An Independent Component Analysis Based Approach for Filtering Spam

Divya Arora[1], Gunjan Rehani[2]

[1]MTech. Student, SPITM, Sonepat;
[2]A.P.(CSE Depth.), SPITM, Sonepat;

Abstract: Spam NET has been used for detecting spam using neural network. While using this approach, PCA is used for feature detection. PCA is dimensionality reduction technique which takes strongly correlated components as input vector. False positive rate is very sensitive to differences in features of mails. So to eliminate this criteria, FastICA is proposed that can be used which takes independence of source signals into considerations and handle the task of spam filtering efficiently. In this paper we will present, the algorithm its basics and also compare the results of simulation with previous approach.

Keywords: Independent, neural, classifier, extractor, fault tolerance, spam filtering.

I. INTRODUCTION

E-mail has become popular means for personal and business communication due to its fast and free availability as well as low or free cost. But several people and companies misuse this facility to distribute unsolicited bulk messages that are commonly called as spam mails. Spam emails may include advertisements of drugs, software, Nigerian scam, adult content, health insurance or other fraudulent advertisements. Spammers can collect email addresses from chat rooms, some AOL profiles, public networking websites, customer lists, from white and yellow pages, newsgroups, worms etc. Sometimes little bit information about target system is enough to get the email address of him.

Spam detection problem is becoming more serious now days. It consumes more than half bandwidth of mailboxes. Spam frustrates, confuse and annoy email users by wasting valuable resources and time. Spam even provides ways for phishing attacks and distributing harmful content such as viruses, Trojan horses, worms and other malicious code. Without a spam filter, one email user might receive over hundreds of mails daily and find that most of them are of spam category. The spam mails are with no use of email users. Due to this, serious attention has given to this issue in mailboxes. Several technical solutions like commercial and open-source products have been used to alleviate the effect of this issue.[1]

Spam filtering can be of two types: Non-machine learning based and Machine learning based. Early anti-spam techniques like white list, blacklist and set of keywords like “you have won” fall into non-machine learning based techniques. White list contains list of safe senders whereas blacklist contains list of blocked systems or users. As these methods are dependent on lists so these can be easily resolved by spammers. These methods also require manual update and sometime these methods misclassify legitimate mail as spam mail which is more dangerous than no filtering. The British Computer Society (BCS), concluded that misclassification of mails may waste over five million working hours a year by users.

On the other hand, machine learning techniques first analyze the message content and then perform classification of mail as spam or ham. Various machine learning anti-spam methods are:

- Support Vector Machines
- Memory based learning
- Ripper rule based learning
- Boosting Decision based learning
- Bayesian Classifiers
- Fuzzy similarity[2,3]

Although previous study has reported promising detection accuracies but still false positive rate is high. Various techniques have been developed to combat the problem of spam but still effective and efficient technique is required which will have very low false positive and false negative. As single technique is not sufficient to combat this issue so multiple methods should be used. In this research, FastICA will be used for spam detection with neural networks. FastICA is signal processing technique used for analysis of several types of data and feature extraction. It combats the problem faced during using PCA. False positive rate is very sensitive to small differences in the number of principal components in normal subspace in case of PCA. So FastICA can be used to detect spam even those which are independent of each other and have nothing in common. [4,5]

Neural network is information processing system that works on biological nervous system. In this system,

large numbers of processing elements are connected together work to resolve a specific problem. Neural networks build a model that states complex relationship between inputs and outputs. Features of neural network are:

- They are extremely powerful computational devices.
- Massive parallelism makes them very efficient.
- They can learn and generalize from training data – so that there is no need for enormous feats of programming.
- Neural networks are fault tolerant it means graceful degradation in biological systems.
- They are very noise tolerant so they can cope with situations where normal symbolic systems would have difficulty.
- In principle, neural network can do anything a symbolic/logic system can do and more.
- Each neuron in neural network does some amount of information processing.
- It derives input from some other neuron and in return gives its output to other neuron for further processing.

II. BASICS OF APPROACH

As various types of neural networks have already used for classifying mail as spam or ham mail. SpamNET has been used for detecting spam using neural network. While using this approach, PCA is used for feature detection.

PCA is dimensionality reduction technique which takes strongly correlated components as input vector. So to eliminate this criteria, fastICA can be used which takes independence of source signals into considerations. FastICA can be efficiently used for detecting spam originating from varying number of users. FastICA is an algorithm for independent component analysis. Its theoretical consideration is: Let's take a data matrix X which includes independent components i.e. X=SA

where columns of S contains independent components and A is linear mixing of matrix. ICA will try to find out un-mixing matrix W where XW=S.

This matrix maximizes the independence of sources. Spam NET has been used for categorizing mail as spam or ham message. Architecture of spamNET has shown in figure. Extractor is used to covert email into appropriate format that will be fed into fastICA for feature extraction. SpamNET has also short circuit path to output which directly extract out spam mails from useful mails.

Extractor Module

SpamNET contains extractor module which extract out mails on the basis of keywords and concepts. Extractor module will find out following features in mail:

- Any invisible word
- Currency value like $1,0000
- Symbols like *, #, % between words.

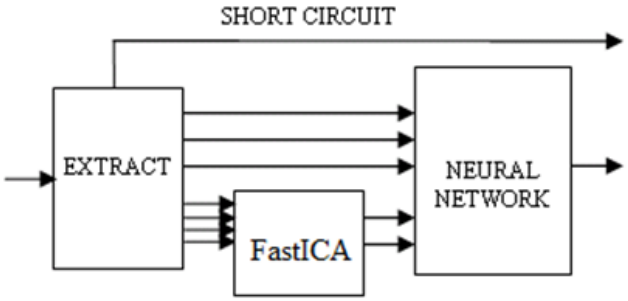

Figure 1 Architecture of SpamNET

If these features are detected by extractor module then mail is directly declared as spam or ham mail by using short circuit. After this module, a matrix is formed which contains frequency of occurrence of words in mails. To make this matrix, FastICA algorithm is used. As convergence is very fast of FastICA, it will adapt itself according to continuously changing environment. Main concept behind using FastICA is that mails may come from varying regions, users. There is possibility that existing spam mails are different from incoming mails which may be spam as well as ham mails. Existing techniques rely on information stored about existing spam mails. These techniques compare keywords or use knowledge base for detecting spam. But with the use of FastICA, spam detection is not relied on existing knowledge base.

Fastica Module

FastICA is statistical signal processing technique which is used for feature extraction, separating mixed data and voice, analysis of several types of data, sensor signal processing. Before implementing this approach, one thing is need to be concerned about that spam will be always of different type. Spam may come from different users or different regions. So algorithm should have capability to detect incoming mail independently of other already detected spam. FastICA will add simple orthogonal zing projection to categorize each time a different mail as spam or ham mail. It assumes that every incoming mail is statistically independent. It minimizes mutual information so that model has ability to adapt itself according to environment.

ICA will make matrix, in which words will be on vertical axis and mails will be on horizontal axis. ICA

can be used for finding underlying factors or features which will indicate that specific mail is spam or ham mail. For extractor module, previous database of mails is required to detect spam. But ICA will perform detection independently. It will detect weird behavior of mails. Then, result will feed into neural networks which will detect spam and extracts out.

Neural Networks: Neural network swaps between two values 1 and -1. 1 indicates that mail is spam mail and -1 indicates mail is not spam. Neural network will detect punctuation signs, capital words, currency values and percentage of colored text. Estimating ICA in original, high-dimensional space may lead to poor results. Measure of non-gaussianity is given by negentropy and computation of negentropy is very difficult.

III. ALGORITHM

Following are the various steps of algorithm[6,7, 8]:

1 Preprocess The Data

Before the FastICA algorithm can be applied, the input vector data $\mathbf{X}$ should be centered and whitened.

a) Centering the data

The input data $\mathbf{X}$ is centered by computing the mean of each component of $\mathbf{X}$ and subtracting that mean. This has the effect of making each component have zero mean.

Thus: $\mathbf{x} \leftarrow \mathbf{x} - E\{\mathbf{x}\}$

b) Whitening the data

Whitening the data involves linearly transforming the data so that the new components are uncorrelated and have variance one. If $\widetilde{\mathbf{x}}$ is the whitened data, then the covariance matrix of the whitened data is the identity matrix:

$$E\{\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T\} = \mathbf{I}$$

This can be done using eigenvalue decomposition of the covariance matrix of the data: $E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{E}\mathbf{D}\mathbf{E}^T$, where $\mathbf{E}$ is the matrix of eigenvectors and $\mathbf{D}$ is the diagonal matrix of eigenvalues. Once eigenvalue decomposition is done, the whitened data is:

$$\mathbf{x} \leftarrow \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x}$$

2 Single Component Extraction

The iterative algorithm finds the direction for the weight vector $\mathbf{w}$ maximizing the non-Gaussianity of the projection $\mathbf{w}^T\mathbf{x}$ for the data $\mathbf{X}$. The function $g(u)$ is the derivative of a nonquadratic nonlinearity function $f(u)$. Hyvärinen states that good equations for f (shown with their derivatives g and second derivatives g') are:

$$f(u) = \log\cosh(u); \quad g(u) = \tanh(u); \quad g'(u) = 1 - \tanh^2(u)$$

$$f(u) = -e^{-u^2/2}; \quad g(u) = ue^{-u^2/2}; \quad g'(u) = (1-u^2)e^{-u^2/2}$$

The first equation is a good general-purpose equation, while the second is highly robust.

1. Randomize the initial weight vector $\mathbf{w}$
2. Let
$$\mathbf{w}^+ \leftarrow E\{\mathbf{x}g(\mathbf{w}^T\mathbf{x})\} - E\{g'(\mathbf{w}^T\mathbf{x})\}\mathbf{w}$$
, where $E\{\cdots\}$ means averaging over all column-vectors of matrix $\mathbf{X}$
3. Let $\mathbf{w} \leftarrow \mathbf{w}^+ / \|\mathbf{w}^+\|$
4. If not converged, go back to 2

3 Multiple Component Extraction

The single unit iterative algorithm only estimates one of the independent components, to estimate more the algorithm must repeated, and the projection vectors decorated. Although Hyvärinen provides several ways of decorating results the simplest multiple unit algorithm follows. $\mathbf{1}$ indicates a column vector of 1's with dimension *M*.

Algorithm FastICA

Input: C Number of desired components

Input: $\mathbf{X} \in \mathbb{R}^{N \times M}$ Matrix, where each column represents an *N*-dimensional sample, where $C < N$

Output: $\mathbf{W} \in \mathbb{R}^{C \times N}$ Un-mixing matrix where each row projects X onto into independent component.

Output: $\mathbf{S} \in \mathbb{R}^{C \times M}$ Independent components matrix, with M columns representing a sample with C dimensions.

for p **in** 1 to C:

$\mathbf{w_p} \leftarrow$ *Random vector of length N*

while $\mathbf{w_p}$ changes

$$\mathbf{w_p} \leftarrow \frac{1}{M}\mathbf{X}g(\mathbf{w_p}^T\mathbf{X}) - \frac{1}{M}g'(\mathbf{w_p}^T\mathbf{X})\mathbf{1}\mathbf{w_p}$$

$$\mathbf{w_p} \leftarrow \mathbf{w_p} - \sum_{j=1}^{p-1} \mathbf{w_p}^T\mathbf{w_j}\mathbf{w_j}$$

$$w_p \leftarrow \frac{w_p}{\|w_p\|}$$

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_C \end{bmatrix}$$

Output:

Output: $S = WX$

IV. RESULT AND ANALYSIS

With the work done when the approach is implemented in MATLAB, we have concluded that the ICA algorithm is one of the efficient algorithm for the data mining and it quite better and efficient than PCA algorithm. We have not just used the ICA algorithm but also we have combined the algorithm with the NEURAL NETWORKS which is getting used these days rapidly for enough development .NEURAL NETWORKS performs the work in segments and the final result is obtained by combining all the segments. Our work concludes that with this hybrid structure, the accuracy can be enhanced to a much better extent as in our result we are getting an average result of 94 to 97 percent based on the number of files selected . Training section includes the Independent Component Analysis algorithm which converts the text into digital signal. The result of ICA algorithm is passed to the neural network. Then the neural network decides that how many iterations it would take to give you the best probabilities result generally 5 to 10 iterations are more than sufficient for neural network to decide what exactly the results analysis could be.

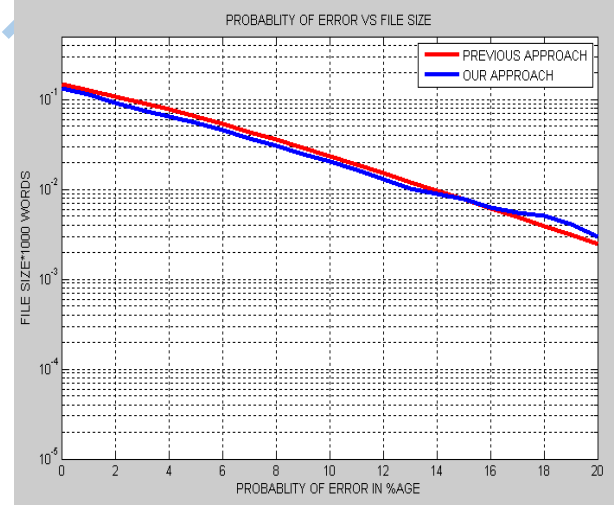

Figure 2. Comparison of PCA and our approach

V. CONCLUSION AND FUTURE SCOPE

With the advent of the study we concluded that our algorithm is much better than earlier approach and very effective to tackle spam mail. Also can filter them efficiently so that appropriate results can be obtained. There are several other algorithms like genetic algorithm, BFO, ACO algorithm etc. is a quite effective algorithm if it can be combined with NEURAL NETWORKS. The drawback of this system is that it has no rule set for processing. In future, if somebody combines Fuzzy Logic with Neural Network this drawback can be removed and the result would be efficient. If somebody wants to experiment, Neural Network classification can be replaced with a number of other Classification for better result.

VI. REFERENCES

[1]. Grigorios Tzortzis and Aristidis Likas, "Deep Belief Networks for spam filtering", 19th IEEE International Conference on Tools with Artificial Intelligence, GR 45110, Ioannina Greece (2007)

[2]. Gaurav Kumar Tak and Shashikala Tapaswi, "Query Based approach towards spam attacks using artificial neural network", International Journal of Artificial Intelligence & Applications, October 2010

[3]. Alex Brodsky (Canada) and Dmitry Brodsky (USA), "A distributed content independent method for spam detection".

[4]. Hyvarienen and E.Oja, Independent Component Analysis and Applications, Neural Networks 13(4-5):411-430, 2000

[5]. Dominic Langlois, Sylvain chartier and Dominique Gosselin, An introduction to Independent Component Analysis: Infomax and FastICA Algorithm (2010)

[6]. Sasmita Kumari Behra (2009) "FastICA for blind source separation and its implementation", Rourkela

[7]. Ann Nosseir , Khaled Nagati and Islam Taj-Eddin," Intelligent Word-Based Spam Filter Detection Using Multi-Neural Networks",IJCSI, Vol. 10, Issue 2, No 1, March 2013

[8]. V.Zorkadis, M.Panayotou, D.A.Karas, "Improved Spam e-mail Filtering Based on Committee Machines and Information Theoretic Feature Extraction", Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July-Aug,2005.

