VOL.4 NO.1 March 2015 - ISSN 2233 – 1859

UOIuBIH
ORSinBIH
Operations Research Society in
Bosnia and Herzegovina

Southeast Europe Journal of Soft Computing

Available online: www.scjournal.ius.edu.ba

# COMPARISON OF MACHINE LEARNING TECHNIQUES IN SPAM E-MAIL CLASSIFICATION

Samed Jukić[1], Jasmin Azemović[2], Dino Kečo[1], Jasmin Kevric[1]

[1]International Burch University, Faculty of Engineering and IT, Francuske revolucije bb, Ilidža 71210 Sarajevo, Bosnia and Herzegovina
[2]Universitz Džemal Bijedić, Faculty of Information Technologies, Sjeverni logor 12, Mostar, Bosnia and Herzegovina

Article Info

*Article history:*
Received 17 Sep.2013
Received in revised form

*Keywords:*
Random Forest (RF), C4.5 Decision Tree, Artificial Neural Network (ANN), Spam Detection, Ensemble methods

## Abstract

E-mail still proves to be very popular and an efficient communication tool. Due to its misuse, however, managing e-mails is an important problem for organizations and individuals. Spam, known as unwanted message, is an example of misuse. Specifically, spam is defined as the arrival of unwelcomed bulk email not being requested for by recipients. This paper compares different Machine Learning Techniques in classification of spam e-mails. Random Forest (RF), C4.5 decision tree and Artificial Neural Network (ANN) were tested to determine which method provides the best results in spam e-mail classification. Our results show that RF is the best technique applied on dataset from HP Labs, indicating that ensemble methods may have an edge in spam detection

## 1. INTRODUCTION

E-mail still represents a common and effective communication tool which is unfortunately susceptible to misuse. The most popular example of misuse is spam, also known as unwanted message. More precisely, spam is defined as the receiving of unwanted bulk commercial messages not demanded by receivers. Spam should not be mistaken with non-commercial solicitations such as political or religious tones even if unwelcomed. Recent studies show that the most popular spamming practice on the internet was still email by a huge margin (Youn & McLeod, 2007; Gaikwad & Halkarnikar, 2014).

Spammers collect e-mail addresses from websites, chatrooms, customer lists and viruses. In last few years, spam emails have grown into a serious threat for security, and act as a really good phishing agent for sensitive data. Furthermore, malicious software is carried to numerous users by spam. Daily, one typical user can receive 10-50 spam emails; around 13 billion of unwanted commercial e-mail (which makes around 50% of all e-mail sent) is sent each day (Grant, 2003; Gaikwad & Halkarnikar, 2014).

Every e-mail user in America received an average of 2200 pieces of spam e-mails in 2002. In 2007 it reached 3600 pieces of spam e-mails due to increase rate of 2% per month. CNNIC conducted a survey revealing that a Chinese received 13.7 spam e-mails weekly. Due to spam e-mails, American enterprises lose up to 9 billion yearly (CNNIC, 2004). Studies reveal that spam e-mails take about 60% of the incoming mails in a corporate network. With inappropriate or no countermeasures, the situation will worsen and, in the end, spam e-mails may destruct the usage of e-mail systems. Many countries are slowly starting to use anti-spam legal measures (Gaikwad & Halkarnikar, 2014).

The main argument supporting spam increase is the fact that spammers do not have any costs for it: “Because email technology allows spammers to shift the costs almost entirely to third parties, there is no incentive for the spammers to reduce the volume” (Hann, Hui, Lai, Lee, & Png, 2006). The issue for spam is the annoying content they carry. However, significant amount of spam contains some offensive materials (Maria & Ng, 2009).

In China, some specialists suggest executing effective anti-spam email measure as early as possible. However, because of the Internet's open architecture, only limited effect was seen in these legal measures by now. Due to that, we should be opting for additional effective methodologies. Currently, majority of systems stop spam messages by means of banning frequent spammers (Gaikwad & Halkarnikar, 2014; Chuan, Xian-liang, Xu, & Meng-shu, 2005).

Automated approaches discriminating between junk and legitimate emails are becoming necessity because of this growing problem (Sahami, Dumaisy, Heckerman, & Horvitz, 1998). Huge number of documents, relatively great number of features and unstructured information are challenges for automated detection of spam email. All of these features may badly impact the performance regarding speed and quality, as the usage increases. Many recent algorithms use just significant features for classification. A huge and different number of features in the dataset and a big number of documents cause a problem to the text and email classification. Since that huge number of features makes most documents indistinguishable, the applicability in datasets using existing classification techniques is limited. Different datasets use classification algorithms such as Support Vector Machine (SVM), Artificial Neural Network (NN), and Naïve Bayesian (NB) classifiers which currently show good classification results (Gaikwad & Halkarnikar, 2014; Youn & McLeod, 2007).

This paper describes the detection of spam messages using various machine learning methods. Random Forest, C4.5 and ANN methods were compared based on different performance evaluation criteria. The organization of the paper is as follows. Section 2 presents background work on detection of spam e-mail, whereas Section 3 describes the Spam dataset and ML techniques applied. Section 4 presents the experimental results. Finally, Section 5 concludes the paper.

## 2. LITERATURE OVERVIEW

A number of early studies have taken advantage of probabilistic Naïve Bayes theory in spam detection. Deshpande et al suggested an anti-spam filtering method based on Naïve Bayes. In addition, the filters were trained on huge amount of non-spam and spam e-mails and tested on unseen incoming e-mail messages. Authors conclude that safety measures are required before a Bayesian anti-spam filter is practically usable but can act as a first pass filter (Deshpande, Erbacher, & Harris, 2007). Obeid suggested a data mining paradigm grounded on Bayesian analysis for filtering spam. The algorithm learns patterns related to legitimate and spam messages and then classifies new e-mail as either legitimate or spam (binary classification). The author demonstrated the capability of filter to detect spam with high accuracy (Obied, 2007). Sahami and thee Microsoft researchers tested techniques for the automatic filter creation for removing unwelcomed mail by employing probabilistic learning methods. They show that superior results are obtained once domain-specific features and the text of e-mail messages is considered together (Sahami, Dumaisy, Heckerman, & Horvitz, 1998). Another approach to automatic e-mail classification using Bayesian Theorem by inspecting its textual contents is presented in (Vira, Raja, & Gada, 2012). An enhanced Bayesian anti-spam mail filter is presented in (Chuan, Xian-liang, Xu, & Meng-shu, 2005). The improvement in total performance is acquired as features are extracted based on word entropy, and vector weights are characterized by word frequency.

A decision tree based ensemble learning paradigm for spam email detection is suggested in (SHI, WANG, MA, WENG, & QIAO, 2012). Public spam e-mail dataset was used to evaluate performance of a few machine learning methods. The suggested ensemble learning technique showed to be mostly superior to benchmark methods. Ozarkar and Patwardhan applied Random Forest and PART Decision Trees to discriminate between legitimate and spam messages in public spam database. Different attribute extraction techniques were implemented. Although this pre-processing step decreased training times, it did not bring substantial improvement in accuracy. However, other benchmark methods were outperformed by Random Forest ensemble (Ozarkar & Patwardhan, 2013). Abu-Nimeh et al compared six data mining methods for phishing detection. Authors produced a Dataset containing 1718 non-phishing and 1171 phishing emails, where each e-mail was characterized by 43 attributes. 10-fold cross-validation was used to evaluate the classifiers performance. Random Forest was again superior to all other algorithms with overall accuracy of 92.28%. The worst performers were Support Vector Machines and Neural Networks. However, one of the disadvantages of Random Forests was high rate of false positives (Abu-Nimeh, Nappa, Wang, & Nair, 2007).

In this paper, we will confirm the superiority of Random Forest ensemble learning over single methods as in (SHI, WANG, MA, WENG, & QIAO, 2012) and (Ozarkar & Patwardhan, 2013). In addition, our work is among a few which included and evaluated ANN for spam detection. Unlike (Abu-Nimeh, Nappa, Wang, & Nair, 2007), our study identified no disadvantages of Random Forests when compared to other benchmark methods.

## 3. DATASET AND MACHINE LEARNING TECHNIQUES

### 3.1. *Dataset*

Email database is acquired from UCI's machine learning data repository (UCI, 2015). HP Labs created and donated the dataset in July 1999. Dataset collection of spam messages is from individuals and postmaster who had filed spam. On the other hand, collection of legitimate messages came from filed work and personal e-mails. In the Spam database there are completely 4601 messages out of which 1813 (39.4%) are characterized as spam. Every e-mail message is characterized as a feature vector comprising of 57 real numbers. Majority of them (47) represent frequencies of certain words. Frequencies of certain characters in the email are stored in the following 6 features. Statistics regarding capital letters constitute the remaining 3 features. These last three features hold the longest, average and sum of lengths of continuous capital letters respectively (Zhao, 2004). The names of all 57 features can be found at (UCI, 2015).

## 3.2. MACHINE LEARNING TECHIQUES

### 3.2.1. *Random Forest (RF)*

Random Forest (RF), proposed by Breiman (Breiman, 2001), is novel, fast, highly accurate, noise resistant classification method. Bagging and random feature selection are combined together in RF. Every tree in the forest is influenced by the values of random vectors sampled separately and has identical distribution as any other tree in the forest (Breiman, 2001). RF consists of outsized number of decision trees where decision tree select their separating features from bootstrap training set $S_i$ where $i$ represent $i^{th}$ internal node. Trees in RF are grown by means of Classification and Regression Tree (CART) method with no pruning. As number of trees in the forest turns into outsized number, generalization error will also increase until it converges to some boundary level (Breiman, 2001). More details about RF can be found in (Breiman, 2001).

### 3.2.2. *C4.5*

The C4.5 calculation uses the same fundamental inductive tree creation approach as ID3, yet extends its abilities to characterization of ceaseless information by gathering together discrete estimations of a trait into subsets or reaches. Another point of interest of C4.5 is that it can foresee values for information with missing properties in light of learning of the important spaces (Dunham, 2003). C4.5 additionally gives an approach to prune or diminish the extent of the tree with no noteworthy lessening in precision. Pruning happens in two structures (Dunham, 2003): subtree substitution and subtree raising. If there should arise an occurrence of the previous, a subtree is supplanted with a leaf node, and in the second system, a subtree is supplanted with its most every now and again utilized subtree (Browne & Berry, 2006).

In both cases, substitution is worthy just when the first tree experiences negligible contortion as an aftereffect of pruning. In circumstances where tree pruning does not adequately diminish the unpredictability of the DT structure, C4.5 produces choice principles in view of the decisions connected with a way, which is characterized as a situated of branches uniting two nodes (Browne & Berry, 2006).

### 3.2.1. *ANN*

An ANN can be characterized as an exceedingly associated cluster of rudimentary processors called neurons. A generally utilized model called the multi-layered perceptron (MLP) is indicated in Figure 1. The MLP comprises of one input layer, one or more hidden layers and one output layer. Every layer utilizes a few neurons and every neuron in a layer is associated with the neurons in the contiguous layer with diverse weights. The attributes (or features) stream into the input layer, go through the hidden layers, and produce an output at the output layer. Except for the input layer, every neuron gets signals from the neurons of the past layer straightly weighted by the interconnect values between neurons. The neuron then creates its output by passing the summed signal through a sigmoid or other types of activation function (Park, El-Sharkawi, Marks II, Atlas, & Damborg, 1991; Sobajic & Pao, 1989).

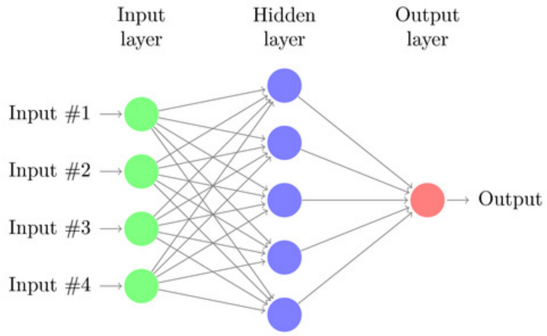

Figure 1: Structure of a Three-Layered Perceptron Type ANN.

## 4. RESULTS AND DISCUSSION

The most commonly used approach for algorithm comparison is the classification performance which is usually not focused on a class (Sokolova, Japkowicz, & Szpakowicz, 2006). For example, accuracy provides no separation among the true labels of different classes (it only evaluates the general performance of the algorithm):

$$accuracy = \frac{tp + tn}{tp + fp + fn + tn} \tag{1}$$

On the other hand, Sensitivity and Specificity represent two measures that evaluate the performance of the classifier on various classes:

$$sensitivity = \frac{tp}{tp + fn} \tag{2}$$

$$specificity = \frac{tn}{fp + tn}, \tag{3}$$

In spam e-mail classification, the Sensitivity shows how good the algorithm is in detecting spam messages, whereas the Specificity is a measure of recognition of legitimate e-mail. In other words, they both evaluate the probability of each label being correct.

There are three more measures that differentiate properly classified samples within various classes: precision, recall, and F-measure. Relation between correctly classified samples and those that are misclassified as positives is called precision.

$$precision = \frac{tp}{tp + fp} \tag{4}$$

A relation between properly classified instances and misclassified instances is called recall.

$$recall=\frac{tp}{tp+fn}=sensitivity \tag{5}$$

$$F-measure=\frac{(\beta^2+1)\cdot precision\cdot recall}{\beta^2\cdot precision+recall} \tag{6}$$

F-measure (or F-score) favors classifiers with greater values of sensitivity and challenges classifiers with greater values of specificity (Sokolova, Japkowicz, & Szpakowicz, 2006).

On the other hand, ROC can provide an extensive estimation of a classifier's effectiveness:

$$ROC=\frac{P(x|positive)}{P(x|negative)} \tag{7}$$

where $P(x|C)$ implies the likelihood that a sample belongs to the class C. In other words, ROC represents a function of the classifier's sensitivity and specificity values.

Table 1 presents performance assessment of three machine learning techniques tested on Spam database: C4.5 decision tree, ANN, and Random Forest (RF). For every algorithm, ROC area and F-Measure can be observed for each class and averaged. More importantly, Table 1 shows the detection accuracy values for spam (Sensitivity) and non-spam (Specificity) messages together with the average detection accuracy of algorithms.

Table 1: Performance evaluation of machine learning methods on Spam database.

| Algorithm | Criteria | Spam | Non-spam | Average |
|---|---|---|---|---|
| C4.5 | ROC Area | 0.939 | 0.939 | 0.939 |
| | F- Measure | 0.911 | 0.942 | 0.930 |
| | Accuracy | 90.80 | 94.40 | 92.98 |
| ANN | ROC Area | 0.959 | 0.959 | 0.959 |
| | F- Measure | 0.884 | 0.927 | 0.910 |
| | Accuracy | 87.10 | 93.70 | 91.05 |
| RF | ROC Area | 0.987 | 0.987 | 0.987 |
| | F- Measure | 0.943 | 0.964 | 0.956 |
| | Accuracy | 93.10 | 97.20 | 95.60 |

All these measures have been obtained by employing 10-fold cross-validation (CV) approach. Dataset is arbitrarily divided into 10 mutually exclusive folds (subsets) of practically the identical size. Nine (9) folds are used for training and remaining one (1) fold is used for testing so the process repeats 10 times. The average of accuracies of each iteration is then reported in Table 1.

All three classifiers have been implemented and tested in software package WEKA (Holmes, Donkin, & Witten, 1994) using default parameters. C4.5 has been used with pruning option disabled. The same holds true for the Random Forest, where the number of generated trees was 100 which will give class label by majority vote. In ANN, the number of input nodes is equal to the number of features used, namely 57. The other parameters of ANN are provided in Table 2.

Table 2: ANN parameters.

| Parameter | Value |
|---|---|
| Hidden Layers | 30 |
| Learning rate | 0.3 |
| Momentum | 0.2 |
| Epoch | 500 |
| Normalize Attributes | YES |
| Validation Set | NO |

The least effective method was ANN with average accuracy of only 91% (87.1% for spam and 93.7% for non-spam). The best result was achieved with Random Forest classifier reaching total accuracy of 95.6% (spam 93.1% and non-spam 97.2%). C4.5 decision tree fits in the middle according to accuracy with 92.98% (90.8% for spam and 94.4% for non-spam e-mails). According to other two measures (ROC and F-measure), Random Forest outperforms the other two where ANN performs the worst. It can be observed that all three algorithms have better Specificity than Sensitivity, i.e. detection accuracy of non-spam outperforms the accuracy of spam messages.

Figure 2 is a graphical representation of accuracy values for all classes and all algorithms from table 1. The observations and conclusion drawn in the previous paragraph are now even more evident.

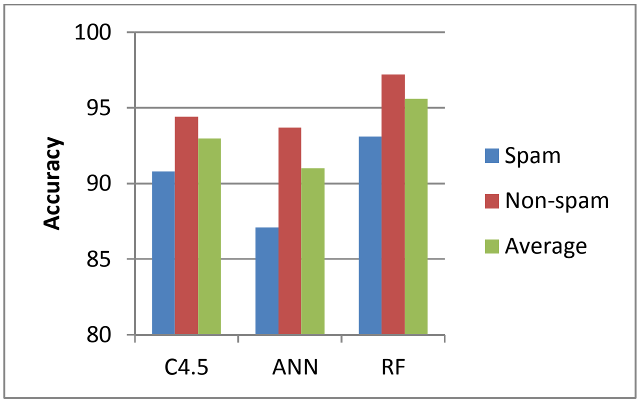

Figure 2: Graphical representation of accuracy values for three machine learning methods.

According to the presented results, emphasis on the following should be stated among the machine learning techniques. Random Forests may be successfully applied in e-mail spam classification due to their stable and high performance presented in Tables 1 and Figure 2. Our results also show that ensemble methods outperform single methods and may have an edge in classification of spam e-mail.

## 5. CONCLUSIONS

E-mail spam detection has gotten a colossal consideration by greater part of the researchers as it serves to recognize the undesirable data and potential dangerous activity. Hence, the greater part of the analysts focuses on discovering the best classifier for recognizing spam messages. This paper portrayed diverse ML systems for spam messages characterization, among which RF proved to be the best one. The upside of RF is that it runs proficiently on huge datasets with high number of samples and attributes, which makes it exceptionally appealing for content classification. In the period of testing the framework different performance measures (ROC area, F-measure, and Accuracy) were taken into consideration. The proposed framework accomplishes average accuracy of 95.56% in spam detection using RF. Future work will incorporate a comparison of ensemble methods in e-mail spam detection.

## REFERENCES

Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). *A Comparison of Machine Learning Techniques for Phishing Detection.* Dallas,TX: SMU HACNet Lab Southern Methodist University.

Breiman, L. (2001). Random Forests. (R. E. Schapire, Ed.) *Machine Learning, 45*, 5–32.

Browne, M., & Berry, M. M. (2006). *Lecture Notes in Data Mining.* Singapore: World Scientific Publishing Co. Pte. Ltd.

Chawla, N., Japkowicz, N., & Kolcz, A. (2004). Special Issue on Learning from Imbalanced Data Set. *ACM SIGKDD Explorations, 6*(1).

Chuan, Z., Xian-liang, L., Xu, Z., & Meng-shu, H. (2005). An Improved Bayesian with Application to Anti-Spam Email. *Journal of Electronic Science and Technology of China, 3*(1).

CNNIC. (2004). *The 13th China Internet Development Status Report[R].*

Deshpande, V. P., Erbacher, R. F., & Harris, C. (2007). An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques. *Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy.* West Point, NY.

Dunham, M. (2003). *Data Mining: Introductory and Advanced Topics.* NJ: Prentece Hall.

Ferry, C., Lachiche, N., Macskassy, S., & Rakotomamonjy, A. (2005). Proc ICML '05 workshop on ROC Analysis in Machine Learning. *ICML.*

Gaikwad, B. U., & Halkarnikar, P. P. (2014). Random Forest Technique for E-mail Classification. *International Journal of Scientific & Engineering Research, 5*(3), 145-153.

Grant, G. (2003). *Spam bill heads to the president.* Retrieved from http://www.nwfusion.com/news/2003/1209spambill.html

Hann, I., Hui, K., Lai, Y., Lee, S., & Png, I. (2006). Who gets spammed? *Communications of the ACM, 49(10), ., 49*(10), 83–87.

Holmes, G., Donkin, A., & Witten, I. H. (1994). WEKA: a machine learning workbench. *Proceedings of the 1994 Second Australian and New Zealand Conference on Intelligent Information Systems*, (pp. 357 - 361). Brisbane , Australia.

Maria, S. P., & Ng, Y.-K. (2009). SpamED: A Spam E-Mail Detection Approach Based on Phrase Similarity. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 60*(2), 393–409.

Obied, A. (2007). *Bayesian Spam Filtering.* Department of Computer Science University of Calgary.

Ozarkar, P., & Patwardhan, M. (2013). Efficient Spam Classification By Appropriate Feature Selection. *International Journal of Computer Engineering and Technology (IJCET), 4*(3).

Park, D., El-Sharkawi, M., Marks II, R., Atlas, L., & Damborg, M. (1991). Electric Load Forecasting Using An Artificial Neural Network. *IEEE Transactions on Power Engineering, vol.6, pp.442-449 (1991)., 6*, 442-449.

Sahami, M., Dumaisy, S., Heckerman, D., & Horvitz, E. (1998). *A Bayesian Approach to Filtering Junk E-Mail.* Stanford, CA: Computer Science Department Microsoft Research Stanford University Redmond.

SHI, L., WANG, Q., MA, X., WENG, M., & QIAO, H. (2012). Spam Email Classification Using Decision Tree Ensemble. *Journal of Computational Information Systems, 8*(3), 949-956.

Sobajic, D., & Pao, Y. (1989). Artificial Neural-Net Based Dynamic Security Assessment for Electric Power Systems. *IEEE Tr. on Power Systems, 4*(1), 220-228.

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond Accuracy, F-score and ROC: a Family of Discriminant Measures for Performance Evaluation. *American Association for Artificial Intelligence*.

UCI. (2015, January). *Machine Learning Repository.* Retrieved from Spambase Data Set: https://archive.ics.uci.edu/ml/datasets/Spambase

Vira, D., Raja, P., & Gada, S. (2012). An Approach to Email Classification Using Bayesian Theorem. *Global Journal of Computer Science and Technology Software & Data Engineering, 12*(13).

Youn, S., & McLeod, D. (2007). *A Comparative Study for Email Classification.* Los Angeles: University of Southern California.

Zhao, C. (2004). Towards better accuracy for Spam predictions.