J. Basic. Appl. Sci. Res., 3(1s)25-30, 2013

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Review of Some Machine Learning Algorithms for Spam Filtering

Nayer Talebi Beyrami[1], Nasim Vasfi Sisi[2] and Mohammad Reza Feizi Derakhshi[3]

[1,2]Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran
[3]Department of Computer, university of Tabriz, Tabriz, Iran

Received: June 10 2013
Accepted: July 7 2013

ABSTRACT

Nowadays spam emails are increasing dramatically and make problems to users. As we know, spam not only damage users' profits, time consuming and band width, but also is has become as a risk to efficiency, reliability and security of network. In this article, we try to study some of the machine learning methods like SVM, Naïve Bayesian and Neural network applied to spam filtering and finally examine the measurement criteria (accuracy, recall, precision) of these three algorithms in terms of performance.
KEYWORDS: spam and ham emails, spam detection, spam filtering, machine learning algorithms.

1. INTRODUCTION

Since spam emails are increasing, so we must be able to deal with them and filter them. One of the methods that can be used to spam filtering is the machine learning methods to email classification and these spam filtering methods based on content are classified into two rule based method and statistic based method. Statistic based method represents the difference among messages about the possibility of certain events occurrence, such as Naïve Bayesian and SVM methods, while rule based method represents a range of knowledge on a set of heuristic rules like neural network [1].

In the present review, in section 2 the neural network and then in section 3 the SVM algorithm and some methods to improve SVM such as weighted SVM, PDWSK and GA-SVM are explained and in section 4 we will refer to Naïve Bayesian algorithm.

2. Neural Network

Neural network is a rule based spam filtering system which has many capabilities like self learning, self compatibility and fault tolerance. The function of this system is to obtain the features of emails whether they are spam or not by using feature extraction algorithm and then filter these rules using neural network algorithm and obtain a set of rules. Also, we must consider a threshold score for spam and compare the entered email with this threshold score, if the score of entered email is higher than the threshold, the email is entered into spam class, otherwise this email is placed in the valid email class. This system has three modules shown in fig. 1 [2]:

- Rule extraction module
- Rule optimization module
- Rule filtering module

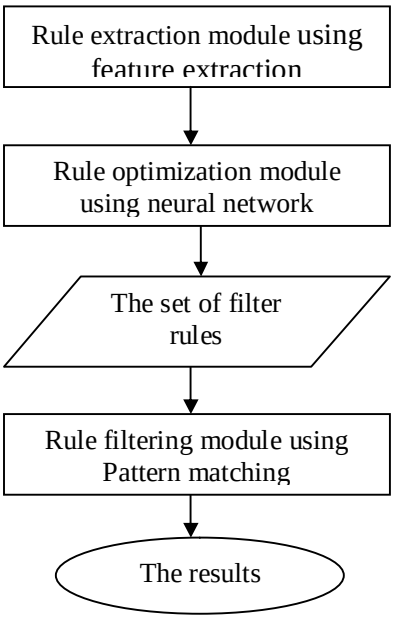

Fig 1. Flow chart of the rule-based spam filtering system [1].

Corresponding author: Nayer Talebibeyrami, Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran, Email address: Talebi.beyrami@yahoo.com

The function of the first module is to extract spam features and give a score for each item in the set of features by using information gain algorithm and then sort these items in a descending manner based on their scores, some of these features are determined based on information gain procedure. The function of rules improvement module is to use neural network technology to improve spam filtering rules [1].

Back propagation is a usual method for neural network training which is based on supervised learning method. BP network has necessarily multilayer neurons (usually, an input layer, a hidden layer and an output layer) [2].

Following construction of the neural network, we train a set of training data through neural network so that after learning the score of each rule would be improved and test data would be classified in terms of these rules. This system trains continuously the neural network until the system reaches a persistent state and the input of each neuron is a Net [1]. The function of the third module includes three following steps:

- The process of rule filtering
- Rule tree
- Format of the rule

Measurement criteria of FP, FN and Accuracy as equations 1, 2 and 3 [1]:

$$\text{Accuracy} = \frac{n_{ham \to ham} + n_{spam \to spam}}{N_{ham} + N_{spam}} \quad (1)$$

$$\text{False positive} = \frac{n_{ham \to spam}}{N_{ham}} \quad (2)$$

$$\text{False Negetive} = \frac{n_{spam \to ham}}{N_{spam}} \quad (3)$$

Luo , Liu , Yan and He (2011) [1] English corpus has been used to perform a test and according to their findings, the threshold value has a great influence on classification, so a method has been purposed to select the threshold value, then Corpus Spam Assassin and CCERT 2005-jul Corpus have been used for evaluation and the threshold value has been improved from 0.8 to 0.99. In order to evaluate the obtained results, the purposed system has been compared with Spam Assassin system and results given in Table 1. Results show that the purposed system increases the precision and decrease the FN and FP value in comparison with Spam Assassin system [1].

Table 1. COMPARISON OF SPAMASSASSIN AND Qin Luo, et al System [1].

	Accuracy	False negative	False positive
SpamAssassin	97.4%	4.7%	0.42%
Qin Luo, et al System	98.5%	3.1%	0.23%

3. SVM

SVM is a supervised learning method which separates each group into binary classes [8]. SVM is based on decision planes concept which determines the boundaries of decision making. A decision planes separates a set of objects with different class members. This algorithm finds an optimal hyper plane with maximum margin to separate two classes. Fig. 2 shows a hyper plane which separates points from each other [5].

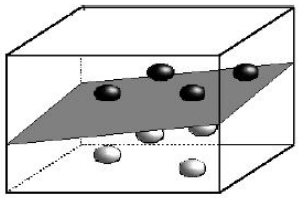

Fig 2. An SVM separating black and white points in 3 dimensions [2].

SVM is commonly used in spam filtering and it is a new developed pattern recognition method from statistical learning theory. Now, SVM has focused on two aspects: one is to study SVM model to create “online model” which is required for spam filtering and another is to research on kernel methods to complete use of syntax and semantic information of messages’ text in spam filtering [3].

WSK(Word Sequence Kernel): The superiority of WSK method in solving finite samples problem is non linear and multi dimensional pattern recognition. Most of the used kernels are distance based kernel. These kernels ignore text syntax and lead to losing mass of semantic information. To solve this problem, word sequence kernel was purposed which is able to extract information from text syntax and also dependency based word sequence kernel had been purposed. PDWSK could show semantic similarity better than WSK and increase the SVM filter precision. To confirm the efficiency of PDWSK, the efficiency of PPDWSK filtering is compared with polynomial kernel and string sequence

kernel (SSK) and words sequence kernel. The evaluation indexes are precision, accuracy, recall and F1 (the balance between precision and recall) which are defined as equations 4, 5, 6 and 7 [3].

$$Accuracy = \frac{TruePositive + TrueNegative}{AllTestSamples} \tag{4}$$

$$precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{5}$$

$$\mathrm{Re}call = \frac{TruePositive}{TruePositive + FalseNegative} \tag{6}$$

$$F_1 = \frac{2*(\mathrm{Pr}ecision*\mathrm{Re}call)}{\mathrm{Pr}ecision+\mathrm{Re}call} \tag{7}$$

Yuguo , Zhenfang and Jing (2011) [3] has been used trec07p corpus to evaluate this method. They have divided corpus into five parts where four parts are selected randomly for training set and the last part is considered as a test set and this represents that each sample not only can be applied as a training set, but also it can be a test set. The comparison results are as Table 2 [3]:

Table 2. Filtering accuracy of SVM under different kernel functions [3].

KERNELS	Precision	Recall	F1
RBF(X2)	89.60	88.75	89.17
Polynominal	84.48	84.58	84.53
SSK	90.30	91.68	90.98
WSK	92.68	91.49	92.08
PDWSK	93.64	92.21	92.92

The studying of kernel function in SVM and PDWSK shows that these functions are effective in spam filtering, and the time complexity of PDWSK is better than SSK, but still it is worse than polynomial kernel and RBF [3].

WSVM: the standard SVM was purposed in 1995, this method has many advantages to solve non linear data sets and pattern recognition with high dimensions. To solve the unbalanced classification problem, the WSVM has been purposed. Since the difference of each class's importance leads to unbalanced problems, so to improve the accuracy of SVM classification method and reduce the error of classifications in emails a set of variables has been added to the standard SVM which lead to create WSVM method. As we know, the standard SVM equation is as equation (8) [4].

$$\begin{aligned} &\min && \tfrac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. && y_i[(w\cdot x_i)+b] \geq 1-\xi_i \end{aligned} \tag{8}$$

Where $\xi_i \geq 0, i = 1,2,...$, w is coefficients vector, the constant value of C is as punishment factor and it is used for handling non separable data. Equation (9) is used for classification [4].

$$F(x) = \mathrm{sgn}\left|(w^*.x) + b^*\right| \tag{9}$$

The WSVM has added two variables of σ representing importance (weight) of each class and variable $s_i > 0$ representing importance (weight) of each email to the equation of standard SVM. After adding variables, the accurate possibility of emails classification is increased and improved as equation (10) [4].

$$\begin{aligned} &\min && \tfrac{1}{2}\|w\|^2 + C\sigma\sum_{i=1}^{l} s_i\xi_i \\ &s.t. && y_i(w^T\varphi(x_i)+b) \geq 1-\xi_i \end{aligned} \tag{10}$$

Where y_i is the label of sample i and $\varphi(x_i)$ is the non linear kernel function. Then by using Lagrange function, SVM equation can be improved. The equation (11) shows the improved function [4]:

$$F(x) = \mathrm{sgn}\left[\sum_{i=1}^{l} y_i\alpha_i^* k(x_i, x) + b^*\right] \tag{11}$$

This method reduces the inaccurate classification in valid emails, while in this method the amount of accuracy is reduced slightly. The results of tests on weighted SVM are as Table (3) [4}:

Table 3. Experimental results of WSVM [4].

$\sigma+$	$\sigma-$	R	P	A
2	1	96.50%	97.47%	97.00%
5	1	93.00%	98.41%	95.75%
10	1	89.50%	99.44%	94.50%

Hybrid GA-SVM: since in large data set of emails the SVM classification method is inefficient and does not give precise results, a feature selection technique has been purposed based on genetic algorithm and SVM (GA-SVM) which improves the accuracy and computation time and it has used spam assassin dataset to evaluate the performance of purposed system. The results given by Temitayo , Stephen and Abimbola (2012) [6] showed that SVM consumes more time and memory to classify large data[6].

The genetic algorithm includes a subset of hard optimization techniques which focuses on an application of selection, mutation and crossover for a population of hard competitive solutions. The genetic algorithm is applied to solve optimization problems in a parallel manner. The genetic algorithm method is used to select a finite subset which is suitable for SVM classification. By setting GA and accuracy parameters, the SVM classification is improved; therefore the genetic algorithm is used to optimize SVM. The main steps of GA-SVM method are as follows [6]:

1. E-mail feature extraction.
2. Using Genetic Algorithms to generate and select both the optimal feature subset and SVM parameters at the same time.
3. Classification of the resulting features using SVM.

Temitayo, Stephen and Abimbola (2012) [6] used the spam assassin data set has been to evaluate two methods of SVM classification and GA-SVM classification and MATLAB has been applied to implement the purposed method. the obtained results show that in the improved method, the amount of accuracy has been increased by 93.5%, while in standard SVM the accuracy is 90% and also the computation time has been reduced by 119/562017s, while in standard SVM the computation time is 149.9844s where Table 4 the results [6].

Table 4. Experimental results of GA-SVM [6].

Classifier	Classification Accuracy(%)	Computation Time(s)
SVM	90	149.9844
GA-SVM	93.5	119.562017

4. Naïve Bayesian

This method has been purposed in 1998 to detect spam which is a relevant event based rule and the possibility of an even which would be occurred in the future can be inferred from previous occurrences of that event, as a result it can applied to emails text classification. The user must produce a database with the collected words and marks before classifying the email by using this method and then by using this database, a training set is adjusted and following training, the sum of possibilities of words used in email determines that to which class the email belongs and if the sum of word possibilities exceeds threshold, this email is known as spam and otherwise the email is [5].

Naïve Bayesian model has two event models: multi variable Bernoulli event model and polynomial event model. The filter steps by Bayesian method are as follows [7]:

1. To collect a lot of spam and non-spam emails as a training set from emails and create a studying database.
2. To extract features. At first word division process must be done and then marks are selected as feature.
3. Based on two previous steps, we can estimate whether an email is spam when some features observed in the email. Assume that A shows spam email and $t_1, t_2, ..., t_n$ are features marks. In terms of training set we could determine the feature possibility t_i at each class. Let $P_h(t_i)$ as the possibility of t_i being in the class of valid emails, and $P_s(t_i)$ as the possibility of t_i being in the class of spam emails, consequently the possibility of $P(A|t_i)$ is calculated as equation (12):

$$P(A|t_i) = \frac{P_s(t_i)}{P_s(t_i) + p_h(t_i)} \qquad (12)$$

4. If the features mark in emails is $t_1, t_2, ..., t_n$, and the corresponding value is $p_1, p_2, ...p_n$, the possibility of combination is computed as equation (13):

$$P(A|t_1, t_2, ..., t_n) = \frac{\prod_{i=1}^{n} P_i}{\prod_{i=1}^{n} p_i + \prod_{i=1}^{n} (1 - P_i)} \qquad (13)$$

5. According to predefined rules we could evaluate whether the message is spam by setting the value of threshold $(0 \le T_d \le 1)$. If $P(A|t_1,t_2,...,t_n)$ is greater than T_d, it is known as spam.

Therefore the main factors effective on filter function are as follows [7]:

- Training samples
- E-mail pre-management
- Feature extraction and feature database
- Threshold

In order to optimize the Naïve Bayesian method, assume that some words exist in the database of features which applied in the normal features of database and they are similar to the invalid emails database, the influence of these words is low in classification and even in some cases they lead to interference in possibility computation and reduce the performance of filter. The features of normal emails and spam emails are called positive features and negative identity, respectively and the difference exists in the possibility of same words within two databases is called difference possibility and equation (14) has been used to improve the features [7]:

$$P(t_i \mid C_j) = \frac{count\,(t_i, C_j)}{\sum_{k=1}^{m} count\,(t_k, C_j)} \qquad (14)$$

In equation (14), $count(t_i, C_j)$ shows the number of feature occurrence t_i in class C_j and $\sum_{k=1}^{m} count(t_k, C_j)$ represents the number of occurrence of all features in class C_j. In comparison of the presented model in this article with Bayesian classification model, this model has optimized the features. This system has been optimized by Jiansheng and Xingwen (2010) [7] and implemented with a C++ program via visual studio 2008 and they have used a set of normal and spam emails provided by china education and Research Network Emergency Response Team (CCERT) in July 2005 to evaluate. They have studied the influence of presented algorithm on precision, recall and false positive rate. The obtained result of this method shows that the presented method could improve the precision of classification and reduce the false positive rate, while the recall rate of spam has had no significant change [7]. Tables 5, 6 and 7 present the obtained results. Table 5 and 6 represent the results of traditional Bayesian method and the purposed Bayesian method, respectively and Table 7 compares the used time to filter by both methods.

Table 5. The Accuracy of Test Results: Traditional Method [7].

Training Samples	Spam Precision(%)	Spam Recall(%)	Fallout(%)
400	92.68	82.08	6.09
600	94.27	85.09	5.23
1000	95.74	89.62	4.01

Table 6. The Accuracy of Test Results: Improved Method [7].

Training Samples	Spam Precision(%)	Spam Recall(%)	Fallout(%)
400	95.80	81.87	3.61
600	96.54	84.82	2.96
1000	97.78	89.45	1.97

Table 7 Efficiency Test Results [7].

Training Samples	The Traditional method	The Improved Method
400	23s390ms	21s22ms
600	26s306ms	24s58ms
1000	32s775ms	29s11ms

5.Conclusion

In this article we have reviewed NB, ANN and SVM methods and studied some of the optimized methods of these algorithms. By selecting an appropriate threshold, the neural networks increase the precision value and optimize FN and FP rates. Results show that kernel SVM performs well in solving the problem of finite samples and non linear and high dimensional pattern recognition and weighted SVM could control the error rate of classification class and by increasing the weight of valid emails it increases the precision, while the recall rate is reduced slightly. GA-SVM method is suitable for classification of large datasets and improves the computation time and consuming memory. GA-SVM has higher recognition rate over SVM. The Naïve Bayesian method pointed in this article reduces the processing time over traditional Naïve Bayesian and improves the amount of precision, recall and fallout. As a future work, we can test these three algorithms on a same dataset to compare performance and selection of the best algorithm.

REFRENCES

1. Luo, Q. , Liu, B. , Yan, J. , He, Zh, Design and Implement a Rule-Based Spam Filtering System Using Neural Network, International Conference on Computational and Information Sciences, 2011.
2. Kumar, P.M. , Kumaresan,P , YokeshBabu, S, Accuracy Analysis of Neural Networks in removal of unsolicited e-mails. International Journal of Computer Applications, 2011. 16(3).
3. Yuguo, L. , Zhenfang, Zh. , Jing, Zh, A word sequence kernels used in spam-filtering". Scientific Research and Essays, 2011. P. 1275-1280.
4. Xiao-li, Ch. , Pei-yu, L. , Zhen-fang, Zh. , Ye, Q, A Method of Spam Filtering Based on Weighted Support Vector Machines, 2009.
5. Awad, W.A. , ELseuofi, S.M, Machine Learning methods for E-mail Classification. International Journal of Computer Applications, 2011. 16(1).
6. Temitayo, F. , Stephen,Q. , Abimbola, A, Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification". Computer Engineering and Intelligent Systems, 2012. 3(3).
7. Jiansheng,W. , Xingwen, Zh, Improvement of Chinese Spam filtering method based on Bayesian Classification, 2010.
8. Upasana., Chakravarty, S, A Survey of Text Classification Techniques for E-mail Filtering. Second International Conference on Machine Learning and Computing, 2010.