Comparative Study on Email Spam Classifier using Data Mining Techniques

R. Kishore Kumar, G. Poonkuzhali, P. Sudhakar, *Member, IAENG*

Abstract— **In this e-world, most of the transactions and business is taking place through e-mails. Nowadays, email becomes a powerful tool for communication as it saves a lot of time and cost. But, due to social networks and advertisers, most of the emails contain unwanted information called spam. Even though lot of algorithms has been developed for email spam classification, still none of the algorithms produces 100% accuracy in classifying spam emails. In this paper, spam dataset is analyzed using TANAGRA data mining tool to explore the efficient classifier for email spam classification. Initially, feature construction and feature selection is done to extract the relevant features. Then various classification algorithms are applied over this dataset and cross validation is done for each of these classifiers. Finally, best classifier for email spam is identified based on the error rate, precision and recall.**

Index Terms— **classifier, e-mail, feature construction, feature selection, relevance analysis, spam**

I. Introduction

Due to the intensive use of internet, email has become one of the fastest and most economical mode of communication. This enables internet user to easily transfer information from anywhere in the world in a fraction of second. However, the increase of email users have resulted in the dramatic increase of spam emails during the past few years. E-mail spam, also known as junk e-mail or unsolicited bulk e-mail (UBE), is a subset of spam that delivers nearly identical messages to numerous recipients by e-mail. Definitions of spam usually include the aspects that e-mail is unsolicited and sent in bulk. E-mail spam has steadily grown since the early 1990s. Botnets, networks of virus-infected computers, are used to send about 80% of spam.

Spammers collect e-mail addresses from chatrooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. Since the cost of the spam is borne mostly by the recipient, many individual and business people send bulk messages in the form of spam. The voluminous of spam emails a strain the Information Technology based organizations and creates billions of dollars lose in terms of productivity. In recent years, spam emails lands up into a serious security threat, and act as a prime medium for phishing of sensitive information [13]. Addition to this, it also spread malicious software to various user. Therefore, email classification becomes an important research area to automatically classify original emails from spam emails. Spam email also fascinate problem for individuals and organizations because it is prone to misuse. Automatic email spam classification [18] contains more challenges because of unstructured information, more number of features and large number of documents. As the usage increases all of these features may adversely affect performance in terms of quality and speed. Many recent algorithms use only relevant features for classification. Even though more number of classification techniques has been developed for spam classification, still 100% accuracy of predicting the spam email is questionable. So Identification of best spam algorithm itself became a tedious task because of features and drawbacks of every algorithm against each other.

In this paper, spam dataset from UCI machine learning repository [23] is taken as input data for analyzing the various classification techniques using TANAGRA [22] data mining tool. In this work, feature construction and feature selection is done first to select the relevant features for classification. After feature extraction, 15 different classification algorithms are taken for evaluation. In this evaluation process, different features are considered for choosing best spam filtering algorithm. Finally performance evaluation is done to analyze the various classification algorithms to select the best classifier for spam emails.

Outline of this paper:
Section 2 presents related works on email spam classification, section 3 presents framework implementation of the proposed system, section 4 presents feature relevance analysis, Section 5 presents study on classification algorithms, Section 6 gives experimental results and performance evaluation. Finally section 7 presents conclusion.

II. Related Works

Email spam is one of the major problems of the today’s Internet, bringing financial damage to companies and annoying individual users. Among the approaches developed to stop spam, filtering is the one of the most

R. Kishore Kumar is with the Department of Computer Science & Engineering, Sri Sivasubramaniya Nadar College of Engineering, Old Mahabalipuram Road, SSN Nagar -603 110, Tamil Nadu, India (e-mail: rskishorekumar@yahoo.co.in).

G. Poonkuzhali is with the Department of Computer Science & Engineering, Rajalakshmi Engineering College, Affiliated to Anna University, Chennai, India, phone: +91 9444836861; (e-mail: poonkuzhali.s@ rajalakshmi.edu.in).

P. Sudhakar is with the Department of Computer Science and Engineering , Kamaraj College of Engineering, Tamil Nadu, India (e-mail: sudhakar.asp@gmail.com).

important technique. Spam mail, also called unsolicited bulk e-mail or junk mail that is sent to a group of recipients who have not requested it. The task of spam filtering is to rule out unsolicited e-mails automatically from a user's mail stream. These unsolicited mails have already caused many problems such as filling mailboxes, engulfing important personal mail, wasting network bandwidth, consuming users time and energy to sort through it, not to mention all the other problems associated with spam [11]. Two methods of machine classification were described in paper [4]. The first one is done on some rules defined manually. The typical example is the rule based expert systems. This kind of classification can be used when all classes are static, and their components are easily separated according to some features. The second one is done using machine learning techniques. Paper [15] formalizes a problem of clustering of spam message collection through criterion function. The criterion function is a maximization of similarity between messages in clusters, which is defined by *k*-nearest neighbor algorithm. Genetic algorithm including penalty function for solving clustering problem is offered. Classification of new spam messages coming to the bases of antispam system.

A novel distributed data mining approach, called Symbiotic Data Mining (SDM) [7] that unifies Content-Based Filtering (CBF) with Collaborative Filtering (CF) is described. The goal is to reuse local filters from distinct entities in order to improve personalized filtering while maintaining privacy. In paper [26] the effectiveness of email classifiers based on the feed forward back propagation neural network and Bayesian classifiers are evaluated. Results are evaluated using accuracy and sensitivity metrics. The results show that the feed forward back propagation network algorithm classifier provides relatively high accuracy and sensitivity that makes it competitive to the best known classifiers. A fully Bayesian approach to soft clustering and classification using mixed membership models based on the assumptions on four levels: population, subject, latent variable, and sampling scheme was implemented in [8]. In paper [1]-[3], automatic anti-spam filtering becomes an important member of an emerging family of junk-filtering tools for the Internet, which will include tools to remove advertisements. The author separate distance measures for numeric and nominal variables, and are then combined into an overall distance measure. In another method, nominal variables are converted into numeric variables, and then a distance measure is calculated using all variables. Paper [24] analyzes the computational complexity and scalability of the algorithm, and tests its performance on a number of data sets from various application domains. The social networks of spammers [12] by identifying communities of harvesters with high behavioral similarity using spectral clustering. The data analyzed was collected through Project Honey Pot [14], a distributed system for monitoring harvesting and spamming. The main findings are (1) that most spammers either send only phishing emails or no phishing emails at all, (2) that most communities of spammers also send only phishing emails or no phishing emails at all, and (3) that several groups of spammers within communities exhibit coherent temporal behavior or have similar IP addresses [5]. It is demonstrated that both methods obtain significant generalizations from a small number of examples; that both methods are comparable in generalization performance on problems of this type; and that both method asset reasonably efficient, even with fairly large training sets [6]. Spam classification [21] is done through Linear Discriminant Analysis by creating a bag-of words document for every Web site.

III. Framework of the proposed system

The overall design of the proposed system is given in Fig. 1 and each of these components is addressed in the following sections briefly.

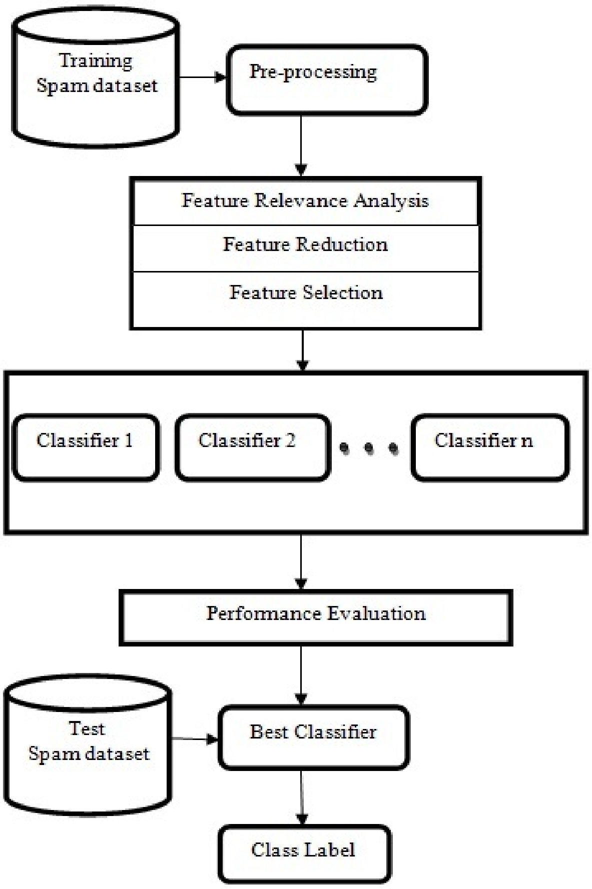

Fig 1. Architectural design of the proposed system

A. *Spam Dataset*

The spam dataset was taken from UCI machine learning repository and was created by Mark Hopkins, Erik Reeber, George Forman, Jaap Suermondt. Hewlett-Packard Labs. This dataset contains 4601 instances and 58 attributes (57 continuous input attribute and 1 nominal class label target attribute. The attribute description [23] are given in Table I.

B. *Pre-processing*

Today, most of the data in the real world are incomplete containing aggregate, noisy and missing values. As the quality decision depends on quality mining which is based on quality data, pre-processing becomes a very important tasks to be done before performing any mining process. Major tasks in data pre-processing are data cleaning, data integration, data transformation and data reduction. In this dataset data normalization is done before performing feature relevance analysis.

TABLE I
Attribute description

IV. Feature relevance analysis

Attribute Number	Attribute Type	Attribute Description
A1 to A48	char_freq_CHAR	percentage of characters in the e-mail that match CHAR
A49 to A54	capital_run_length_average	average length of uninterrupted sequences of capital letters
A55	capital_run_length_longest	length of longest uninterrupted sequence of capital letters
A56	capital_run_length_longest	length of longest uninterrupted sequence of capital letters
A57	capital_run_length_total	total number of capital letters in the e-mail
A58	Class attribute	denotes whether the e-mail was considered spam (1) or not (0)

Feature relevance analysis has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities. The main objective of feature selection is to choose a subset of input variables by eliminating features, which are irrelevant or of no predictive information. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results. Feature selection in supervised learning has a main goal of finding a feature subset that produces higher classification accuracy.

A. Feature Reduction Techniques

Correspondence Analysis

Correspondence Analysis is a multivariate statistical technique proposed by Hirschfield and later developed by Jean-Paul Benzécri. It is conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarizing a set of data in two-dimensional graphical form.[25]

Canonical Discriminant Analysis

Canonical Discriminant Analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. Given a nominal classification variable and several interval variables, canonical discriminant analysis derives canonical variables (linear combinations of the interval variables) that summarize between-class variation in much the same way that principal components summarize total variation [9].

Principal Component Analysis

Principal Component Analysis is a dimensionality reduction technique which enables to visualize a dataset in a lower dimension without loss of information. It is appropriate when obtaining measures on a number of observed variables and wish to develop a smaller number of artificial variables (called principal components) that will account for most of the variance in the observed variables [10].

B. Feature Selection Algorithms

Fisher filtering is a supervised feature selection algorithm [22] which processes the selection independently from the learning algorithm. It follows univariate Fisher's ANOVA ranking which ranks the inputs attributes according to their relevance without considering the redundancy aspects of input attributes.

ReliefF algorithm detects conditional dependencies [20] between attributes and provides a unified view on the attribute estimation in regression and classification. It is not limited to two class problems, is more robust and can deal with incomplete and noisy data.

STEPDISC (Stepwise Discriminant Analysis) [22] procedure performs a stepwise discriminant analysis to select a subset of the quantitative variables for use in discriminating among the classes. The set of variables that make up each class is assumed to be multivariate normal with a common covariance matrix. The STEPDISC procedure can use forward selection, backward elimination, or stepwise selection.

Runs filtering are a non parametric test [22] for predictive attribute evaluation. It is an univariate attribute ranking from runs test. It is a supervised feature selection algorithms based upon a filtering approach i.e. processes the selection independently from the learning algorithm. This component ranks the inputs attributes according to their relevance without considering redundancy aspect. A cutting rule enables to select a subset of these attributes.

After performing feature relevance analysis, various classification algorithms are applied over this training dataset before filtering irrelevant attributes as well as on the relevant attributes after filtration.

V. Study on classification algorithms

C4.5

It works similar to ID3 and builds the decision trees, using the concept of information entropy. Information entropy is a measure of the uncertainty associated with a random variable. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets. Its criterion is the normalized information gain that results from choosing an attribute for splitting the data. Information gain is the difference in entropy associated with an attibute. The attribute with the highest normalized information gain is chosen to make the decision [20].

C-RT & CS-CRT

The CART method under Tanagra is a very popular classification tree learning algorithm. CART builds a decision tree by splitting the records at each node, according to the function of a single attribute it uses the gini index for determining the best split. The CS-CRT is similar to CART but with cost sensitive classification [22].

ID3

In ID3 decision tree, each node corresponds to splitting attribute. It uses information gain to determine the splitting attribute. The attribute with the highest information gain is taken as the splitting attribute. Information gain is the difference between the amount of information needed to make a correct prediction before and after splitting. Information gain can also be defined as the different between the entropy of the original segment and the accumulated entropies of the resulting split segments. Entropy is the measure of disorder found in the data. ID3 can handle high-cardinality predictor variables. A high-cardinality predictor is a variable which has different possible values thus having different possible ways of performing a split [8].

K-NN

The *k*-nearest neighbor algorithm (*k*-NN) is a method for classifying objects based on closest training examples in an n-dimensional pattern space. When given an unknown tuple the classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuple are the k nearest neighbour of the unknown tuple [19].

LDA

Linear discriminant analysis (LDA) is a classical statistical approach developed by R.A Fisher for classifying samples of unknown classes, based on training samples with known classes. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements [19].

Log Regression TRIRLS

LR-TRIRLS stands for Logistic Regression with Truncated Regularized Iteratively Re-weighted Least Squares. LR-TRIRLS uses IRLS a quasi-Network method to learn the LR parameters, with some modifications. It is also used for fitting Generalized Linear Models (GLiMs) to data. It is an iterative method for solving a weighted least squares (WLS) linear regression problem on each iteration.

Multilayer Perceptron

It is the most popular network architecture in today world. The units each perform a biased weighted sum of their inputs and pass this activation level through a transfer function to produce their output. The units are arranged in a layered feed forward topology. The network has a simple input-output model, with the weights and thresholds. Such networks can model functions of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptrons is the design specification of the number of hidden layers and the number of units in these layers [20].

Naïve Bayes Continuous

The Naïve Bayes Classifier technique [16] is based on the Bayesian theorem and is particularly suited inputs which has dimension high. Despite of its simplicity, It can often outperform more sophisticated classification methods. The Naive Bayes continuous which works similarly with continuous variable as input [20].

PLS-DA & PLS – LDA

PLS-DA is a regression technique usually designed to predict the values taken by a group of dependent variables from a set of independent variables. For the prediction of continuous target variable, the PLS regression can be adapted to the prediction of one discrete variable - i.e. the supervised learning framework in different ways .which is referred as "PLS Discriminant Analysis". In PLS-LDA (PLS Linear Discriminant Analysis) is similar to PLS-DA where the number of descriptors is moderately high in relation to the number of instances.

Random Forest Tree(Rnd Tree)

A Random Tree consists of a collection or ensemble of simple tree predictors, each capable of producing a response when presented with a set of predictor values. For classification problems, this response takes the form of a class membership, which associates, or classifies, a set of independent predictor values with one of the categories present in the dependent variable. For regression problems, the tree response is estimated for the dependent variable given by the predictors [20].

SVM

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. The standard SVM takes a set of input data and predicts, for each given input, which of two possible classes comprises the input, making the SVM a non-probabilistic binary linear classifier.

VI. Experimental results and Performance evaluation

The spambase dataset is downloaded from the UCI machine learning repository[23] in the form of text file. This dataset contains 57 input attributes of continuous format and 1 target attribute in discrete format. Then feature construction is done for feature transformation. Since the training dataset contains all the input attributes as continuous and target attribute as discrete, the following four feature selection algorithms namely, Fisher filtering, ReliefF, Runs Filtering and Step disc are executed on this dataset for retrieving relevant features and the results are given in Table II. Classification algorithms such as Naive bayes continuous, ID3 ,K-NN, multilayer perceptron, C-SVC, Linear discriminant analysis, CS-MC4, Rnd tree, PLS-LDA, PLS-DA etc, are applied to each of the above filtering algorithms and the results are given in Fig. 2.

Fisher filtering produces above 95% accurate results for 3 classifiers (C4.5, CS-MC4 and Rnd – tree classification algorithms); above 90% accuracy for 8 classifiers and above 85% for 6 classifiers. ReliefF filtering produce above 95% accuracy for only 1 classifier (Rnd Tree); above 90% accuracy for 6 classifiers; above 85% for 6 classifier and above 80% for 4 classifiers. Runs filtering and Stepwise discriminant analysis provides best result for 2 classifiers (C4.5 and CS-MC4); above 90% for 10 classifiers and above 85% accuracy for 5 classifiers. Runs filtering and Step disc feature selection algorithms almost provide the same result. From the results, the Rnd tree classification is considered as a best classifier, as it produced 99% accuracy through fisher filtering feature selection.

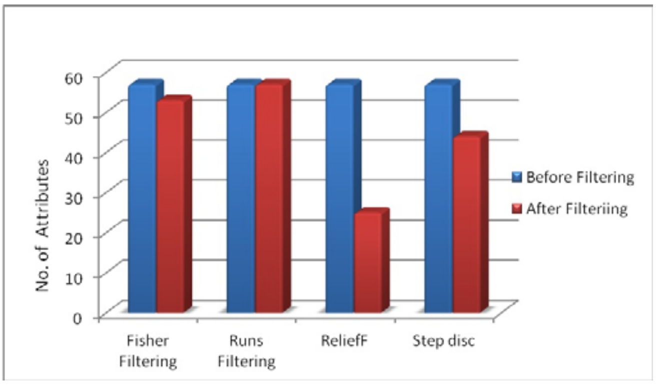

Fig 2. Feature Selection

TABLE II
RESULTS OF CLASSIFICATION ALGORITHMS

Classification Algorithms	Error rate before filtering	Error rate after filtering			
		Fisher filtering	ReliefF	Runs Filtering	Step disc
C4.5	0.0367	0.0363	0.0513	0.0367	0.0372
C-PLS	0.0898	0.1024	0.1463	0.0898	0.0919
C-RT	0.0596	0.0535	0.0739	0.0596	0.0659
CS-CRT	0.0596	0.0535	0.0739	0.0596	0.0659
CS-MC4	0.0385	0.0385	0.0676	0.0385	0.0396
CS-SVC	0.0767	0.0815	0.1206	0.0767	0.0782
ID3	0.0863	0.0863	0.1050	0.0863	0.0895
K-NN	0.0569	0.0609	0.0824	0.0596	0.0650
LDA	0.1113	0.1139	0.1519	0.1113	0.1119
Log Reg TRIRLS	0.1389	0.1448	0.1821	0.1389	0.1413
Multilayer Perceptron	0.0461	0.0541	0.0815	0.0393	0.0519
Multilogical Logistic Regression	0.0687	0.0689	0.1117	0.0687	0.0706
Naïve Bayes Continuous	0.1126	0.1135	0.1413	0.1126	0.1171
PLS-DA	0.1121	0.1248	0.1526	0.1121	0.1174
PLS-LDA	0.1121	0.1243	0.1524	0.1121	0.1171
Rnd Tree	0.0117	0.0089	0.0324	0.0117	0.0100
SVM	0.0924	0.0930	0.1361	0.0924	0.0950

A. Error rate

Error rate of a classifier was defined as the percentage of the dataset incorrectly classified by the method. It is the probability of misclassification of a classifier

$$\text{Error rate} = \frac{\text{No. of incorrectly classified samples}}{\text{Total No. of Samples in the Class}}$$

B. Accuracy

Accuracy of a classifier was defined as the percentage of the dataset correctly classified by the method. The accuracy of all the classifiers used for classifying spam dataset is represented graphically in Fig 2.

$$\text{Accuracy} = \frac{\text{No. of correctly classified samples}}{\text{Total No. of Samples in the Class}}$$

C. Recall

Recall of the classifier was defined as the percentage of errors correctly predicted out of all the errors that actually occurred.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

D. Precision

Precision of the classifier was defined as the percentage of the actual errors among all the encounters that were classified as errors.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

The terms positive and negative refer to the classifier's prediction, and the terms true and false refer to classifier’s expectation. The precision and recall and error rate for the Rnd tree classifier is done and the results are given in Fig 3.

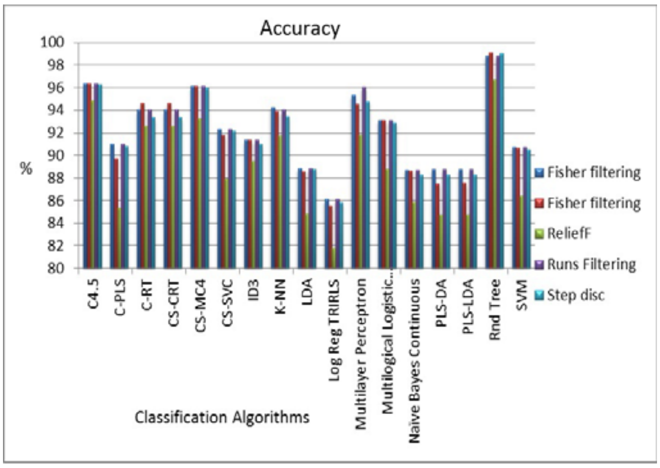

Fig 3. Accuracy of the Spam classification algorithm

TABLE IV
RESULTS OF BEST CLASSIFIER (Rnd Tree classifier)

Error rate			0.0089			
Values prediction			Confusion Matrix			
Value	Recall	1-Precision		Spam	No Spam	Sum
Spam	0.9901	0.0127	Spam	1795	18	1813
No Spam	0.9918	0.0065	No Spam	23	2765	2788
			Sum	1818	2783	4601

VII. Conclusion

Email spam classification has received a tremendous attention by majority of the people as it helps to identify the unwanted information and threats. Therefore, most of the researchers pay attention in finding the best classifier for detecting spam emails. From the obtained results, fisher filtering and runs filtering feature selection algorithms performs better classification for many classifiers. The Rnd tree classification algorithm applied on relevant features after fisher filtering has produced more than 99% accuracy in spam detection. This Rnd tree classifier is also tested with test dataset which gives accurate results than other classifiers for this spam dataset.

Acknowledgment

The authors would like to thank Dr. K.Sarukesi, Vice chancellor of Hindustan Institute of Technology and Science, Chennai – India Dr. Ponnammal Natarajan, Advisor, (Research and Development), Rajalakshmi Engineering College and Dr. K. Ravi, Associate Professor and Head, Department of Mathematics, Sacred Heart College,Tirupattur, India for their intuitive ideas and fruitful discussions with respect to the paper's contribution.

References

[1] Androutsopoulos .I, J. Koutsias, K.V. Chandrinos, G. Paliouras, and C.D. Spyropoulos. An Evaluation of Naive Bayesian Anti-Spam Filtering. Proceedings of the Workshopon Machine Learning in the New Information Age, 11th European Conference on Machine Learning, Barcelona, Spain, pages 9–17, 2000.

[2] Androutsopoulos I., J. Koutsias, K.V. Chandrinos, and C.D. Spyropoulos. An Experimental Comparison of Naive Bayesian and Keyword-Based Anti-Spam Filtering with Encrypted Personal Messages. Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Athens, Greece, 2000.

[3] Apte, C. and F. Damerau. Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems, 12(3):233–251, 1994.

[4] Biro. I, J. Szabo, and A. A. Benczur. Latent Dirichlet," location in Web Spam Filtering". In Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2008.

[5] A. Bratko, B. Filipic, G. Cormack, T. Lynam, and B. Zupan. "Spam Filtering Using Statistical Data Compression Models", The Journal of Machine Learning Research, pp., 2673–2698, 2006.

[6] Cohen, W, Learning rules that classify e-mail. In Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access. Palo Alto, California, 1996.

[7] C. Paulo, L. Clotilde, S. Pedro et al., "Symbiotic data mining for personalized spam filtering," in Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, (IEEE/WIC/ACM), pp. 149–156, 2009.

[8] E.A. Erosheva and S.E. Fienberg. Bayesian mixed membership models for soft clustering and classification. Manuscript, 2004.

[9] http://en.wikipedia.org/wiki/Linear_discriminant_analysis.

[10] Husna, H. Phithakkitnukoon, S. Palla, S. Dantu, R., "Behavior analysis of spam botnets" in *IEEE xplore*, ISBN 978-1-4244-1796-4 , Issue date 6-10 Jan. 2008.

[11] Kh. Ahmed, "An overview of content-based spam filtering techniques," Informatica, vol. 31, no. 3, pp. 269–277, 2007.

[12] K. S. Xu, M. Kliger, Y. Chen, P. J. Woolf, and A. O. Hero, "Revealing social networks of spammers through spectral Clustering," in Proceedings of the IEEE International Conference on Communications, (ICC '09), Dresden, Germany, June 2009.

[13] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing email detection based on structural properties," in *Proc. 9th Annual NYS Cyber Security Conf.*, Jun. 2006.

[14] M. Prince, L. Holloway, E. Langheinrich, B. M. Dahl, and A. M. Keller, "Understanding how spammers steal your e-mail address: An analysis of the first six months of data from Project Honey Pot," in Proc. 2nd Conf. Email and Anti-Sp*am*, Jul. 2005.

[15] Perkins, A. The classification of search engine spam. http://www. ebrand management.Com/ white papers/spam classification.

[16] Phimphaka Taninpong, Sudsanguan Ngamsuriyaroj," Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification" , 2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing. *Materials Research*, Vols. 171-172, pp. 543-546, 2011.

[17] RasimM. Alguliev, Ramiz M. Aliguliyev, and Saadat A. Nazirova," Classification of Textual E-Mail Spam Using Data Mining Techniques", Applied Computational Intelligence and Soft Computing,Volume 2011.

[18] R.Deepa Lakshmi, N.Radha," Supervised Learning Approach for Spam Classification Analysis using Data Mining Tools", (IJCSE) International Journal on Computer Science and Engineering. Vol. 02, No. 08, 2010, 2760-2766.

[19] R. Geetha Ramani, G. Sivagami, Parkinson Disease Classification using Data Mining Algorithms, International Journal of Computer Applications (0975 – 8887) Volume 32– No.9, October 2011.

[20] Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical Data through Data Mining Algorithms: Multi-class Categorization of Breast Tissue Data, International Journal of Computer Applications (0975 – 8887) Volume 32– No.7, October 2011.

[21] S. Nazirova, "Mechanism of classification of text spam messages collected in spam pattern bases," in Proceedings of the3rd International Conference on Problems of Cybernetics and Informatics, (PCI '10), vol. 2, pp. 206–209, 2010.

[22] Tanagra-Data Mining tutorials http://data-mining-tutorials.blogspot.com/

[23] UCI Machine Learning Repository – Spambase Dataset http://archive.ics.uci.edu/ml/datasets/Spambase

[24] X. Li and N. Ye, "A supervised clustering and classification algorithm for mining data with mixed variables," IEEE Transactions on Systems, Man, and Cybernetics Part A, vol. 36, no. 2, pp. 396–406, 2006.

[25] Yang Kyu shin," The exploratory Analysis of spam mail data using correspondence Analysis" *Journal of Korean data & Information Science Society*, 2005, vol.16, No 4,pp. 735~744.

[26] Yue Yang, Elfayoumy. S," Anti-Spam Filtering Using Neural Networks and Bayesian Classifiers", Proceedings of the 2007 IEEE International Symposium on Computational Intelligence in Robotics and Automation, Jacksonville, FL, USA, June 20-23,2007