal (2012) used Distinguishing Feature Selector (DFS) and Support Vector Machine (SVM). They found that the testing result was 71%. In his study Wang *et al*. (2005) used Support Vector Machines (SVM) based on Structural Risk Minimization (SRM) and integrating feature selection using Genetic Algorithm (GA). The experimental result was 87.89 for testing. Fagbola *et al*. (2012) used GA as feature selection and SVM as classifier. They found that the testing result was 93.5% using GA-SVM and the testing result was 90.0% of accuracy using SVM. In our study the result for testing is 93.55% of accuracy using FS-SVM while the result for testing is 93.55% of accuracy using SVM. The comparison indicates that our news che me is out perform than others in term of accuracy.

From Fig. 10 comparisons several methods with our scheme, Uysal and Gunal (2012) used Distinguishing Feature Selector (DFS) and Support Vector Machine (SVM). They found that the time cost is 79.9 sec. in his study Fagbola *et al*. (2012) used GA as feature selection and SVM as classifier. They found that the time cost is 119.562 sec when used GA-SVM and the time cost are 149.984 sec when used SVM. In our study the time cost is 63.09 when using FS-SVM while the time cost is 169.42 when using SVM. The comparison indicates that our new scheme is outperform than others in term of computational time.

The contributions of this study: The final results after using feature selection and SVM improved accuracy and reduced time-cost and give a better result than others. The advantages and contributions of this study can be summarized by these points:

- The new spam detection scheme is based on the hybrid between the one-way ANOVA F-test as a feature selector and SVM based on normalized poly kernel as classifier which none of the previous studies introduced.
- The combined method achieved better results in terms of classification accuracy and computational time on the spam base standard dataset.
- Our proposed method obtained good results by reducing the time cost and increase the accuracy compared with current spam detection methods suchas Fagbola *et al*. (2012) and Uysal and Gunal (2012).
- Enhancement of accuracy score by combing the one-way ANOVAF-test measurement as a feature selector to the SVM and proved that via t-test significance test algorithm.

STATISTICAL TESTING

In this study two statistical techniques were used: t-test and correlation coefficient for significance testing for accuracy prospective:

t-test for significant: t-test is the statistical technique used to select words that are more likely to be spam or non-spam. The formula of the t-test is shown as following:

Table10: Statistical significant testing using t-test

	Paired differences							
				95% confidence interval of the difference				
	Mean	S.D.	S.E. mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1 SVM-FSSVM	1.162E-03	0.1167	1.721E-03	-2.21E-03	-4.536E-03	0.675	4600	0.005

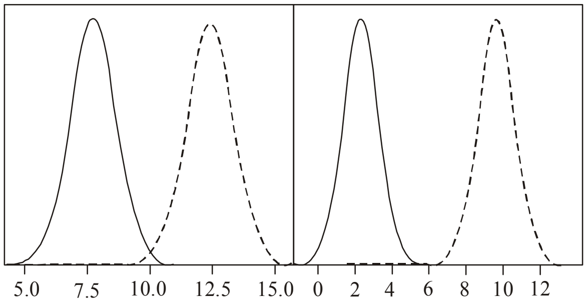

Fig. 11: t-test statistics

$$Z_0 = \frac{|\overline{X}_l - \overline{X}_S|}{\sqrt{\frac{\sigma^2{}_l}{n_l} + \frac{\sigma^2{}_s}{n_s}}}$$

From the equation $\overline{X_l}$ is the mean of non-spam mail, $\overline{X_s}$ is the mean of spam mail, σ_l is the variance of non-spam mail, σ_s is the variance of spam mail, n_l is the number of non-spam mail and n_s is the number of spam mail.

From Fig. 11, x-axis describes the frequency of each word in each email and y-axis describes the frequency of email that has this word. For example, if we have the frequencies of occurrence of the word “all” in spam and non-spam mail as 10 and 12, respectively, it is likely that the word “all” is not playing as a main feature in characterizing spam mail and non-spam mail. On the other hand, if we get the frequencies equal to 2 and 10 for the word “make” in spam mail and non-spam mail respectively, it is likely that the word “make” is significant in determining spam and non-spam mail. Thus, the word “make” should be selected as one of the features to classify spam mails. t-test statistical technique is the method that can be used to separate non-spam mail and spam mail when the difference between averages of spam and non-spam mail is high and the difference of variance is low. From Fig. 11, the left-hand side shows the low z_0 that graph has overlap area, so we cannot separate between spam and non-spam emails. The right-hand side of the figure shows the high z_0 withhigh contrast between spam and non-spam mails.

In this study, t-test is used to show the significance size of our proposed method. Table 10 shows that our proposed method is statistically significant.

Table 10 above displays the number of cases, mean the value, standard deviation and standard error and significance values for the pairs of variables, before and after optimization using feature selection (SVM, FSSVM) compared in the Paired Samples t-test method. The Paired-Samples t-test methods evaluate the means of two variables that stand for the same group at different times. The average values for the two variables (SVM, FSSVM) are showed in the Paired Samples Statistics in Table 10. Since the Paired Samples t-test evaluate the means of the two variables, it is expedient to know what the mean values are. A low significance value for the t-test (typically less than 0.05) indicates that there is a significant between the two variables. We can see Table 10, the terms VAR is (0.005), which shows that our proposed method obtained significant results in SVM and FSSVM. In addition, the significance value is high in SVM and FSSVM values and the confidence interval for the mean difference does not contain zero. We can therefore conclude that there is significant between results before and after optimization.

Correlation coefficient: Correlation Coefficient (CC) is a statistical tool used to study correlations between the set of variables. For example, a CC is used to learn a relationship between two variables and then the user can take decision on these relationships. Pearson’s, Kendall and Spearman’s correlation coefficients are well known CC types (Mohammed *et al*., 2010). CC they are descriptive statistical measures that demonstrate the strength or degree of relationship between two or more variables.

In this study, we used the Pearson’s-CC in order to assess the significance of our proposed method. Karl Pearson (Pearson, 1920) proposed Pearson’s correlation coefficient. It gauges the linear relationship between two variables and the relationship is located between -1 and 1. The CC obtains the highest value i.e., +1 if the variable is growing in relationship and on other hand CC obtains -1 if variable is decreasing in relationship.

The Pearson’s correlation coefficient ranges from -1 to +1 and can calculate using the below equation:

$$p = \frac{N\sum_i xy - \sum_i x \sum_i y}{\sqrt{(N\sum_i x^2 - (\sum_i x)^2)\,(N\sum_i y^2 - (\sum_i y)^2)}}$$

where,

X = The column result before improvement
y = The column result after improvement

Table 11: Statistical significant testing using correlation coefficient

Correlations coefficient	SVM	FSSVM
SVM Pearson correlation	1.0000	0.9600
Sig. (2-tailed)	0	0.0002
N	4601	4601
FSSVM Pearson correlation	0.9600	1.0000
Sig. (2-tailed)	0.0002	0
N	4601	4601

Correlation coefficient is significant at the 0.01 level

In this study, Pearson's correlation coefficient is used to show the significance size of our proposed method. Table 11 shows that our proposed method is statistically significant.

Table 11 reflects about comparison among original SVM and our improved SVM algorithm. According to Table 11 result SVM is improved because the correlation coefficient result is 0.0002 and this result is less than 0.01 that mean there significant.

CONCLUSION AND RECOMMENDATIONS

This study proposed a new scheme using Feature Selection (FS) based on one-way ANOVA F-test and SVM based on normalized poly kernel approach for spam detection. In general, the proposed scheme offers the following advantages: Improved classification accuracy, reducing computational time and reduce the false positive rate. The classification accuracy of a new scheme is 93.55%, the time cost is 63.09 sec and false positive is 0.04. The result of a new scheme using a hybrid of (feature selection and SVM) was compared with spam detection using SVM and others, it gives better classification accuracy, the computational time and false positive rate. There are adopted methods for ranking the 52 attributes to determine the most effective spam predictors, such as the t-statistic measure of the significance of the difference between two means of each attribute for the spam and non-spam subsets of the data. Performance was compared with others classifiers the algorithm reported in the literature using difference email dataset. The comparison indicates that the new scheme provides better classification accuracy, computational time and false positive rate than others. For future study, we plan to implement the feature selection based on differential evolution algorithm to gives quite efficient and quite accurate results. Furthermore, feature selection using to estimate the optimal features.

ACKNOWLEDGMENT

This study was financially supported in part by IDF in Universiti Teknologi Malaysia. The authors would like to thank the Research Management Centre (RMC) Universiti Teknologi Malaysia and Algraf Technical College-Sudan for their supports.

REFERENCES

Alguliyev, R. and S. Nazirova, 2012. Two approaches on implementation of CBR and CRM technologies to the spam filtering problem. J. Inform. Secur., 3(1): 11-17.

Alper, K.U. and S. Gunal, 2012. A novel probabilistic feature selection method for text classification. Knowl. Based Syst., 36: 226-235.

Bulletin, K.S., 2012. Retreieved form: http://www.securelist.com/en/analysis/204792224/Spam_report_February_2012.

Chen, J., H. Huang, S. Tian and Y. Qu, 2009. Feature selection for text classification with Naïve Bayes. Expert Syst. Appl., 36(3, Part 1): 5432-5435.

Chhabra, P., R. Wadhvani and S. Shukla, 2010. Spam filtering using support vector machine. Special Issue IJCCT, 1(2, 3, 4): 161-171.

Fagbola, T., S. Olabiyisi and A. Adigun, 2012. Hybrid GA-SVM for efficient feature selection in e-mail classification. Comput. Eng. Intelli. Syst., 3(3): 17-28.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3: 1289-1305.

Golovko, V., S. Bezobrazov, P. Kachurka and L. Vaitsekhovich, 2010. Neural network and artificial immunesystems for malware and network intrusion detection. Adv. Mach. Learn. II, 263: 485-513.

Guzella, T.S. and W.M. Caminhas, 2009. A review of machine learning approaches to Spam filtering. Expert Syst. Appl., 36(7): 10206-10222.

Ji, Z. and D. Dasgupata, 2004. Augmented negative selectionalgorithmwithvariable-coverage detectors. Proceeding of the Congress on Evolutionary Computation, CEC 2004, pp: 1081-1088.

Jin, Q. and M. Ming, 2011. A method to construct self set for IDS based on negative selection algorithm. Proceeding of the International Conference on Mechatronic Science, Electric Engineering and Computer (MEC), pp: 1051-1053.

Lai, C.C. and C.H. Wu, 2007. Particle swarm optimization-aided feature selection for spam email classification. Proceeding of the 2nd International Conference on Innovative Computing, Information and Control, pp: 165.

Lee, S.M., D.S. Kim, J.H. Kim and J.S. Park, 2010. Spam detection using feature selection and parameters optimization. Proceeding of the International Conference on Complex, Intelligent and Software Intensive Systems (CISIS), pp: 883-888.

Liang, J., S. Yang and A. Winstanley, 2008. Invariant optimal feature selection: A distance discriminant and feature ranking based solution. Pattern Recog., 41(5): 1429-1439.

Long, X., W.L. Cleveland and Y.L. Yao, 2011. Methods and Systems for Identifying and Localizing Objects based on Features of the Objects that are Mapped to a Vector: Google Patents.

Ma, W., D. Tran and D. Sharma, 2009. A novel spam email detection system based on negative selection. Proceeding of the 4th International Conference on Computer Sciences and Convergence Information Technology, ICCIT '09, pp: 987-992.

Mark Hopkins, E.R., G. Forman and J. Suermondt, (year). Spambase Dataset. Retreived form: ftp:// ftp.ics.uci.edu/pub/machine-learningdata bas es/ spambase/.

Marsono, M.N., 2007. Towards improving e-mail content classification for spam control: Architecture, abstraction and strategies. Ph.D. Thesis, University of Victoria.

Méndez, J.R., F. Fdez-Riverola, F. Diaz, E. Iglesias and J. Corchado, 2006. A comparative performance study of feature selection methods for the anti-spam filtering domain. Lect. Notes Comput. Sci., 4065: 106-120.

Mohammad, A.H. and R.A. Zitar, 2011. Application of genetic optimized artificial immune system and neural networks in spam detection. Appl. Soft Comput., 11(4): 3827-3845.

Mohammed, M., A. Shawkat and T. Kevin, 2010. Improved C4.5 algorithm for rule based classification. Proceedings of 9th Artificial Intelligence, Knowledge Engineering and Database Conference (AIKED'10), pp: 296-301.

Morariu, D., L. Vintan and V. Tresp, 2006. Evolutionary feature selection for text documents using the SVM. Informatics, 15: 215-221.

Nazirova, S. and R. Alguliyev, 2012. Two Approaches on implementation of CBR and CRM technologies to the spam filtering problem. J. Inform. Secur., 3(1): 11-17.

Noble, W.S., 2006. What is a support vector machine? Nature Biotechnol., 24(12): 1565-1567.

Parimala, R. and R. Nallaswamy, 2011. A study of spam e-mail classification using feature selection package. Global J. Comput. Sci. Technol., 11(7).

Pearson, K., 1920. Notes on the history of correlation. Biometrika, 13(1): 25-45.

Saad, O., A. Darwish and R. Faraj, 2012. A survey of machine learning techniques for Spam filtering. Int. J. Comput. Sci. Network Secur., 12(2): 66.

Salcedo-Campos, F., J. Díaz-Verdejo and P. García-Teodoro, 2012. Segmental parameterisation and statistical modelling of e-mail headers for spam detection. Inform. Sci., 195(0): 45-61.

Salehi, S. and A. Selamat, 2011. Hybrid simple artificial immune system (SAIS) and particle swarm optimization (PSO) for spam detection. Proceeding of the 5th Malaysian Conference in Software Engineering (MySEC), pp: 124-129.

Sanasam,R.S., H.A. Murthy and Timothy A. Gonsalves, 2010. Feature Selection for Text Classification Based on Gini Coefficient of Inequality. Proceeding of the 4th Workshop and Conference on Feature Selection in Data Mining, pp: 76-85.

Shang, W., H. Huang, H. Zhu, Y. Lin, Y. Qu and Z. Wang,2007.A novel feature selection algorithm for text categorization. Expert Syst. Appl., 33(1): 1-5.

Sirisanyalak, B. and O. Sornil, 2007. An artificial immunity-based spam detection system. Proceeding of the IEEE Congress on Evolutionary Computation, pp: 3392-3398.

Sun, J., C. Zheng, X. Li and Y. Zhou, 2010. Analysis of the distance between two classes for tuning SVM hyperparameters. IEEE T. Neural Networks, 21(2): 305-318.

Tu, C.J., L.Y. Chuang, J.Y. Chang and C.H. Yang, 2007. Feature selection using PSO-SVM. IAENG Int. J. Comput. Sci., 33(1): 111-116.

Uğuz, H., 2011. A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowl. Based Syst., 24(7): 1024-1032.

Unler, A. and A. Murat, 2010. A discrete particle swarm optimization method for feature selection in binary classification problems. Euro. J. Operat. Res., 206(3): 528-539.

Unler, A., A. Murat and R.B. Chinnam, 2011. < i> mr</i>< sup> 2</sup>< i> PSO</i>: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. Inform. Sci., 181(20): 4625-4641.

Uysal, A.K. and S. Gunal, 2012. A novel probabilistic feature selection method for text classification. Knowl. Based Syst., 36: 226-235.

Wang, L., 2005. Support Vector Machines: Theory and Applications. Springer Verlag, Berlin, New York.

Wang, X.L. and I. Cloete, 2005. Learning to classify email: A survey. Proceeding of 2005 International Conference on Machine Learning and Cybernetics, pp: 5716-5719.

Wang, H.B., Y. Yu and Z. Liu, 2005. SVM classifier incorporating features electionusing GA forspam detection. Lect. Notes Comput. Sci., 3824: 1147-1154.

Wang, Z.Q., X. Sun, X. Li and D.X. Zhang, 2006. An efficient SVM-based spam filtering algorithm. Proceeding of the International Conference on Machine Learning and Cybernetics, pp: 3682-3686.

Wood, P., 2012. Symantec Intelligence Report: September 2012.

Wu, X., V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, J.M. Geoffrey, N. Angus, L. Bing and S.Y. Philip, 2008. Top 10 algorithms in data mining. Knowl. Inform. Syst., 14(1): 1-37.

Xiao-Li, C., L. Pei-Yu, Z. Zhen-Fang and Q. Ye, 2009. A method of spam filtering based on weighted support vector machines. Proceeding of the IEEE International Symposium on IT in Medicine and Education, pp: 947-950.

Yang, Y. and J.O. Pedersen, 1997. A comparative study on feature selection in text categorization. Proceeding of the 14th International Conference on Machine Learning, pp: 412-420.

Yang, Z., X. Nie, W. Xu and J. Guo, 2006. An approach to spam detection by naive Bayes ensemble based on decision induction. Proceeding of the 6th International Conference on Intelligent Systems Design and Applications, ISDA'06., pp: 861-866.

Ying, K.C., S.W. Lin, Z.J. Lee and Y.T. Lin, 2010. An ensemble approach applied to classify spam e-mails. Expert Syst. Appl., 37(3): 2197-2201.

Youn, S. and D. McLeod, 2007. A comparative study for email classification. Adv. Innovat. Syst. Comput. Sci. Software Eng., pp: 387-391.

Yun, C., D. Shin, H. Jo and J. Yang, 2007. An experimental study on feature subset selection methods. Proceeding of the 7th IEEE International Conference on Computer and Information Technology, pp: 77-82.

Zhu, Z., 2008. An email classification model based on rough set and support vector machine. Proceeding of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD'08., pp: 236-240.