A Comparative Study between Naïve Bayes and Neural Network (MLP) Classifier for Spam Email Detection

Amit Kumar Sharma
Faculty
Department of Computer Science,
Govt Engineering College,
Jhalawar, Rajasthan, India

Sudesh Kumar Prajapat
Assistant Professor
Department of Computer Science,
Indira Gandhi National Tribal
University, Amarkantak, MP, India

Mohammed Aslam
Department of Computer Science,
Central University of Rajasthan,
Ajmer, Rajasthan,
India

ABSTRACT

The continue demands of internet and email communication has creating spam emails also known unsolicited bulk mails. These emails enter bypass in our mail box and affect our system. Different filtering techniques are using to detect these emails such as Random Forest, Naive Bayesian, SVM and Neural Network. In this paper, we compare the different performance matrices using Bayesian Classification and Neural Network approaches of data mining that are completely based on content of emails. Proposed method are based on data mining approach, that provides an anti spam filtering technique that segregate spam and ham emails from large dataset. Methodologies that are used for the filtering methods are machine learning techniques using ANN and Bayesian Network based solutions. This approach practically applied on Trec07 dataset.

Keywords

Spam Filtering, Feature Selection, Stemming, Features Reduction, Naive Bayes, Neural Network, MLP.

1. INTRODUCTION

The continuous growth of internet technology, there are many possible ways of communication. In email system we receive daily different messages in bulk. Spam emails is a big problem which have several harmful affects such as wastage of user time, economic loss, loss of work productivity, extend virus, Trojans and degrade users trust. Spam stands for “Self Promotional Advertising Messages” but now the most popular definition is “unsolicited bulk mail” which causes email system overload in bandwidth and server load capacity which results in increase annual cost [14]. Phishing spam emails are also serious threat for security of end users that try to get personnel and confidential information like passwords and account numbers through spoof messages from on-line business transactions.

Spam filtering techniques are classified to segregate ham and spam emails. These techniques mainly focus on three levels as email address, the subject of message and message contents. Content based spam filtering is one of the most effective solutions to detect spam. It is based on features selection and text classification methods such as Decision Tree, Naive Bayesian classifier, Random Forest, Neural Network and SVM etc [7] [22] [23].

The goal of this paper is to compare the different spam filtering technique using the naive bayesian classifier and neural network [2][9]. Therefore, it is important to understand the spam filtering based on the best-practice solutions of present days; we compare the different performance metrics using the bayesian classification and neural network approach.

1.1 Source of Spam

Source of spam are Social Networking [17], Botnet, Internet Chain Process, Backscatter [21], Unsecured Networks and Open Relays etc.

1.2 Type of Spam

There are various types of spam. Some of spam are as: Phishing Mails, Email Scam, Trojan, Web Spam and Attachment Spam [6] etc.

1.3 Attacks on Spam Filter

There are various types of attacks some are: Tokenization attack and Text obfuscation attacks etc.

2. SPAM FILTERING TECHNIQUES

All There are various techniques available to spam filtering such as: Origin based filtering, Filtering based on traffic analysis, rule based spam filtering and Content based spam filtering [4] [7] [14].

Content based spam filtering is the most effective filtering technique, it is based on body of the email, body of the email is user interactive part because user always found their interesting items in content of emails. Content based spam filtering happens after a message received. This is based on known keywords or words are presence in the subject and body of the message. This techniques use different classification techniques that are:

2.1 Naïve Bayesian

A naive bayes classifier is a simple probabilistic classifier that is based on applying bayes theorem with strong (naive) independence assumptions. A more descriptive term for the probability model would be independent feature model [9] [19].

Bayes Theorem: Prob (B given A) = Prob (A and B)/Prob (A).

2.2 Decision Tree

In decision tree structure, each internal node (non leaf node) denotes a test on an attribute, every branch represents an outcome of the test data, and each leaf node holds a class label. The topmost node in a tree is the root node. There are various algorithms available for making decision tree, such as CART, ID3 and C4.5 etc. These are the greedy (i.e. non-backtracking) approach in which decision tree construct in top down divide and conquer manner.

2.3 SVM

In SVM is a new method for the classification of both linear and non-linear data. SVM are supervised learning models and it associated with learning algorithms that analyse data and recognize patterns [10]. The basic SVM takes a set of input data, for each given input, which has two possible class forms the output making it a non-probabilistic binary linear classifier.

2.4 Neural Network

A neural network is a set of connected input or output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural Network learning is also referred to connections between units [2] [22] [23].

3. RELATED WORK

Classification techniques have been applied in textual as well as image spam filtering process. During literature survey, we can see a proposed method where variant of naive bayes classifier have been applied for spam detection [2]. A compared classification strategy including Naive Bayes, Neural Network, Decision Tree and SVM were tested on different dataset on emails [20]. In which, J48 and NB classifier provides better results compare to NN and SVM. A textual classification method defined by K-NN and Genetic Algorithm for solving clustering problem [1]. A suggestion to combine cluster analysis based on sparse representation with clustering algorithm also provides spam detection.

4. PROBLEM FORMULATION

Email facilities are misusing for distributing unsolicited /inappropriate messages and documents by the hacker also known as spammer. The spam can be sent with almost no cost to the sender. In fact, others are paid the costs associated with the spam, such as the Internet Service Provider (ISP) and the receiver. Besides, it is difficult to have a legal action against spammers for preventing the receipt of spam within that jurisdiction. When this situation occurs, user will face a lot of troubles in receiving mail from others because the size of mail account is limited as well as user cannot send his mail out due to mail traffic. Moreover, user will waste much time to clean out the mailbox if he does not fix any device or software, which can detect whether the mail is junk mail or real mail. Therefore, the spam filter is needed in order to let the system to check the e-mails before downloading them. In other words, spam is harmful because it utilizes resources for other tasks, such as bandwidth, screen area, disk space, and user’s time. In addition, spam can be disreputable or entire illegal. For instance, various frauds, illegal products, and other inappropriate materials are advertised via spam.

Furthermore, user will feel difficult to search his desired e-mails if someone broadcasts unsolicited mass e-mail or news group postings simply because he wants to spread messages. This is referring to the “signal-to-noise ratio”. The purpose of spam filter is to help user to keep the Internet useful information readily available and keeps ”junk mail” to a minimum level. The main problem of the existing filter’s software is that they cannot be trained and learn instead of fixing a set of filter’s rules. It is tedious and difficult to construct robust rules to detect the naturally changeable junk mail too.

Therefore, our main problem is find best solution to detect spam emails with better accuracy in low cost and secure our system with viruses, phishing attacks, save user time and gain belief user trust.

5. OBJECTIVE AND SCOPE

The objective of this project is to classify and make analysis of spam and non-spam (ham) through using ANN models, such as multilayer perceptron and comparison of it with nave bayesian classifier. This research focuses pattern classification of e-mail content in order to determine whether it is a spam or a non-spam. When a set of data samples is given, the network will carries out training to learn the pattern of e-mails. The trained network (filter) has to decide on which type of dataset categories (spam versus non-spam) could be matched most closely when testing with the test set. The test set, which indicates advertisement, business’s information, pornographic issues etc and will be classified as spam (filtered out). The rest of the mails are classified as ham mails. There are two specific objectives in this project:

- To implement the ideas of multi-layer perceptron network for spam filtering.

- To evaluate the performance of multilayer perceptron neural network and naive bayesian models using keywords selection method as well as to quantify their results by statistical measures.

The scope of this study is focus on specified ANN model as mentioned previously, which are the multilayer perceptron and the naïve bayesian classification. The architectures and learning algorithms of ANN models in classification mails problem will be investigated. The trained network that obtained from training phase will be used in testing phase. Then the comparison of both models will be analysed. Besides, the project also concerns about Naive Bayesian classifier, which famous applied to spam detection application. However, it is mentioned in theoretically.

6. PROPOSED WORKING MODEL

The proposed methodology will be used for implementation this proposed work for spam filtering will be shown with the help of the following flow chart [Fig 1]. Proposed working model is based on data mining approach for classify ham and spam emails. It has data selection, data pre-processing, data classification and data analysis. In data selection we are working on Trec07 dataset, these provided by us government [12]. Data pre-processing has achieved by feature extraction [13] [15], stop word removing [18], stemming [5] [3] [11] and feature reduction [15] [16] techniques and results saved in .arff file in matrix form. For data classification here we are applying naive bayes classifier and multilayer perceptron (A technique of NN). In last we compare the results and understand what would be better technique that might me applied for detects spam.

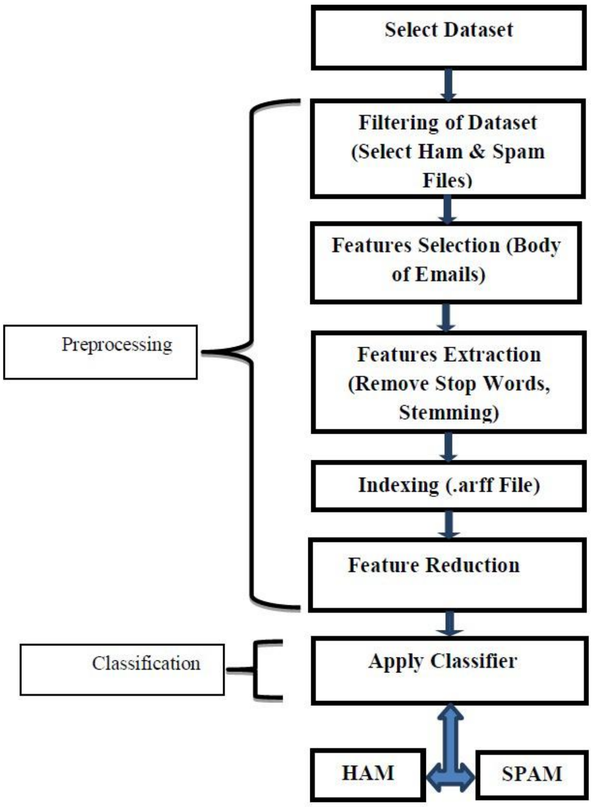
Select Dataset
Filtering of Dataset (Select Ham & Spam Files)
Features Selection (Body of Emails)
Features Extraction (Remove Stop Words, Stemming)
Indexing (.arff File)
Feature Reduction
Preprocessing
Classification
Apply Classifier
HAM
SPAM

Fig 1 Proposed Working Model for Ham and Spam Classification.

7. IMPLEMENTATION AND EXPERIMENTAL ANALYSIS

7.1 Experimental Setup

Experiments were carried out on Trec07 data sets, which are publicly available. We are using WEKA data mining tool for analysing the results. System configuration was 4GB RAM, Core2duo 2:00G$_H$ processor having window 7 installed.

7.2 Implementation by using MLP

After successfully practically implementation of pre-processing step of data mining, .arff file open in WEKA data mining analysing tool and select MLP classifier for classifying data.

Following input parameter are used to train the MLP network, see in Table 1:

Table 1. Input Parameter for Reduce Dataset

Total Number of Attributes Used	99
Total Number of Instances Used	100
Number of Hidden Layer Used	1
Learning Rate	0.3
Momentum	0.2

Following output are given by the uses of Multilayer Perceptron, see in Table 2:

Table 2. Output Using MLP Applied on Reduced Dataset

Number of Epoch (Iteration)	100
Correctly Classified Instances	93
Incorrectly Classified Instances	7
Time Taken to Build Network Model	10.94 Sec
Error Per Epoch	0.021445

Confusion Matrix generated for the reduced dataset by using MLP:

```
        Confusion Matrix

      a    b   ←   classified as

     48    2    |a = ham

      5   45    |b = spam
```

7.3 Implementation by using NB Classifier

After successfully practically implementation of pre-processing step of data mining, .arff file open in WEKA data mining analysing tool and select Naive Bays classifier for classifying data items.

Following output produced while applying the Naive Bays classifier on reduced dataset, see in Table 3:

Table 3. Output Using NB Classifier on Reduced Dataset

Total Number of Instances	100
Correctly Classified Instances	88
Incorrectly Classified Instances	12
Time Taken to Build Network Model	0.14 Sec

Confusion Matrix generated for the reduced dataset by using NB Classifier:

```
        Confusion Matrix

      a    b   ←   classified as

     46    4    |a = ham

      8   42    |b = spam
```

7.4 Result Evaluation

We are measured the precision, recall and accuracy, after then check the effectiveness of the performance of both classifier on reduced dataset. Comparison of both classifier, see in Table 4:

Table 4. Performance Measurement of Both Classifiers

Approach	Accuracy	Precision	Recall	Time Taken To Build Model
MLP	93	93.2	93	10.94 Sec
NB	88	88.2	88	0.14 Sec

We observe that MLP classifying filtering approach at user level better classify ham as well as spam emails but it take more time to build model comparative NB classifier.

8. CONCLUSION AND FUTURE WORK

8.1 Conclusion

The application of neural networks to detecting spam is definitely something that can and is being pursued as a viable option. However, to obtain optimum performance, we have to do sufficient amount of data analysis. Also, this data analysis has to be general so as to block a wider variety of spam.

The basic principal used in any spam filtering technique, whether heuristic or keyword based is identical: spam messages generally look different than good messages and detecting these differences is a good way to identify and stop spam. The difference between these technologies really comes down to the problem of distinguishing between these two classes of email. The neural networks approach is more refined, more mathematical and potentially far more accurate and reliable in accomplishing this task.

Although no single technology can achieve one hundred percent spam detection with zero false positives (despite vendor claims), machine-learned heuristics in general and neural networks in particular have proven extremely effective and reliable at accurately identifying spam and minimizing errors to an acceptable minimum.

8.2 Future Work

The following future work can be done on the basis of this project:

- Applying of the different network such as back propagation network, RBF network to detect spam.

- We can also do implementation on two or more hidden layer to provide robustness.

- Fuzzy logic is another important content-based method to distinguish spam. A fuzzy logic approach to the same problem can bring some new insights into the problem.

- A combinational approach can be used to achieve higher classification rates (using header filters, content based filters and user specific information).

9. REFERENCES

[1] Rasim M Alguliev, Ramiz M Aliguliyev, and Saadat A Nazirova. Classification of textual e-mail spam using data mining techniques. Applied Computational Intelligence and Soft Computing, 2011:10, 2011.

[2] T.A. Almeida and A. Yamakami. Content-based spam filtering. In Neural Networks (IJCNN), The 2010 International Joint Conference on, pages 1-7, 2010.

[3] Veena H Bhat, Vandana R Malkani, PD Shenoy, KR Venugopal, and LM Patnaik. Classification of email using beaks: Behaviour and keyword stemming. In TENCON 2011-2011 IEEE Region 10 Conference, pages 1139-1143. IEEE, 2011.

[4] Godwin Caruana and Maozhen Li. A survey of emerging approaches to spam filtering. ACM Computing Surveys (CSUR), 44(2):9, 2012.

[5] Duke education. Stemming code. URL http://www.cs.duke.edu/courses/compsci308/cps108/fall07/code/stemmer/ code.pdf.

[6] George Giannakopoulos, Petra Mavridi, Georgios Paliouras, George Papadakis, and Konstantinos Tserpes. Representation models for text classification: a comparative analysis over three web document types. In Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, page 13. ACM, 2012.

[7] YiShan Gong and Qiang Chen. Research of spam filtering based on bayesian algorithm. In Computer Application and System Modeling (ICCASM), 2010 International Conference on, volume 4, pages V4-678-V4-680, 2010.

[8] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[9] Biju Issac and Wendy J Jap. Implementing spam detection using bayesian and porter stemmer keyword stripping approaches. In TENCON 2009-2009 IEEE Region 10 Conference, pages 1-5. IEEE, 2009.

[10] R Kishore Kumar, G Poonkuzhali, and P Sudhakar. Comparative study on email spam classifier using data mining techniques. In Proceedings of the International MultiConference of Engineers and Computer Scientists, volume 1, 2012.

[11] M.F.Porter. Porter stemming algorithm. URL http://tartarus.org/martin/PorterStemmer/def.txt.

[12] NIST (National Institute of Standard and Technology) US govt. Trec07 dataset. URL http://trec.nist.gov/data/spam.html.

[13] R Parimala and R Nallaswamy. A study of spam e-mail classification using feature selection package. Global Journal of Computer Science and Technology, 11(7), 2011.

[14] Noemi Perez-Diaz, David Ruano-Ordas, Florentino Fdez-Riverola, and Jose R Mendez. Sdai: An integral evaluation methodology for content-based spam filtering models. Expert Systems with Applications, 2012.

[15] Aziz Qaroush, Ismail M Khater, and Mahdi Washaha. Identifying spam email based-on statistical header features and sender behavior. In Proceedings of the CUBE International Information Technology Conference, pages 771-778. ACM, 2012.

[16] Alessandro Rozza, Gabriele Lombardi, and Elena Casiraghi. Novel ipca based classifiers and their application to spam filtering. In Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on, pages 797-802. IEEE, 2009.

[17] Zac Sadan and David G Schwartz. Social network analysis of web links to eliminate false positives in collaborative anti-spam systems. Journal of Network and Computer Applications, 34(5):1717-1723, 2011.

[18] Onix text retrieval toolkit. Stop word lists. URL http://www.lextek.com/manuals/onix/stopwords.html.

[19] Jiansheng Wu and Tao Deng. Research in anti-spam method based on bayesian filtering. In Computational Intelligence and Industrial Application, 2008. PACIIA '08. Pacific-Asia Workshop on, volume 2, pages 887-891, 2008.

[20] Seongwook Youn and Dennis McLeod. A comparative study for email classification. In Advances and Innovations in Systems, Computing Sciences and Software Engineering, pages 387-391. Springer, 2007.

[21] Backscatter. Source of Spam. http://www.spamresource.com/2007/02/backscatter-whatis-it-how-do-i-stop-it.html

[22] Owen Kufundirimbwa and Richard Gotora. Spam detection using artificial neural network. JPESR, ISSN:2315-5027, Vol 1, Issue 1, PP 22-29, June 2012.

[23] Laurence Fausett. Fundamental of Neural Network. Architecture, Algorithms and Application, Page 24-26, 2006.