Middle-East Journal of Scientific Research 23 (5): 874-879, 2015
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2015.23.05.22221

A Probabilistic Neural Network Based Classification of Spam Mails Using Particle Swarm Optimization Feature Selection

[1]S. Kumar and [2]S. Arumugam

[1]Anna University, Chennai, India
[2]Nandha Engineering College, Erode, India

Abstract: Email has gained the explosive growth in the communication of people across the world. This worldwide communication also has some disadvantages like Spam mails. The spammers spread the useless, unwanted mails and even malicious contents to the usersemails. This increasing number of spam mails increases the need for the spam detection architecture with the machine learning classification. The proposed spam detection architecture composed of a feature selection process to minimize the error rate, a redundancy removing method and finally a classification system for categorizing the spam mails from the legitimate mails. The incoming mails are preprocessed by using the three traditional steps such as Tokenization, Stemming and the Stop Word Removal. The Vector Quantization (VQ) process is utilized to remove the redundancy in both the training and preprocessed data. Then the preprocessed redundancy removed training and testing data are given to the feature selector called the familiar Particle Swarm Optimization (PSO) algorithm which mines the optimal features suitable for the classification. Finally, along with the selected features, the Probabilistic Neural Network (PNN) classifies the spam mails from the legitimate mails with more accuracy and precision.

Key words: Spam Mails · Legitimate Mails · Spammers · Tokenization · Stemming · Stop Word Removal · Vector Quantization · Particle Swarm Optimization · Probabilistic Neural Network

INTRODUCTION

Email spam becomes the unavoidable message that the mails users always identified as a problem. This spam mail is not only causing disturbance to the users, it also becomes dangerous to the users because of its content that may be a spiteful code [1]. The mail users spend more time to sort these junk mails as it results in more time consuming and subject to cost efficiency. These unsolicited messages gobble more storage space and time to maintain and transmit. Since 1990, the spam volume has started to grow and reached between 50-80% of total email traffic.To overcome these spam mail issues, many spam detection methods, namely keyword based filtering, email abstraction and heuristic based filtering are implemented. These methods are good in their performance and time consuming rate, but still experiences specific fall in accuracy [2].

In order to tackle this drawback, instead of these methods, the machine learning algorithms are proposed. In the machine learning process for spam filtering, the user defines “what isspam” and forms the training dataset [3]. Themachine learning spam filter can able to learn through those training datasets. Some of the Machine Learning Algorithms are well-versed in the text message classification which is a process of categorizing the text into one of the categories available. Among the different classifiers that do not provide the accurate prediction, the probabilistic classifiers like Bayesian classifier brings the better accuracy [4]. The paper utilizes the probabilistic method which is based on the Bayesian filter for classification and a swarm intelligence algorithm for the feature selection process which makes the classification more reliable.

The proposed system utilizes the Probabilistic Neural Network (PNN) which is a machine learning algorithm in order to assign the text message to a particular category (spam or non-spam). The incoming mails may be large in size that leads to the high processing time and consume more storage space for processing. Hence the data (mail) preprocessing and feature selection operation has to be done before the classification process. The data

Corresponding Author: S. Kumar, Anna University, Chennai, India.

preprocessing is performed using the traditional steps such as Tokenization, Stemming and the Stop Word Removal [5]. The tokenization is the process of reducing the message into discrete words for making the message assessing easier. Stemming refers to the reduction of tokenized words in its real form. Then the stop word removal is a process of elimination of the unwanted words that increase the size of the vector space. The preprocessed data and the training dataset are processed using k-means algorithm for the issue of Vector Quantization (VQ) [6]. Considering the individual bag full of words as features in accounting, the optimal features (words) selection process is accomplished by using the Particle Swarm Optimization (PSO) [7].It removes the irrelevant features in terms, it reduces the dataset size eventually.The final optimized features are given to the Probabilistic Neural Network (PNN) [8] which classifies the message into either of the two categories (non-spam or spam) using the inbuilt sub layers and the reduced training dataset.

Related Work: The spam machine detection is performed by using the well-known Page Rank and HITS algorithms. These algorithms provide better results when combined with the K-Means clustering algorithm. [9] introduce a new method for the weight selection process of the algorithms. Since the weight precision brings the result with more accuracy. The K-Means algorithm plied in the network having huge traffic and cluster the IP address into normal and anomalous and these IP addresses are given to the next set of algorithms that detects the spam machines in the huge traffic network.

Phishing is a method of stealing the information of the web user account, identity information and the log on credentials. This leads to the hack the users private information, many techniques are present to over the phishing problem. But still many attacks such as retrieving the bank details, pin numbers leads to the necessity of phishing filters that provides global security and increase the communication security. [10] Lists the features that recognize the phishing mails for depth analysis of the incoming emails and find the indicators and logos that are hidden in the mail.

The information retrieval is a big issue when considering the huge database. The text extraction for such huge database is done using many algorithms which are well-versed in the similarity computation. When the clustering is a process of information retrieval, then the similarity measures plays a vital role in that. Therefore, the paper [11] proposes a new SR-Method which finds the similarity of two documents taken for text clustering. This system utilizes four different parameters and the measure using the SR-method is given to the K-means clustering algorithm to group the document. The final results show that the proposed method provides 85% accuracy than the existing methods.

The paper [12] introduce a new data mining tool called TANAGRA which examines the data set and provide the prominent email spam classifier. In order to make the classification easier and less time consuming, the efficient features extraction and feature selection methods are implemented to bring out the optimal features. Then different classification methods are used on the dataset. This way of assessing the methods results in the finding the best email spam classifier based on low error rate, recall and precision.

The paper [13] utilized the rough set theory for a spam detection method implemented in the chat system. The spam messages become more suspicious issue for the internet communication system with unwanted message that disturbs the people in web. hence the paper discuss the previously used techniques and methods for spam detection and find new efficient technique based on the rough set theory and performs better than the existing methods.

Web search engine page becomes the starting page of every browser. The users simply give the keywords to search and the feature selection methods are used here to render the needed feature. In particular the powerful feature clustering methods are used to minimize the dimensionality. The author [14] suggested the id's of group the features into clusters using FAST algorithm. The data can be uploaded from any database independent of the document format and finally the document is converted into XML format and then uploaded.

Proposed System

Mail Text Cleansing: The incoming mail has to be processed under many refining stages where the noises and irrelevant data are removed. The spam mail classification from the legitimate mail is a complex process. In order to make this process simple, the data preprocessing is incorporated in the beginning of the classification process. The incoming mail is a text document and cleansing of this text is carried out using three methods in sequence such as (i) Tokenization, (ii) Stop word removal and (iii) Stemming [15].

- The tokenization is a process of converting the mail text into a number of individual words called tokens. The tokenizer in this process is capable of recognizing the HTML tags; and parses the text by

utilizingthe URL encoding methods and eliminates the symbols and punctuations. The individual tokens are included in the vector space that helps the classification process. This process is just extracts the words from the text document without considering its importance. This may lead to the necessity of removing the irrelevant words from the text document (mail body).

- The stop word removal is an activity that eliminates the particular functional words that are useless and having no meaning in the process of text mining and information retrieval. In addition, the stop word removal finds and removes the frequent preposition, articles and formal grammatical words. This process has the dictionary that has the list of around 300-400 words and checks whether those words exist in the document. The matching words with the list are eliminated from the mail text document, which reduces the vector space size considerable.
- The stemming process minimizes the each individual word to its root word. In simple words, it removes the suffix and prefix of the words in the mail text. The main concept behind the stemming is that there are many words that may have the same root word, for instance, the word "receive" is the stem or root word for receive, received, receiver, receiving. While minimizing each with it's root word, the vector space size is minimized which in turn minimized the searching time of the classification in advance.

Redundancy Eliminationusing Vector Quantization: Thespam detection architectureincorporates the Probabilistic Neural Network (PNN) for the classification of legitimate mails and the spam mails. Although the PNN is a growing eminent tool for solving many classification problems, the training dataset of the PNN may be contaminated with duplications or redundancies. In order to solve this redundancy issue in advance, the most prominent vector quantization process is utilized. The vector quantization is a process of moving the large volume of features vectors from a space to a number of clusters where the k-means algorithm is used to remove the redundancy by computing the mean square error between the centroid and the cluster member [6].

Initially, the training dataset features and the preprocessed features are clustered separately into k groups based on their nearest neighbor condition. Then the cluster centroid is selected randomly in each cluster. The modification of k-means clustering algorithm is that instead of finding the optimal cluster centroid, the initially elected cluster centroid is compared with each and every cluster member. Compute the mean square error between each cluster member (feature) and the cluster centroid. Let x_n be the set of feature and is represented as $x_n = x_{n0}, x_{n1},\ldots,x_{nk-1}$ and let y_n be the centroid in a cluster. The mean square error between a single cluster member and the cluster centroid is computed by using the Euclidean distance between the two vectors as

$$D(x_n,y_n) = \|x_n - y_n\|^2 = \sum_{j=1}^{k}(x_{nj} - y_{nj})^2 \quad (1)$$

If the mean square error equals to zero, then it means that both the features are same and one of the features are removed from the vector space. Once all the clusters finish their search for redundancy, the mean square error for the all the cluster centroids are computed and the redundancies are removed.

Feature Selection: Feature Selection is a significant method in the process of pattern classification and it is also an issue of reducing the extracted features without reducing the classification accuracy. The objective of any feature selection algorithm is (i) to minimize the number of features and (ii) to maximize the classification accuracy [16]. Particle Swarm Optimization (PSO) is one of such algorithms and also it is an evolutionary computation technique. It spread the extracted features in the search space. The extracted features as a whole is called as a population and the individual features are called as particles. Each particle is free to fly over the search space with the given velocity to search for the best position. Therefore, each particle has a velocity and a current position. The current position in the search space is represented as $X_i^t = (x_1^t, x_2^t, \ldots, x_N^t)$, where $x_j^t \in [l_j, u_j], 1 \le j \le N$, and l_j and u_j lower and upper bound. The current velocity of a particle allows it to move in the specific direction and mostly the velocity forwards the particle to the best solution. The current velocity of the particle is represented as $V_i^t = (v_1^t, v_2^t, \ldots, v_N^t)$. The algorithm for the computing the reduced feature set using the PSO is shown in the Figure 1.

In the starting stage, the particles spread over the search space having N dimension. The fitness value is computed for each particle.The fitness value is computedfor all the particles in every iteration. If the fitness value of the current particle is better than the pbest value, then the current particle is termed as pbest and it position and fitness value is stored. Then the fitness value of the current particle is compared with all the particles in tsearch space. If the current fitness value is better than the global best fitness value (gbest), then

```
∀i: X_i ← randomPosition; ∀i: X_i ← randomVelocity;
fit ← bestFit(X);
globalbest ←fit;
pbest← bestpos(X);
∀i: P_i ← X_i
while (end criterion met)
        for i=1, i++, i=N
          if (fitness(i) > fit) /* for local best */
             fit ← fitness(i)
        pbest ←X_i
          if (fitness(i) > globalbest) /* for global best */
            globalbest ← fitness(i)
            gbest ← X_i;
        RF ← getReducedFeature (X_i) /* reduced features get after the
                                                    end condition met*/
UpdateVelocity();
UpdatePosition();
```

Fig. 1: Algorithm for reduced feature set computation using PSO

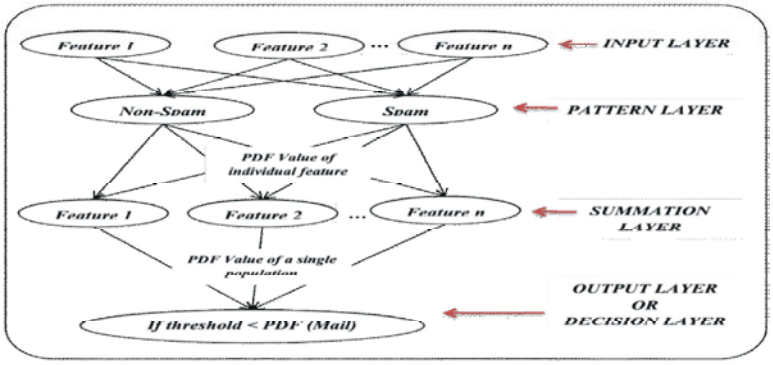

Fig. 2: Email classification using the Probabilistic Neural Network

the current particle fitness value is set as global best fitness value. The velocity and the position of the particle are known and updated by the following equations.

$$\text{Velocity}_i^{t+1} = V_i^t + c_1 r_1 (P_i - X_i^t) + c_2 r_2 (P_g - X_i^t) \quad (2)$$

$$X_i^{t+1} = X_i^t + \text{Velocity}_i^{t+1} \quad (3)$$

where r_1 and r_2 are the random functionswithin the range, c_1 and c_2 are the constants for speed up the process, P_g be the best position among all the particles and $P_i$denotes the previous best position of the i^{th} particle.

The process of finding the global best values is iterative until the end criterion is met. In the process of spam mail feature selection, the end criterion is the k number of features. Once the user specified number of features is obtained, then the iteration stops.

Classification: The Probabilistic Neural Networks (PNNs) [17] are a combination of feed forward neural network and the Radial Basis Function (RBF) which are closely related to the Bayes' decision theory that easily classifies the input pattern. Since it is based on the RBF, the pattern layer in the PNN is capable of estimating the Probability Density Function (PDF) of each feature. Basically the PNN is four-layer architecture comprises of an Input Layer, two hidden layers, namely Pattern and Summation Layer and finally the Decision (output) Layer. The architecture of the PNN is depicted in the Figure 2.

The input layer just gathers and distributes the input data (optimized features) to the neuron in the pattern layer. The pattern layer is built with the trained dataset of spam and non-spam classes and is formed based on the Radial Basis Function which can able to find the Probability Density Function of each feature. The PDF of each feature is computed as

$$f_k(X) = \frac{1}{2\pi^{p/2}\sigma^p} e^{-\frac{\|X - X_k\|^2}{2\sigma^2}} \quad (4)$$

where X be the testing data feature, X_k denotes the kth training data feature, p be the dimension of the input feature and I be the smoothing parameter. Once the PDF for all the optimized feature is computed by using the training dataset samples in the pattern layer, the summation layer gathers the spam and ham probability of each feature in separate neurons. Therefore the PDF of the particle population is computed as

$$g_i(X) = \frac{1}{(2\pi)^{p/2}\sigma^p}\frac{1}{n_i}\sum_{k=1}^{n_i} e^{-\frac{\|X-X_k\|^2}{2\sigma^2}} \quad (5)$$

Finally the decision layer analyzes the output of the summation layer with the PDF value and the pattern layer spam and ham probability of each feature of the population. The threshold value is fixed in this layer and compares the threshold value with PDF value along with the spam/ham probability. If the PDF value of the population (mail) is greater than the threshold value, then that mail is marked as spam mail. If the PDF value of the population (mail) is smaller than the threshold value, then that mail is marked as ham mail.

Result Analysis: In the proposed system, the incoming mails are preprocessing which in turn remove the noise and also extract the possible features. Then the extracted features after removing the redundancies will feed into the Particle Swarm Optimization (PSO) to minimize the features set. Here the PSO is compared with the Bat Algorithm (BA) and the Levenberg-Marquardt (LM). The Figure 3 shows that the PSO is acted best when compared with these algorithms. The PSO reduced the feature set in minimum time.

Minimizing the error rate is an essential task of classification process. The probabilistic neural network (PNN) minimizes the error rate in Figure 4 when compared with existing classification methods such asBasic Local Alignment Search Tool (BLAST) search and Bayesian Filter.

The Table 1 provides the details of total emails with the spam and non-spam mails considered inevaluating the proposed system.

Once the execution of the proposed system is completed, the spam and legitimate mails are classified. The Table 2 shows the classified emails using the proposed system.

Once the execution of the proposed system is completed, the spam and legitimate mails are classified. The table 2 shows the classified emails using the proposed system and gives 90% accuracy, 93.23% specificity and 91.42% sensitivity.

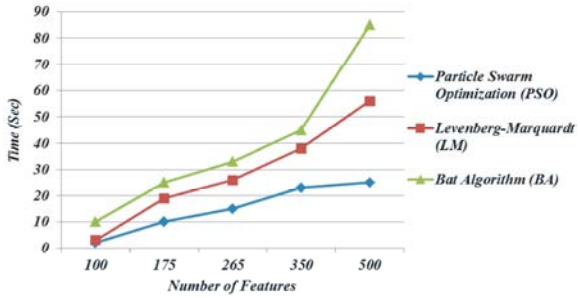

Fig. 3: Feature selection algorithm comparisons

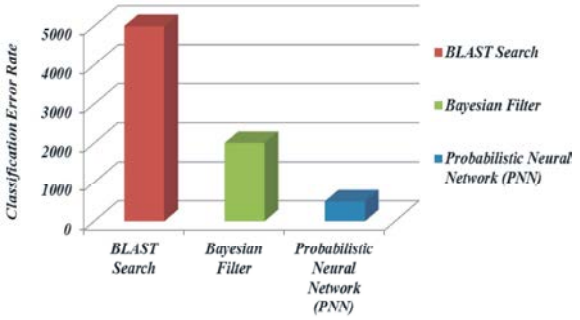

Fig. 4: Classification algorithm comparisons in terms of error rate

Table 1: Unclassified E-mails before processing

Total E-mails	Spam E-mails	Non-Spam E-mails
90	35	55

Table 2: Classified E-mails using proposed system

Classification Status	Number of E-mails
E-mails are spam E-mails and identified as spam (TP)	32
E-mails are spam E-mails but identified as non-spam (FN)	3
E-mails are non-spam E-mails and identified as non-spam (TN)	49
E-mails are non-spam E-mails and identified as spam (FP)	6

CONCLUSION

The spam mail classification is proposed using the new combination of techniques. Features are extracted using the initial Tokenization, Stemming and Stop Word Removal. This feature extraction stage is in turn act as a data preprocessing stage. The extracted features and the training data are noised with the redundant data which will be removed by using the Vector Quantization (VQ) process. The redundancy removed features may huge in size and this will be reduced using the Particle Swarm Optimization (PSO). The features sufficient for the classification process alone selected for the classification process. Finally the classification of text mail is done using the Probabilistic Neural Network (PNN). The Particle Swarm Optimization feature selection is compared with the existing Bat Algorithm (BA) and Levenberg-Marquardt

(LM). This proves that the PSO renders better results in feature selection in a minimized time. The Probabilistic Neural Network (PNN) is compared with the existing BLAST and Bayesian Filter. The comparison shows that the PNN classifies the data with the less error rate. At last the overall classification status is shown.

REFERENCES

1. Sathiya, V., M. Divakar and T.S. Sumi, 2011. Partial Image Spam E-Mail Detection Using OCR, International Journal of Engineering Trends and Technology, ISSN:2231-5381, 1(1): 55-59.
2. Tran Tich Phuoc, Pohsiang Tsai and Tony Jan, 2008. An Adjustable Combination of Linear Regression and Modified Probabilistic Neural Network for Anti-Spam Filtering, International Conference on Pattern Recognition, IEEE, ISSN: 1051-4651,1-4.
3. Roy, S., A. Patra, S. Sau, K. Mandal and S. Kunar, 2013. An Efficient Spam Filtering Techniques for Email Account, American Journal of Engineering Research (AJER), ISSN: 2320-0936, 2(10): 63-73.
4. Alfonso Ibanezn, Concha Bielza and Pedro Larranaga, 2014. Cost-Sensitive Selective Naive Bayes Classifiers for Predicting the Increase of the H-Index for Scientific Journals, Neuro computing, Elsevier, 135(5): 42-52.
5. Sagar Imambi, S. and T. Sudha, 2011. A Novel Feature Selection Method for Classification of Medical Documents from Pubmed, International Journal of Computer Applications, ISSN: 0975 – 888, 26(9): 29-33.
6. Manjot Kaur Gill, Reetkamal Kaur and Jagdev Kaur, 2010. Vector Quantization based Speaker Identification, International Journal of Computer Applications, ISSN: 0975-8887, 4(2).
7. Bilal M. Zahran and Ghassan Kanaan, 2009. Text Feature Selection using Particle Swarm Optimization Algorithm, World Applied Sciences Journal 7-Special Issue of Computer & IT, ISSN 1818-4952, 69-74.
8. Ciarelli Patrick Marques, Elias Oliveira, Claudine Badue and Alberto Ferreira De Souza, 2009. Multi-Label Text Categorization Using a Probabilistic Neural Network, International Journal of Computer Information Systems and Industrial Management Applications,ISSN: 2150-7988, 1: 133-144.
9. Tala Tafazzoli and Seyed Hadi Sadjadi, 2009.A Combined Method for Detecting Spam Machines on A Target Network, International Journal of Computer Networks and Communications (IJCNC), 1(2).
10. Lalitha, P. and Sumalatha Udutha, 2013. New Filtering Approaches for Phishing Email, International Journal of Computer Trends and Technology (IJCTT), ISSN: 2231-2803, 4(6): 1733-1736.
11. Yadav Poonam, 2014. SR-K-Means Clustering Algorithm for Semantic Information, International Journal of Inventions in Computer Science and Engineering, ISSN: 2348-3431, 1(9).
12. Kumar R. Kishore, G. Poonkuzhali and P. Sudhakar, 2012. Comparative Study on Email Spam Classifier using Data Mining Techniques, Proceedings of the International Multi Conference of Engineers and Computer Scientists, IISN: 2078-0958, 1.
13. Roy Sanjiban Sekhar, Saptarshi Charaborty, Swapnil Sourav and Ajith Abraham, 2013. Rough Set Theory Approach for Filtering Spams from boundary messages in a Chat System, 13th International Conference on Intelligent System Design and Application (ISDA), IEEE, ISBN: 4799-3515, 28-34.
14. Pratheeba R. and R. Purushothaman, 2014. Modeling Smarty Web Search Engine Using Xml Clustering, ISSN: 2348-3431, 1(2).
15. Basavaraju, M. and Dr. R. Prabhakar, 2010. A Novel Method of Spam Mail Detection using Text Based Clustering Approach, International Journal of Computer Applications, Foundation of Computer Science, 5(4): 15-25.
16. Chih-Chin Lai and Chih-Hung Wu, 2007. Particle Swarm Optimization-Aided Feature Selection for Spam Email Classification, Second International Conference on Innovative Computing Information and Control, ICICIC 07, IEEE publication, ISBN: 7695-2882.
17. Reza Narimani and Ahmad Narimani, 2013. Classification credit dataset using particle swarm optimization and probabilistic neural network models based on the dynamic decay learning algorithm, Automation, Control and Intelligent Systems, 1(5): 103-112.