Scientific Journal of Impact Factor (SJIF): 4.72

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 4, Issue 11, November -2017

An Improved Mechanism using J48 Classifiers for Spam Filtering

[1]Shiva Sharma, [2]U.Datta

[1]*Dept. of CSE/IT, Maharana Pratap College of Technology, Gwalior, India*
[2]*Dean Academic, Maharana Pratap College of Technology, Gwalior, India*

Abstract— *The internet has become an inherent part of daily day to day life and email has become a robust tool for information exchange. There has been a prominent growth in spam in recent years as the importance and applications of the web and e-mail has grown. Spam has been a serious and annoying problem for decades. Even though plenty of solutions have been put forward, there still remains a lot to be promoted in filtering spam emails more efficiently. Nowadays a major problem in spam filtering as well as text classification in natural language processing is the huge size of vector space due to the numerous feature terms, which is usually the cause of extensive calculation and slow classification. Support vector machine (SVM) takes a set of input data and output the prediction that data lays in one of the two categories i.e. it classify the data into two possible classes.*

Keywords—*Social Network, Email, Spam, Social Spam, Support Vector Machine, SMO.*

I. INTRODUCTION

While online social networks are now recognized as popular communication channels, survey results reveal that email still reigns as the most popular form of Internet communications in our daily life. Email is a cost-effective method of communication commonly found in all areas of industries. Education industry is not an exception. Workforce in education industry spends fair amount of time in front of computer chasing up on emails. This is more so with jobs that deal with high volume of emails each day such as administrator in education industry. Managing incoming email is a critical matter to many because emails can herald important meetings, work messages, lunch, industry related information, upcoming events which many cannot afford to miss [1]. Millions of people use email correspondence for communication across the globe and it is a critically vital application for many businesses. One of the long-last problem in email systems is the existence of spam emails. Spam emails, also known as junk emails, are unsolicited bulk emails received by massive recipients. The sender's addresses of spam emails are usually concealed and the spams are not requested by any of the recipients. Spam emails may contain malware, in the form of scripts or executable files, or contain disguised links that lead to phishing web sites. Considerable amount of unsolicited mail flows into user's mail boxes on a daily basis. A major negative aspect since the past decade has been bulk spam or phishing mail. Besides such unsolicited spam emails being wearisome for many email users, it also puts pressure on the IT infrastructure of organizations and costs businesses billions of dollars in lost efficiency. Increasing need of effectively filtering spam has become vital [2].

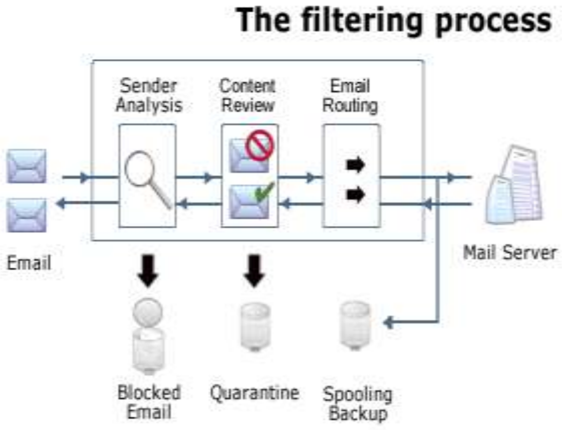

Fig.1 Spam Filtering Process

Social spam is an e-crime on social networking sites with contents such as comments, post, chat, etc. There are many spamming activities going through social media such as malicious links posting, insulting posts, hate speech, fake friends**,** deceitful reviews, etc. Motivation of these spamming activities can be either private or commercial. Previously, emails were the major object of spammers however it is slowly reduced with the advancement of spam filters that can filter almost 95% of spam content mails. On the other hand, growth of social networking sites and its weak security measures attracted many spammers and made it a vigorous field of concern for research community. In past decade, lot of research activity has been performed to control spamming activities using supervised and unsupervised mechanisms. But typical enhancement of technologies made spamming actions difficult to tackle with existing mechanisms. This made the researchers to come up with the advanced mechanism improving the existing one through further modification. Majority of the social media rely on the user community to tackle with spam issues because of its dynamicity. In past few years, enormous supervised approaches had been proposed for spam filtering but immutable nature of these approaches needs the model to be retrained every time to classify a new variety of spam which is inefficient [3].

There are various definitions for spam and its difference from valid mails. The shortest definition of spam is 'an unwanted electronic mail'. A major problem with introduction of spam filtering is that a valid email may be labeled spam or a valid email may be missed. To not filter spams causes problems; not only will inboxes be completely occupied by spam, but it will result in more serious problems including reduction of bandwidth and storage. There are techniques to identify emails received in the form of spam, as follows: black list/white list, Bayesian classifying algorithm, and keyword matching and header information analysis. A white list is a list of addresses from which users tend to receive emails. Users can also add email addresses, domain inputs or domains of functions. An advantage of white list is that it allows users or administrators to put email addresses of favorite people into the list in order to make sure that valid emails received from addresses in the white list are not labeled spam when receiving emails from different senders. A black list is a list of addresses from which users do not tend to receive emails. The header reviewing process of an email involves a series of rules implemented as follows. An email will be labeled junk and transferred to a spam folder if its header is congruent to a header of training data in the black list. Otherwise, it will be transferred to the white list [4].
Following are some of the focuses in the research of email analysis data mining in emails' datasets:

1. A major research subject in email classification is to classify emails into spam or no spam emails. This can be further used for the real time prediction of spam emails.
2. Some research papers tried to classify emails based on similar threads or subjects. Some email systems such as Gmail connect emails related to each other (e.g. by reply or forward events) together.
3. Some email classification research papers tried to classify emails based on the gender of the sender given some of the common aspects that may distinguish emails from females or males.
4. Email classification can be also used to automatically assign emails to predefined folders.
5. Rather than spam and non spam emails, emails can be also classified into: Interesting and uninteresting emails.
6. Features are extracted from the email content or body, title or subject or some of the other Meta data that can be extracted from the emails such as: sender, receiver, BCC, date of sending, receiving, number of receivers, etc. The method to extract feature can be based on words, bags of words, etc.
7. Email clustering is also considered to cluster emails into different subjects or folders.
8. The time information in emails (e.g. when: sent, received, etc.) is used also in some research papers to classify emails [5].

II. Literature Survey

S. Youn et al. [6] explored how the classification performance can be affected by the size of dataset. Accuracies of 92.7%, 97.2%, and 95.8% where achieved for a dataset of size 1000 by SVM, NB, and J48 respectively. However, accuracy of J48 increased by 1.8%, whereas that of SVM and NB dropped by1.8% and 0.7% respectively on increasing the dataset size to 5000. Furthermore, they found out that by increasing feature size, accuracy too increased.

F. Benevenuto et al. [7] provided a heuristic for classifying an arbitrary video as legitimate or spam. In the proposed method, the dataset of Youtube users is collected and manually classified to be spam or legitimate. Support Vector Machine (SVM) is used as a data classifier with a 5- fold cross validation. The mechanism achieved a true positive rate of 43.9% and the accuracy of 87% for detecting spam.

Yudong Zhang et al. [8] proposed a spam filtering method namely binary PSO with mutation operator (MBPSO) for feature selection. In this model, decision tree is chosen as a classifier model with the training algorithm C4.5. The achieved specificity, sensitivity and accuracy are 97.51%, 91.02%, and 94.27% respectively.

Biro I et al. [9] two ways are described for classification. First is done with some rules which can be defined manually, like rule headquartered trained method. This process of classification is utilized when lessons are static, and their addons are simply separated in accordance with the features. Second is completed with the support of present computing device learning methods. According to the trained [10] clusters of spam emails are created with the help of criterion perform. Criterion function is outlined because the maximization of similarity between messages in clusters and this similarity is calculated utilizing ok-nearest neighbor algorithm.

Kufandirimbwa et al. [11] spam is detected using artificial neural network. In this paper author designed the artificial neural network spam detector using the perceptron learning rule. Perceptron employs a stochastic gradient method for training, where the true gradient is evaluated on a single training example and the weights are adjusted accordingly until a stopping criterion is met.

Mohammad et al. [12] fuzzy clustering procedure is used. On this paper author evaluated the usage of fuzzy clustering and textual content mining for spam filtering. Fuzzy clustering is scalable and effortless to update process. This is trained offers with the examination of use of fuzzy clustering algorithm to construct a spam mail filter. Classifier has been proven on one-of-a-kind data units and after testing Fuzzy C-approach making use of Heterogeneous value difference Metric with variable percentages of spam mail and used a regular model of assessment for the hindrance of spam mail classification.

Blanzieri and Bryl et al. [13] presented a technical report in 2008 to survey learning algorithms for spam filtering. The paper discussed several aspects related to spam filtering such as the proposals to change or modify email transmission protocols to include techniques to eliminate or reduce spams. Some methods focused only on content while others combined header or subject with content. Some other email characters such as size, attachments, to, from, etc. were also considered in some cases. Feature extraction methods were also used for both email content, attached and embedded images.

D. Punuskis et al. [14] artificial neural network (ANN) was applied to filter unsolicited emails. Replacing the frequency of words in the content with explanatory properties of obscure patterns fashioned by the spammers was their major contribution. Their data set consisted of about 1800 spam and 2800 valid emails. A maximum precision value of 91% after training it using 57 email parameters was obtained by ANN.

Joachims et al [15] in this thesis, had proposed the use of Support Vector Machines (SVMs) for text classification and had also demonstrated that SVMs could offer better performance for text classifiers as compared to other well known machine learning techniques such as Naïve-Bayes classifiers and kNN classifiers.

Mountrakis et al [16] in this thesis, presented a review over remote sensing implementations of support vector machines. This review is timely due to the exponentially increasing number of works published in recent years. Most of the findings show that there is empirical evidence to support the theoretical formulation and motivation behind SVMs. The most important characteristics is SVM's ability to generalize well from a limited amount and quality of training data. Compared to alternative methods such as BPN, SVM's can yield comparable accuracy using a much smaller training sample size.

M. Kepa et al. [17] in this thesis, authors presented a hybrid classification model that uses k-nearest neighbor and support vector machine techniques. This method is two stage approach based on the one-vs-near scheme was tested on big datasets. In the first stage, the kNN classifier is used to compute the category neighbor list which is learning phase. The kNN figures the distance between every centroid in the form of an ordered list which is used in second stage classifier. The second stage SVM uses the saved neighbor list to limit the dataset used for training the classifier for a single category.

Choy Y.K. et al. [18] in this thesis, in its several proposed approach there is an approach to select the structure of the RBF networks based on the support vectors (SVs) of the support vector machines. In this paper, the modeling of the relationship between rainfall and river discharges of the Fuji River using the SVRBFN is presented. The main advantage of this approach is that the structure of the network can be obtained objectively, as the SVs of the SVM are obtained from the constrained optimization for a given error bound.

Lai H.C. et al. [19] In this paper four simple dynamic methods and two supervised learning techniques including a linear regression model, a quadratic regression model, an original grey prediction model, a back-propagation neural network model, and an epsilon-SVM regression model were investigated for the forecasting of flood stage one hour ahead for early warning of flooding hazards.

III. BACKGROUND

A. SUPPORT VECTOR MACHINE

Support vector machines are supervised learning models with associated learning models that analyze data and are mainly used for classification purpose. Support vector machine (SVM) takes a set of input data and output the prediction that data lays in one of the two categories i.e. it classify the data into two possible classes. Given a set of training examples, each marked as belonging to one of the two classes, an SVM training algorithm build a model that assign new data in one class or the other. Basically SVM is a representation of the examples as points in space, mapped so that new examples of the separate classes are clearly classified as belonging to one of the two categories. A support vector machine performs classification by constructing an N-dimensional hyper plane that optimally categorizes the data in two categories. SVM are set of related supervised learning methods used for classification and regression. SVM map input vector to a higher dimensional plane where a maximal separating hyper plane is constructed. Two parallel hyper planes are constructed on each side of the hyper plane that separates the data. The separating hyper plane is the hyper plane that maximizes the distance between the two hyper planes. Larger the margin or distance better the generalization error of the classifier [20].

B. SEQUENTIAL MINIMAL OPTIMIZATION

Sequential Minimal Optimization (SMO) is a simple algorithm that can quickly solve the SVM QP problem without any extra matrix storage and without using numerical QP optimization steps at all. SMO decomposes the overall QP problem into QP sub-problems, to ensure convergence. The noteworthy features of the SMO algorithm are computational speed and ease of implementation. In addition, SMO requires no extra matrix storage at all. Thus, very large SVM training problems can fit inside of the memory of an ordinary personal computer or workstation. Because no matrix algorithms are used in SMO, it is less susceptible to numerical precision problems.

Algorithm: Simplified SMO
Input: C: regularization parameter
tol: numerical tolerance
max passes: max # of times to iterate over α's without changing
(x (1), y(1)), . . . ,(x (m) , y(m)): training data [21]

IV. PROPOSED WORK

J48 Decision Tree

Classification is the process of building a model of classes from a set of records that contain class labels. Decision Tree Algorithm is to find out the way the attributes-vector behaves for a number of instances. Also on the bases of the training instances the classes for the newly generated instances are being found. This algorithm generates the rules for the prediction of the target variable. With the help of tree classification algorithm the critical distribution of the data is easily understandable. J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible.

This algorithm it generates the rules from which particular identity of that data is generated. The objective is progressively generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Basic Steps in the Algorithm:

(i) In case the instances belong to the same class the tree represents a leaf so the leaf is returned by labeling with the same class.

(ii) The potential information is calculated for every attribute, given by a test on the attribute. Then the gain in information is calculated that would result from a test on the attribute.

(iii) Then the best attribute is found on the basis of the present selection criterion and that attribute selected for branching.

The system's input is a set of training samples/emails that include both spam and non-spam emails. The contents of every email in the training set will be first segmented. Then, the feature vector of every email will be extracted by the feature selection module. Because most of the features present redundancy and inconsistency, we adopt a feature selection method that is based on the information gain (IG). Specifically, we compute the IG for every feature vector, no matter whether it corresponds to a spam or a regular email. These feature vectors are then ordered based on their IG values, in a decreasing order.

Proposed Algorithm:

```
Step:1    Input datasets from the database
Step:2    Perform Pre-processing over datasets (Training and Testing)
Step:3    Execute tokenization, stop word removal and stemming
Step:4    Perform feature selection over the subset of the overall data
Step:5    Testing Phase
              a.   Input testing dataset
              b.   Testing instances contains set of unlabelled incoming spam and legitimate mails
              c.   Classify testing instances
Step:6    Apply J48 Algorithm
Step:7    Create root node and label with splitting attribute
Step:8    D = Database created by applying splitting predicate to D
Step:9    If stopping point reached for this path
Step:10   Stop
```

V. Result Analysis

In the implementation of the proposed work, we used WEKA which show HAM and SPAM. Enron datasets are used for the detailed analysis of the mails. WEKA (Waikato Environment for Knowledge analysis) is a fashionable suite of machine learning software written in Java, developed on the University of Waikato, New Zealand. WEKA is free software to be had beneath the GNU General Public License. In the first approach, conventional batch training is performed on J48. All training examples are presented at the same time and resultant support vectors are used to discriminate e-mails from testing sets. J48 is trained incrementally in the second and third approaches.

For filter evaluation, we use six large and well-known ENRON data-sets. This data-set contains pre-processed email messages with removal of attachments. The data-set belongs to six e-mail directories farmer-d, kaminski-v, kitchenl, williams-w3, beck-s, and lokay-m, named as ENRON1 to ENRON6. The composition of each directory is given in table I along with training and testing sets sizes we consider for our experiment.

Table I: Dataset Size

Datasets	Total Mails		Training Set Size		Testing Set Size	
	Spam	Ham	Spam	Ham	Spam	Ham
Enron 1	1500	3672	500	1224	100	250
Enron 2	1496	4361	400	1100	110	330
Enron 3	1500	4012	500	1300	100	270
Enron 4	4500	1500	1000	500	350	100
Enron 5	3675	1500	1225	500	245	100
Enron 6	4500	1500	1000	500	350	100

Spam Recall (SRe) is the percentage of all spam emails that are correctly classified as spam.
SRe=No. of Spam Correctly Classified/ Total No. of messages

Spam Precision (SPr) is the number of relevant documents identified as a
SPr=No. of Spam Correctly Classified / Total No. of messages classifies as spam

Table II: Performance Measures for ENRON Datasets

Datasets	**Performance Measures**	**Results achieved in each Training mode**		
		Conventional Batch Training	**Incremental Training with same features**	**Incremental Training with Updated Features**
ENRON 1	SPr	0.861	0.835	0.771
	SRe	0.930	0.910	0.871
	LPr	0.971	0.963	0.944
	LRe	0.940	0.928	0.894
	MCC	0.851	0.818	0.740
ENRON 2	SPr	0.730	0.880	0.770
	SRe	0.736	0.800	0.763
	LPr	0.912	0.935	0.920
	LRe	0.909	0.964	0.923
	MCC	0.644	0.789	0.688
ENRON 3	SPr	0.755	0.728	0.707
	SRe	0.830	0.830	0.810
	LPr	0.935	0.934	0.926
	LRe	0.900	0.885	0.877
	MCC	0.709	0.688	0.660
ENRON 4	SPr	0.929	0.940	0.997
	SRe	0.937	0.943	0.871
	LPr	0.773	0.798	0.925
	LRe	0.750	0.790	0.989
	MCC	0.695	0.735	0.848
ENRON 5	SPr	0.948	0.924	0.942
	SRe	0.967	0.996	0.979
	LPr	0.916	0.988	0.943
	LRe	0.870	0.800	0.853
	MCC	0.850	0.852	0.858
ENRON 6	SPr	0.947	0.953	0.937
	SRe	0.890	0.977	0.919
	LPr	0.969	0.912	0.735
	LRe	0.986	0.830	0.783
	MCC	0.896	0.835	0.686

A. Confusion matrix:

A confusion matrix illustrates the accuracy of the solution to a classification problem. Given n classes a confusion matrix is a m x n matrix, where Ci,j indicates the number of tuples from D that were assign to class Ci,j but where the correct

class is Ci . Obviously the best solution will have only zero values outside the diagonal. A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier. The entries in the confusion matrix have the following meaning in the context of our study:
1. a is the number of correct predictions that an instance is negative,
2. b is the number of incorrect predictions that an instance is positive,
3. c is the number of incorrect of predictions that an instance negative, and
4. d is the number of correct predictions that an instances positive.

Some standards and terms:
1. True positive (TP): If the outcome from a prediction is p and the actual value is also p, then it is called a true positive.
2. False positive (FP): However if the actual value is n then it is said to be a false positive.
3. Precision and recall: Precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

Both precision and recall are therefore based on an understanding and measure of relevance. Precision can be seen as a measure of exactness or quality, whereas recall is a measure of completeness or quantity. Recall is nothing but the true positive rate for the class.

Conclusion

The most common scam mails is the fraud job offer mails, most of them are using the logos of multinational companies and higher official names and signatures. The only way to identify the fraud mails and legitimate mails is that the email ids of multinational companies' newer use Gmail, Hotmail or Yahoo, they will have their official mail account. The performance testing on the designed email spam filter is to calculate the accuracy, reliability and other factors.

References

[1] M. K. Chae, Abeer Alsadoon, P.W.C. Prasad, Sasikumaran Sreedharan, "Spam Filtering Email Classification (SFECM) using Gain and Graph Mining Algorithm", IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), January 2017.

[2] Simranjit Kaur Tuteja, Naga raju Bogiri, "Email Spam Filtering using BPNN Classification Algorithm", International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT), IEEE, pp. 915-919, 2016.

[3] Mohit Agrawal, R. Leela Velusamy, " R-SALSA: A Spam Filtering Technique for Social Networking Sites", IEEE Students' Conference on Electrical, Electronics and Computer Science, 2016.

[4] Ali Shafigh Aski, Navid Khalilzadeh Sourati, "Proposed efficient algorithm to filter spam using machine learning techniques", Pacific Science Review A: Natural Science and Engineering 18, (2016), 145-149.

[5] Izzat Alsmadi, Ikdam Alhami, "Clustering and classification of email contents", Journal of King Saud University – Computer and Information Sciences, Elsevier, (2015) 27, 46–57.

[6] S. Youn & D. McLeod, "*A comparative study for email classification", In* Advances and Innovations in Systems, Computing Sciences and Software Engineering, 2007.

[7] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross, "*Identifying video spammers in online social networks*," Proceedings of the 4th international workshop on Adversarial information retreival on the web, ACM, pp. 45-52, 2008.

[8] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "*Binary PSO with mutation operator for feature selection using decision tree applied to spam detection*," Knowledge-Based Systems, 64, pp. 22-31, 2014.

[9] Biro I, Szabo J, Benczur A, and Siklosi D. LinkedLatent Dirichlet Allocation in Web Spam Filtering[A].In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIR Web), Madrid, Spain, 2009.

[10] Perkins A. "The classification of search engine spam" http://www.ebrandmanagement.com/whitepapers/spam classification, 2001.

[11] Kufandirimbwa O, Gotora R. Spam detection using Artificial Neural Networks [J]. In Online Journal of Physical and Environmental Science Research, 2012, 1:22-29.

[12] Mohammad N.T.A Fuzzy clustering approach to filter spam E-mail [A]. Proceedings of World Congress on Engineering, vol. 3, WCE-2011.

[13] Enrico Blanzieri, Anton Bryl, 2008. A survey of learning-based techniques of email spam filtering, Technical Report # DIT-06-056. Pranjal S. Bogawar, Kishor K. Bhoyar, 2012. Email mining: a review, IJCSI Int. J. Comput. Sci. Issues 9(1), No 1, January 2012.

[14] D. Punuskis, R. Laurutis & R. Dirmeikis, "*An artificial neural nets for spam e–mail recognition,*" Electronics and Electrical Engineering, 2006.

[15] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Technical Report 23, Universitat Dortmund, LS VIII, 1997.

[16] Mountrakis G., Im J. and Ogole C., 2010, Support vector machines in remote sensing: A review, ISPRS Journal of Photogrammetry and Remote Sensing, doi:10.1016/j.isprsjprs.2010.11.001.

[17] M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.

[18] Choy Y.K. and Chan W.C., 2010, Modeling of river discharges and rainfall using radial basis function networks based on support vector regression, International Journal of Systems Science, vol.34, numbers14-15, pp763-773.

[19] Lai H.C. and Tseng H.M., 2010, Comparison of regression models, grey models, and supervised learning models for forecasting flood stage caused by typhoon events, Journal of the Chinese Institute of Engineers, 33:4, 629-634.

[20] Megha Rathi and Vikas Pareek, "Spam Mail Detection through Data Mining – A Comparative Performance Analysis", I.J. Modern Education and Computer Science, 2013, 12, 31-39.

[21] Platt, John. Fast Training of Support Vector Machines using Sequential Minimal Optimization, in Advances in Kernel Methods – Support Vector Learning, B. Scholkopf, C. Burges, A. Smola, eds., MIT Press (1998).