ISSN 2348 – 9928
IJAICT Volume 3, Issue 9, September 2016 Doi:01.0401/ijaict.2014.07.02 Published on 05 (9) 2016

MACHINE LEARNING ALGORITHMS USED IN SPAM FILTERING – A STUDY

Mr. T. Kumaresan,
Assistant Professor (Sr grade),
Computer Science and Engineering,
Bannari Amman Institute of Technology,
Sathyamangalam, TamilNadu, India

Ms. S. SanjuShree,
PG Scholar,
Information Technology,
Bannari Amman Institute of Technology,
Sathyamangalam, TamilNadu, India

Ms. K. Suhasini,
PG Scholar,
Information Technology,
Bannari Amman Institute of Technology,
Sathyamangalam, TamilNadu, India

Dr. C. Palanisamy,
Professor & Head,
Information Technology,
Bannari Amman Institute of Technology,
Sathyamangalam, TamilNadu, India

Abstract— Spam is an unsolicited bulk mail. Due to increased communication within shorter duration and for longer distance and fastest medium email is considered. Now a day's spam became a big problem of internet and electronic communication. There are several techniques developed to solve the problem and to fight with them. There are many spam filters used to filter the spam from the received mail. In this paper the overview of machine learning algorithms like Bayesian classification, k-NN, ANN and SVM and their applicability are described. The performance of various algorithms are compared on the spam corpus are evaluated.

Keywords—E-Mail Classification, Machine Learning Algorithms and Spam.

I. INTRODUCTION

In recent years, emails become a big trouble over the internet. Spam is an unsolicited commercial or bulk mail, is a bane of email communication. It is a waste of time, more communication bandwidth, storage space and lost productivity are the ways they pay the fee to distribute their materials. The spam problem is uncontrollable if the spam continues to grow. The two general techniques used in email filtering are knowledge engineering and machine learning. In knowledge engineering technique, a set of rules has to be specified according to which emails are identify as spam or ham but by applying this technique there is no likely results must be shown. This technique is not convenient for most of the users and it is a waste of time because it should be updated and maintained continually. In machine learning technique, it doesn't require any rules. By comparing with knowledge engineering technique, machine learning technique is more efficient because it uses a set of pre classified email messages. Machine learning technique has been studied and there are lots of algorithms used to filter the email. The algorithms include Naïve Bayes, support vector machine, neural networks K-nearest neighbor, rough sets and the artificial immune system.

In today's business emails 70% are spam and there occur a serious problem associated with the growing rate of spam which is given above[1][2].

II. ALGORITHMS

This section gives a brief overview of the underlying theory and algorithms we consider. We shall discuss the Naïve Bayesian Classifier, Neural Network Classifier, [3]the K-NN Classifier, and the Support Vector Machine.

2.1 Naïve Bayes Classifier

Bayesian classifier is working on the dependent events and the probability of an event occurring in the future that can be detected from the previous occurring of the same event. This technique can be used to classify spam e-mails; words probabilities play the main rule here. If some words occur frequently in spam but not in ham, then this incoming e-mail is probably spammed. Naïve bayes classifier technique has become a very popular method in mail filtering software. Bayesian filter should be trained to work effectively. Every word has certain probability of occurring in spam or ham e-mail in its database. If the total words probabilities exceed a certain limit, the filter will mark the e-mail to either category. Here, only two categories are necessary: spam or ham [4][5].The naïve Bayes classifier is a statistical algorithm which provides more precise results.

2.2 Artificial Neural Networks Classifier

An artificial neural network (ANN), usually called neural network, is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks[6][7]. A neural network consists of an interconnected group of artificial neurons, and it processes

information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data. By definition, a "neural network" is a collection of interconnected nodes or neurons. See Fig. 1. The best known example of one is the human brain, the most complex and sophisticated neural network.

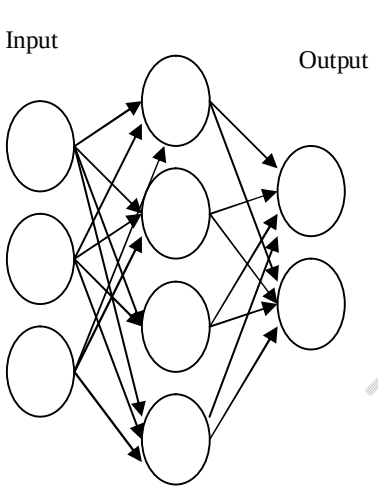

Fig 1: An artificial neural network is an interconnected group of nodes

Spam presents a unique challenge for traditional filtering technologies: both in terms of the sheer number of messages (millions of messages daily) and in the breadth of content (from pornographic to products and services, to finance). Today's economic fabric depends on email communication which is equally broad and plentiful and whose subject matter contextually overlaps with that of many spam messages and you've got a serious challenge [8][9].

2.3 K-Nearest Neighbor Classifier

The k-nearest neighbor (K-NN) classifier is considered an example-based classifier, that means that the training documents are used for comparison rather than an explicit category representation, such as the category profiles used by other classifiers. As such, there is no real training phase. When a new document needs to be categorized, the k most similar documents (neighbors) are found and if a large enough proportion of them have been assigned to a certain category, the new document is also assigned to this category, otherwise not. Additionally, finding the nearest neighbors can be quickened using traditional indexing methods [10]. The KNN algorithm gives consistent results. A disadvantage of the basic majority voting classification occurs when the class distribution is skewed. To overcome skew is by abstraction in data representation.

2.4 Support Vector Machine Classifier

Support vector machines (SVM) are relatively new technique that have rapidly gained popularity because of the excellent results they have achieved in a wide variety of machine learning problems, and because they have solid theoretical underpinnings in statistical learning theory. Support vector machine (SVM) algorithms divide the n-dimensional space representation of the data into two regions using a hyper plane. This hyper plane always maximizes the margin between the two regions or classes. The margin is defined by the longest distance between the examples of the two classes and is computed based on the distance between the closest instances of both classes to the margin, which are called supporting vectors. Instead of using linear hyper planes, many implementations of these algorithms use so-called kernel functions. These kernel functions lead to non-linear classification surfaces, such as polynomial, radial or sigmoid surfaces.

III. MACHINE LEARNING METHODS PERFORMANCE

3.1 Experimental Implementation

In order to test the performance of above mentioned four methods, some corpora of spam and legitimate emails had to be compiled; there are several collections of email publicly available to be used by researchers. Spam Assassin (http://spamassassin.apache.org) will be used in this experiment, which contains 6000 emails with the spam rate 37.04%. Thus we have divided the corpora into training and testing sets keeping, in each such set, the same proportions of ham (legitimate) and spam messages as in the original example set. Each training set produced contained 62.96% of the original set; while each test set contain 37.04% as Table I.

Table I - Corpora of Spam and Ham Messages

Message Collection	Training Set	Testing Set
Ham Messages	2378	1400
Spam Messages	1398	824
Total Messages	3776	2224

In addition to the body message of an email, an email has another part called the header. The job of the header is to store information about the message and it contains many fields like the field (From) and (Subject), we decided to divide the email into 3 different parts. The first part is the (Subject) that can be considered as the most important part in the email, it is noticed that most of the new incoming emails have descriptive

Subjects that can be used to clearly identify whether that email is Spam or Ham.

The second part is (From) which is the person that taking the responsibility of the message, this field we store it in a database and use it after the decision of the classifier has been taken, that is the way to compare the field (From) stored in the database to the field (From) in the new incoming email, if they are the same so the decision of the new incoming email is Spam. The (Body) is the third part which is the main part of the message. Furthermore we applied two procedures in the preprocessing stage [11][12]. Stopping is employed to remove common word. Case-change is employed to change the (Body) into small letters. The experiment is performed with the most frequent words in spam email; we select 100 of them as features.

3.2 Algorithm Steps

1) Email Preprocessing

The content of email is received through our software, the information is extracted then as mentioned above, then the information (Feature) extracted is saved into a corresponding database. The spam message is converted into a feature vector with 21700 attributes. If the word is present in a spam message, then an attribute value is set to 1 or to 0 otherwise. This extraction of feature scheme was used for all the algorithms.

2) Description of th Feature Extracted

Feature extraction module extract the spam text and the ham text, then produce feature dictionary and feature vectors as input of the selected algorithm, the function of feature extraction is to train and test the classifier [13]. For the train part, this module account frequency of words in the email text, we take words which the time of appearance is more than three times as the feature word of this class. And denote every email in training as a feature vector.

3) Spam Classification

Through the steps above, we take standard classification email documents as training document, pretreatment of email, extract useful information, save into text documents according to fix format, split the whole document to words, extract the feature vector of spam document and translate into the form of vector of fix format. We look for the optimal classification using the selected algorithm which is constructed using the feature vector of spam documents [14].

4) Performance Evaluation

In order to test the performance of above mentioned four methods, we used the most popular evaluation methods used by the spam filtering researchers. Spam Precision, Spam Recall, Accuracy.

3.3 Performance Comparison

We summarize the performance result of the four machine learning methods in term of spam recall, precision and accuracy. Table 2 and Figure 2 summarize the results of the four classifiers by selecting the top 100 features (the most relevant word). In term of accuracy we can find that the Naïve bayes method is the most accurate while the k-nearest neighbor give us approximately the same lower percentage, while in term of spam precision we can find that the Naïve bayes method has the highest precision among the four algorithms while the k-nearest neighbor has the worst precision percentage, and finally we can find that the recall is the less percentage among the four classifiers while the Naïve bayes still has the highest performance.

Table II- Performance of Four Machine Learning Algorithms by Selecting Top 100 Features

Algorithm	Spam Recall (%)	Spam precision (%)	Accuracy (%)
NB	98.46	99.66	99.46
SVM	95.00	93.12	96.90
KNN	97.14	87.00	96.20
ANN	96.92	96.02	96.83

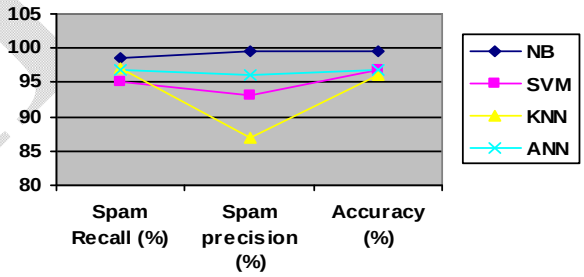

Fig 2 : Performance analysis of various algorithms

In general it (KNN) was poor, and it has the worst precision percentage. The performance of the artificial neural network is the most simple and fastest algorithm. While in term of spam recall almost all four algorithms are same and it shows some slight variation in it [15].

By comparing K-nearest neighbor and Naïve Bayes algorithm, the knn has high variance and low bias but naïve bayes has low variance and high bias when the surface is linear. When compared with performance the k nearest neighbor classifier and naive Bayes often start with an advantage on SVM when the training sets are composed of a small number of documents.

Support vector machine and artificial neural network algorithms are compared, the classification accuracy for SVM is better than the ANN algorithm. The pink color indicates the SVM algorithm and blue color indicates the ANN algorithm. SVM algorithm has low spam precision percentage when compared to ANN algorithm but accuracy of SVM algorithm is good [16].

The yellow color indicates the KNN algorithm and pink color indicates the SVM algorithm. The spam precision is too low for KNN algorithm when compared to SVM algorithm but accuracy is nearly same[17][18]. SVM method is simple and less expensive to build but KNN and Naïve Bayes algorithms are expensive to build. KNN is prone to over fitting because of its non-linear nature. SVM algorithm performs well on datasets which has many attributes and there are few cases that are available for training process but there is a size and speed limit during training and testing phase of an algorithm.

IV. CONCLUSION

Spam is becoming a serious problem to the Internet circle, threaten the efficiency of the users. In this paper we review some of the most popular machine learning methods and of their applicability to the problem of spam e-mail classification. Descriptions of the algorithms are presented, and the comparison of their performance on the Spam Assassin spam corpus is presented, the experiment showing a very promising results specially in the algorithms that is not popular in the commercial e-mail filtering packages, spam recall percentage in the four methods has the less value among the precision and the accuracy values, while in term of accuracy we can find that the Naïve bayes method has a very satisfying performance among the other methods, more research has to be done to escalate the performance of the Naïve bayes.

By comparing these four machine learning algorithms, the naïve bayes algorithm gives better accuracy than others. Among the four machine learning methods KNN algorithm has the worst precision percentage. Although methods used by us have many advantages, it certainly does come with some disadvantages. The disadvantage of text filtering is that they are time consuming.

V. FUTURE WORK

In future work, we need to improve the precision percentage of the KNN classifier method by using enhanced algorithms. It is an adaptable and scalable project thus we would like to detect threats found in emails that are viruses.

REFERENCES

[1] Ahmed Khorsi, "An Overview of Content-Based Spam Filtering Techniques", Informatics 31 (2007) 269-277 269

[2] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, "Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis", IET Computers & Digital Techniques, 2008

[3] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch"Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis" IEEE GLOBECOM, 2008

[4] Guzella, T. S. and Caminhas, W. M."A review of machine learning approaches to Spam filtering." Expert Syst. Appl., 2009

[5] Wu, C."Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks" Expert Syst., 2009

[6] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malic. "SVM-KNN: Discriminative nearest neighbour classification for visual category recognition", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006

[7] Carpinteiro, O. A. S., Lima, I., Assis, J. M. C., de Souza, A. C. Z., Moreira, E. M., & Pinheiro, C. A. M. "A neural model in anti-spam systems.", Lecture notes in computer science.Berlin, Springer, 2006

[8] Cormack, Gordon. Smucker, Mark. Clarke, Charles "Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011

[9] Androutsopoulos, G. Paliouras, ―Learning to filter spam E-mail: A comparison of a naïve bayesian and a memory based approach. In Proceedings of the Workshop on Machine Learning and Textual Information Access, fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2000), pages

[10] Hassanien, H. Al-Qaheri, Machine Learning in Spam Management, IEEE TRANS.,VOL.X, NO. X, February 26, 2009.

[11] Yoo, S., Yang, Y., Lin, F., and Moon, I. "Mining social networks for personalized email prioritization". In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France), June 28 - July 01, 2009

[12] Toshihiro Tabata, "SPAM mail filtering: commentary of Bayesian filter," The journal of Information Science and Technology Association, Vol.56, No.10, pp.464- 468, 2006

[13] N. Cristianini, B. Schoelkopf**,** "Support vector machines and kernel methods, the new generation of learning machines". Artificial Intelligence Magazine, 23(3):31–41, 2002

[14] J. Hidalgo**,** "Evaluating cost-sensitive unsolicited bulk email categorization". In Proceedings of SAC-02, 17th ACM Symposium on Applied Computing, pages 615–620, Madrid, ES, 2002

[15] Miller, "Neural Network-based Antispam Heuristics", Symantec Enterprise Security (2011), www.symantec.com Retrieved December 28, 2011

[16] N. Cristiatnini and J. Shawe-Taylor. An introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 2003. http://www.support-vector.net.

[17] Almeida,tiago. Almeida, Jurandy.Yamakami, Akebo " Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers" Journal of Internet Services and Applications, Springer London , February 2011

[18] Alwin, and J. Krosnick, ―The reliability of survey attitude measurement: The influence of questions and respondent attributes, Sociological Methods Research, 1991, 20(139).