International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 03 – Issue 02 || February 2018 || PP. 29-33

A Review on Spam Email Detection Using Fuzzy C-Mean with Machine Learning

[1]Gauri Katyarmal, [2]Dr.G.R.Bamnote
[1]ME student Dept. of CSE, PRMIT & R Badnera, Amravati, India
[2]Professor, Dept. of CSE, PRMIT & R Badnera , Amravati, India

Abstract: Today, most Internet users use email to communicate electronically. They depend on the Internet to deliver their important emails safely and to the right recipients. However, the fast growth of Internet users and their use of email together with the exponential increase of unsolicited users sending spam have made the email system less reliable. An email can falsely be marked by a spam filter on its way to the recipient or even get buried among junk mail in the recipient's inbox. There are several intelligent anti-spam filters which use different methods to detect spam including and fuzzy logic systems. Fuzzy set is an effective technique for spam detection and email classification. The proposed system enables the user to have more control over the various categories of spam and allows for filter personalization. This proposed work applied fuzzy logic to classify spam. This work used a fuzzy inference system to classify spam mails. This work has a list of spam words and spammer's email addresses in the database. This method extracts features from the email which, this work compares them against a list of spam features stored in the database ranked with its values and categorize the words and addresses in accordance to the ranking. Fuzzy inference system finally classified the spam mails as least dangerous or moderate or most dangerous spam mail.
Keywords: Email, spam, fuzzy logic, filter.

1. Introduction

Nowadays email is becoming fastest and economical mode of communication. All the growing use of email has led to increased rate of spam emails. As it is information all the users rely on emails to communicate with the whole world. Business organization, individuals and all corporate industries are communicating with emails so that it is important part concerning with education, business and personal usage. Spam are nothing but the unsolicited bulk emails (UBE) and it's another part is unsolicited commercial email .These spam emails not only consume the user's time but also the energy to recognize the undesired messages. It is wasting the network bandwidth. Content Based Spam Filter: Content Based filter works on content of emails i.e., text, URLs, main headers like subject for classification purpose. It is the method used to filter spam. The emails include two parts such as Body of the message and Header. Header stores the information about message like from whom it is received, date and time of emails received, sender etc. Now emails ambiguous data is removed by preprocessing then text is extracted.

Spam mail causes e-mail server engines to overload in band-width and server storage capacity. In addition, another type of spam mail i.e. phishing mail are becoming a serious threat for the security of end users, since they try to convince users and gain their personal information and account credentials to commit fraud [3]. Currently, many mail sever engines such as G-mail, Yahoo etc. are using multiple authentication techniques, analyzing the content of email, maintaining black list and white list for text categorization [1]. For dealing with image spam and attachments of word document and pdf, they are using OCR tool for extracting text embedded into natural images to obfuscate spam filters and make further classification of mail as a spam or ham mails [15]. OCR is widely acceptable and robust because of its vivacious capabilities since, it can take input of almost every image format and produces output in the desired format. Also, it has shown approximately 99.8% (ABBYY Fine Reader tool) accuracy in fetching the correct text which is embedded on image.

2. Related Work

Many studies conducted to detect email spam. Mostly used machine learning classifier as main algorithm and combine it with parameter optimization attribute selection and threshold scheme that commonly used to improve the detection result. To date, the detection accuracy is still challenging; since the best accuracy that fully accepted from all evaluation is still not reached.

BijuIssac and Wedy J Jap have proposed cost-sensitive three-way (Bayesian, thresholds, probability) to filtering spam emails which can reduce the error rate of miss classifying a legitimate email for spam and show better performance for aspects of cost-sensitive. They split the dataset into 2 parts, training part and testing part with composition 80% and 20% respectively. The dataset used is composed of three different datasets, spam

base from the UCI Machine Learning Repository, PU1 corpus and Ling-Spam corpus with accuracy 89.8 and 93.94% respectively.

Combination of Negative Selection Algorithm (NSA) with the differential evolution (DE) has been proposed by George Giannakopoulous. DE implemented the random generation phase detector distance NSA and the results can be maximized and overlapping of the detector can be minimized. DE implemented to improve the generation of detectors at the stage of the NSA, while local outlier factor (LOF) is used as a fitness function. The proposed method using spam base dataset from the UCI machine learning repository and the results yield accuracy about 83.06%.[6]

R Kishore Kumar, G Poonkuzhali and P Sudhakar propose binary search strategy subset by particle swarm optimization with mutation operator (MBPSO) that begins on a wrapper-based feature selection to extract features with the basic decision tree classifier (C4.5) and weighting parameters. The proposed method used different data value or instances with spambase but using the same spam base dataset standard format. They collect 6000 email during year 2012 and accuracy gained by 94.27%.[9]

A novel model proposed by YiShan Gong and Qiang Chen that improves the random generation of a detector in NSA with the use of stochastic distribution to model the data point using particle swarm optimization (PSO) was implemented. Local outlier factor is introduced as the fitness function to determine the local best (Pbest) of the candidate detector that gives the optimum solution. Spambase dataset from the UCI machine learning repository is used as dataset. The proposed method, NSA-PSO produces a level of accuracy of 91.22%.[7]

Another PSO was implemented to improve the random detector generation in the NSA proposed by BijuIssac. The algorithm generates detectors in the random detector generation phase of the negative selection algorithm. The combination of NSA–PSO uses a local outlier factor (LOF) as the fitness function for the detector generation. A distance measure and a threshold value are employed to raise up the distinctiveness between the non-spam and spam detectors after the detector generation. Spam base dataset is used and the proposed method yield accuracy about 83.20%.Classification techniques have been applied in textual and image spam filtering process. During literature survey, we can see a proposed method, where variant of Naïve Bayes classifier have been applied for spam detection [7]. A Compared classification strategy including Naive Bayes, Neural Network, Decision Tree and SVM were tested on different dataset on emails [9]. 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 [2]. A suggestion to combine cluster analysis based on sparse representation with clustering algorithm also provide spam detection.

Today's internet is suffering from major problem known as Email spam .It annoys users and make financial damage to companies. So far developed techniques to stop spam are filtering methods .Spam emails are UBE also known as junk emails, that are send to many recipients who have not requested or subscribe to this. Spam filter removes spam or un-required messages from email inbox. It also has Phishing URLs which redirects users to phishing websites and seeking personal credentials like username and password for financial purpose. The existing work by Rasim M Alguliev, Ramiz M Aliguliyev , did implementation on malicious URL detection in Email. Lexical features, page rank, host information are taken into consideration to classify URLs. Phishtank corpora has been used and Bayesian classification is done to improve the performance of system [1].

T.A.Almeida and A. Yamakami, have presented learning method to filter spam email. The two machine learning algorithm are considered for anti-spam filtering such as Naïve Bayesian and Memory based learning approach and they are compared concerning performance. So, that in both methods spam filtering accuracy has improved and keyword based filter are used widely for email [2].

Veena H Bhat, Vandana R Malkani has given an application for email filtering using a new improved Bayesian filter. They have represented word frequency by vector weights and word entropy is used for attribute selection then formula is derived which improves the performance apparently [3].

Jiansheng Wu and Tao Deng, have shown that phishing websites are hacked as soon as they are identified as phishing campaigns have two hours of average life. So to block and identify such phishing URLs they have extracted features like suspicious characters, number of dots, ip address, hexa decimal character [13]. SeongwookYoun and Dennis McLeod, discovered malicious URLs by enhancing blacklisting. One conflict with this method is that their updating process is fast so they failed to identify phishing URLs in early hours of a phishing attack[14]. YiShan Gong and Qiang Chen, endeavor for a survey to recognize the essential features which can develop accuracy and precision for malicious URLs detection [7]. Fumera, Giorgia, Ignazio Pillai, did a feature extraction on Base64 encoding of image with n-gram technique. A SVM needs to be trained for efficiently detecting spam images from legitimate images. Its seen from experiment that it has improved the performance in terms of Accuracy, Precision and Recall [15].

R. Parimal, has given a new spam detection method by employing Text Categorization, using Supervised Learning with Bayesian Neural Network which uses Rule based heuristic approach and statistical

analysis tests to identify "Spam" [10]. R Kishor Kumar, P Sudhakar, had presented spam detection based on interval type-2 fuzzy sets. This system gives user more control on categories of spam and permits the personalization of the spam filter [10].

3. Proposed Work

In the proposed methodology to detect an email as spam or legitimate mail on the basis of text categorization. Various techniques for pre-processing of email format are applied such as applying stop words removing, stemming, feature reduction and feature selection techniques to fetch the keywords from all the attributes and finally using different classifiers to segregate mail as spam or ham. We have used numeric feature representation technique for feature extraction and fuzzy c-mean algorithm for clustering feature vectors, in various clusters and machine learning technique to make predictive analysis of mail filtering, this will be decision making step.

3.1 Pre-processing

Frequently the texts we have are not those we want to analyse. Having a single file containing the collected works of an author although we are only interested in a single work. Or we may be given a large work broken up into volumes where the division into volumes is not important to us. If we want to break up a long text (such as a book-length work) into smaller chunks so we can get a sense of the variability in an author's writing. If we are comparing one group of writers to a second group, we may wish to aggregate information about writers belonging to the same group. This will require merging documents or other information that were separated initially. This section illustrates these two common pre-processing step: splitting long texts into smaller "chunks" and aggregating texts together. Another important pre-processing step is tokenization.

3.2 Tokenizing

Tokenization is a critical activity in any information retrieval model, which simply segregates all the words, numbers, and their characters etc. from given document and these identified words, numbers, and other characters are called tokens. Along with token generation this process also evaluates the frequency value of all these tokens present in the input documents.

3.3 Stemming

Stemming phase is used to extract the sub-part i.e. called as stem/root of a given word. For example, the words continue, continuously, continued all can be rooted to the word continue. The main role of stemming is to remove various suffixes as result in the reduction of number of words, to have exactly matching stems, to minimize storage requirement and maximize the efficiency of model.

3.4 Chunking

Chunking is a term referring to the process of taking individual pieces of information and grouping them into larger units. By grouping each piece into a larger whole, you can improve the amount of information you can remember. By separating disparate individual elements into larger blocks, information becomes easier to retain and recall.

3.5 Stopping

Removal of stop words – Stop words like "and", "the", "of" are very common in all English sentences and are not very meaningful in deciding spam or legitimate status, so these words have been taken away from the emails. For example stop words include "the, as, of, and, or, to etc. this phase is very essential in the tokenization because it has some advantages: It reduces the size of indexing file and it also improve the overall efficiency and make effectiveness.

3.6 Feature extraction process

Once the dictionary is ready, we can separate word count vector (our feature here) of 3000 dimensions for each email of preparing set. Each **word count vector** contains the frequency of 3000 words in the training file. Of course you might have guessed by now that most of them will be zero. Let us take an example. Suppose we have 600 words in our dictionary. Each word count vector contains the frequency of 600 dictionary words in the training file. Suppose text in training file was "Get the work done, work done" then it will be encoded as[0,0,0,0,0,…….0,0,2,0,0,0,……,0,0,1,0,0,…0,0,1,0,0,……2,0,0,0,0,0]. Here, all the word counts are placed at 296th, 359th, 415th, 495th index of 600 length word count vector and the remaining are zero. The python code will generate a feature vector matrix whose rows denote 800 files of training set and columns denote 4000 words of dictionary. The incentive at record "ij" will be the quantity of events of jth expression of word

reference in ith document.

3.7 Fuzzy c-means clustering

Fuzzy logic principles can be utilize to cluster multidimensional data, assigning each point a participationin each cluster centre from 0 to 100 percent. This can be very powerful compared to traditional hard-threshold clustering where every point is assigned a crisp, exact label. Fuzzy c-means clustering is accomplished via sk fuzzy, c means, and the output from this function can be repurposed to classify new data according to the calculated clusters (also known as prediction) via sk fuzzy c means predict Fuzzy clustering (also referred to as soft clustering) is a form of clustering in which each data point can belong to more than one cluster. Cluster analysis or clustering involves assigning data points to clusters such that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Different similarity measures may be chosen based on the data or the application. In non-fuzzy clustering (also known as hard clustering), data is divided into different clusters, where each data point can only belong to exactly one cluster. In fuzzy clustering, data points can possibly belong to multiple clusters.

3.8 Training the classifiers

In the proposed system Support Vector Machines (SVM).SVMs are supervised binary classifiers which are extremely viable when you have higher number of features. The objective of SVM is to separate some subset of preparing data from rest called the support vectors. The decision function of SVM model that predicts the class of the test data is based on support vectors and makes use of a kernel trick. Once the classifiers are trained, we can check the performance of the models on test-set. We separate word count vector for each mail in test-set and anticipate its class with the prepared SVM model.

Conclusion

In this paper we have proposed procedure to identify an email as spam or ham based on text categorization. Different methods for pre-processing of email organize are connected, for example, applying stop words expelling, stemming, include decrease and highlight choice strategies to bring the catchphrases from every one of the qualities lastly utilizing distinctive classifiers to isolate mail as spam or ham. We have utilized numeric feature representation for highlight extraction and fuzzy c-mean calculation for grouping highlight vectors**.**

References

[1]. Rasim M Alguliev, Ramiz M Aliguliyev, and Saadat A Nazirova. Clas-sification 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, pages1−7, 2010.

[3]. Veena H Bhat, Vandana R Malkani, PD Shenoy, KR Venugopal, andLMPatnaik. Classification of email using beaks: Behavior and keywordstemming. In TENCON 2011-2011 IEEE Region 10 Conference, pages1139−1143. 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]. Dukeeducation.Stemmingcode.URLhttp://www.cs.duke.edu/courses/compsci308/cps108/fall07/code/stemmer/code.pdf.

[6]. George Giannakopoulos, Petra Mavridi, GeorgiosPaliouras, GeorgePapadakis, and KonstantinosTserpes. 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, pagesV4−678−V4−680, 2010.

[8]. BijuIssac 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.

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

[10]. 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.

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

[12]. 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.

[13]. Jiansheng Wu and Tao Deng. Research in anti−spam method based on bayesian filtering. In Computational Intelligence and Industrial Appli-cation, 2008. PACIIA '08. Pacific-Asia Workshop on, volume 2, pages887−891, 2008.

[14]. SeongwookYoun and Dennis McLeod. A comparative study for email classification. In Advances and Innovations in Systems, Computing Sci-ences and Software Engineering, pages 387−391. Springer, 2007.

[15]. Fumera, Giorgio, Ignazio Pillai and Fabio Roli. Spam filtering based on the analysis of text information embedded into images. The Journal of Machine Learning Research 7 (2006): 2699−2720.