*I.J. Modern Education and Computer Science*, **2013, 12, 31-39**
Published Online December 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2013.12.05

# Spam Mail Detection through Data Mining – A Comparative Performance Analysis

Megha Rathi
Department of computer Science Engineering of Jaypee Institute of Information Technology, Noida, India
E-mail: megha.rathi@jiit.ac.in

Vikas Pareek
Department of Computer Science of Banasthali University, Banasthali, India
E-mail: er_pareekvikas@yahoo.co.in

***Abstract —*** As web is expanding day by day and people generally rely on web for communication so e-mails are the fastest way to send information from one place to another. Now a day's all the transactions all the communication whether general or of business taking place through e-mails. E-mail is an effective tool for communication as it saves a lot of time and cost. But e-mails are also affected by attacks which include Spam Mails. Spam is the use of electronic messaging systems to send bulk data. Spam is flooding the Internet with many copies of the same message, in an attempt to force the message on people who would not otherwise choose to receive it. In this study, we analyze various data mining approach to spam dataset in order to find out the best classifier for email classification. In this paper we analyze the performance of various classifiers with feature selection algorithm and without feature selection algorithm. Initially we experiment with the entire dataset without selecting the features and apply classifiers one by one and check the results. Then we apply Best-First feature selection algorithm in order to select the desired features and then apply various classifiers for classification. In this study it has been found that results are improved in terms of accuracy when we embed feature selection process in the experiment. Finally we found Random Tree as best classifier for spam mail classification with accuracy = 99.72%. Still none of the algorithm achieves 100% accuracy in classifying spam emails but Random Tree is very nearby to that.

***Index Terms —*** Classifier, Feature Selection, E-mails, Spam Mails.

## I. INTRODUCTION

E-Mail is an effective way of communication as it saves a lot of time and money this makes it as a favourite means of communication in personal as well as in professional communication. E-mails provide a way for internet users to easily transfer information globally. But there is also a case when your e-mails are affected by attacks whether active or passive. Sometimes we receive e-mail from unknown source and also e-mail comprised of contents which is of no importance to the user. These kind of unwanted mails are better known as Spam Mails. Spam email is the practice of frequently sending unwanted data or bulk data in a large quantity to some email accounts. Spam Mail is a subset of electronic spam involving nearly identical messages sent to various recipients by email. Spam mails also include malware as scripts or other executable file attachment. There are two main types of spam and they have different affects on Internet users. Cancellable Usenet spam is a single message sent to 20 or more Usenet groups. Usenet spams aims at "lurkers", people who read newsgroups but rarely or never post and give their address away. Usenet spam subverts the ability of system administrator to manage the topics they accept on their systems. Another type of Email spam targets individual users with direct mail messages. Email spam list are created by scanning Usenet postings, stealing Internet mailing list. Email spam is any email that meets the following three criteria:

1) Anonymity: The address and identity of the sender are concealed.

2) Mass Mailing: The email is sent to large group of people.

3) Unsolicited: The email is not requested by recipients.

Spam Mail has become an increasing problem in recent years. It has been estimated that around 70% of all emails are spam. As the usage of web expanding, problem of spam mails are also expanding. According to [1] it has been found that on an average 10 days per year waste on dealing with spam mails only. Spam is an expensive problem that costs billion of dollars per year to service providers for lost of bandwidth. Spam is a major problem that attacks the existence of electronic messages. So it is very essential to distinguish emails from spam mails, many methods have been proposed for classification of email messages as spam mail or legitimate mail and it has been found that machine learning algorithm success ratio for classification is very high [2].

Several algorithms are used for classification of spam mails which are extensively utilize and analyze out of which support vector machine, Na ïve Bayes, Decision Tree, Neural network classifiers are well known classifiers. In this paper we experiment our data set with these given algorithms: Naïve Bayes, Bayes Net, Support vector machine (SVM), function Tree (FT), J48, Random Forest and Random Tree. Initially we experiment on entire data set which consists of total 58 attributes and total number of instances is 4601. We apply above mentioned algorithm one by one on the data set and check the result and it is retrieved from the study that out of all these classifiers Random Forest and Random Tree works well and gives accuracy better than other classifiers in detection of spam mails. In order to compare the result that classifiers works well with some attributes selected or not, then we apply Feature selection algorithm on the same dataset (the algorithm we used here is Best First Search algorithm) and apply the same classifiers with features selected. Out of 58 features only 15 features are selected and apply the same above mentioned algorithm on this reduced dataset. From this study it is found that all classifier's accuracy improved when we select features through Best-First algorithm. Again when compared with all classifiers which we experimented on this reduced data set Random Tree shows better results in context of accuracy.

This paper is organized as follows: Section 2 comprised of Background study, Section 3 presents related work, Section 4 presents the Experimental work and results, Section 5 presents Experimental Results and Section 6 presents conclusion and future work.

## II. BACKGROUND STUDY

This section presents an overview of what is Data Mining, different algorithm of data mining, explains Feature selection and most of the terms that we used in this paper.

### *A. Data Mining*

Data Mining is basically the discovery of knowledge from the large database. It is a technique that attempts to find out new patterns in huge data sets. It is mixture of various fields like Artificial Intelligence, Machine Learning, statistics, and Database systems. The main objective of data mining approach is to extract information from a data set and transform it into and understandable form for further use. The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously undetermined interesting patterns. Data Mining is the process of analyzing data from different perspective and summarizing it into useful information and this information can be used to increase revenue, cut costs, for classification, prediction etc. It is the process of finding correlations in large relational databases. While large-scale information technology has been evolving separate transactions and analytical systems, data mining provides the link between the two approaches. Data mining software analyzes relationships in stored data based on end user queries. In general these 4 types of relationships are sought:

1) Classes: Class is used to place the data in predetermined groups.

2) Clusters: Data items are placed in a group according to logical relationships. For example, data can be mined to identify market segments.

3) Associations: Data mining is applied to data set to find out the associations.

4) Sequential Patterns: Data is mined to anticipate behavior patterns and trends.

Basically Data mining involves listed five elements:

1) Extract, transform, and load data on data warehouse system.

2) Store and manage data in multidimensional database system.

3) Provide data access in an easier manner to business analyst and technical professionals.

4) Analyze data by existing tool/application software.

5) Make data in format which is useful to concerned user such as graph or tables.

Sometimes we treat data mining as a synonym for another known term, Knowledge discovery from databases (KDD), because data mining is necessary step in the process of knowledge discovery from the database. Knowledge discovery is a combination of all these steps shown in fig.1.

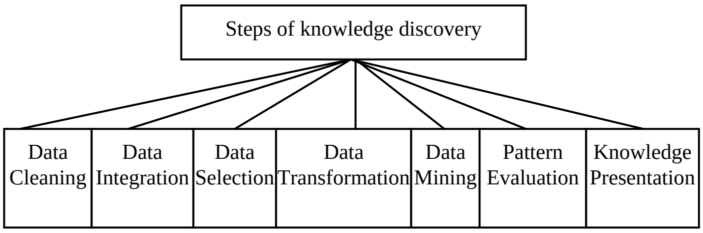


Figure. 1: Knowledge Discovery steps

Data mining involves many different algorithms to achieve the desired tasks. All of these algorithms try to fit a model, the algorithm examine the data and find out the model that is closest to the characteristics of the data being examined. Data mining algorithms characterized based on the purpose of the algorithm to fit a model to the data, based on Preference, and all algorithms require some approach for searching. Fig.2 shows the model than can be either predictive or descriptive.

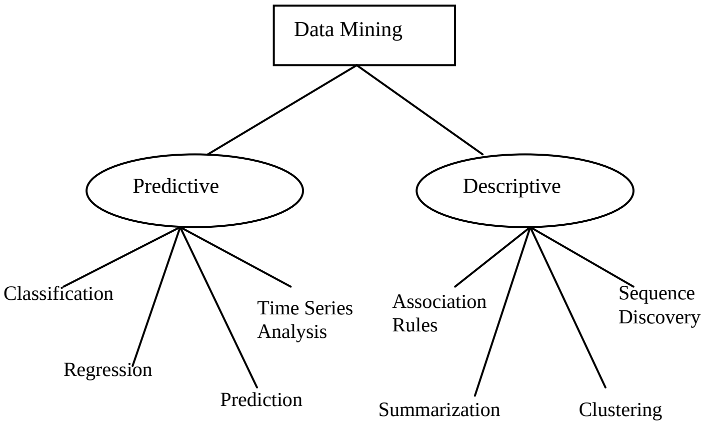


Figure. 2: Data Mining Models and tasks

## B. Algorithm used in this study

### 2.2.1 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 lies 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 [3]. 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.

### 2.2.2 Naïve Bayes

A naïve Bayes classifier is a simple probabilistic classifier with strong independence assumptions. In simple terms, a naïve bayes classifier assumes that the presence/absence of a particular feature of a class is unrelated to the presence/absence of any other feature, given the class variable depending on the nature of probability model, naïve bayes classifier can be trained in supervised learning setting. An advantage of the naïve bayes classifier is that it only requires a small amount of training data to estimate the parameters required for classification. In Bayesian classification we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis to be true. Bayesian classifiers are basically statistical classifiers i.e. they can predict the class membership probabilities, such as the probability that a given sample data belongs to a particular class.

The naïve Bayes technique is based on Bayesian approach hence it is a simple, clear and fast classifier [4]. Before reaching to the main term of Baye’s theorem we will first analyze some terms used in the theorem. P (A) is the probability that event A will occur. P (A/B) is the probability that event A will happen given that event B has already happened or we may define it as the conditional probability of A based on the condition that B has already happened. Bayes theorem is defined in equation 1.

$$P(A/B) = P(B/A)\, P(A)\, P(B) \quad (1)$$

If we consider X to be an object to be classified with the probabilities of belonging to one of the classes C1,C2,C3 etc. by calculating $P(C_i/X)$. Once these probabilities have been computed for all the classes, we simply assign X to the class that has highest probability.

$$P(C_i/X) = [P(X/C_i)\, P(C_i)] / P(X) \quad (2)$$

Where $P(C_i/X)$ is the probability of the object X belonging to a class $C_i$, $P(X/C_i)$ is the probability of obtaining attribute values X if we know that it belongs to class Ci. $P(C_i)$ is the probability of any object belonging to a class $C_i$ without any other information, and P(X) is the probability of obtaining attribute values X whatever class the object belongs to.

### 2.2.3 Decision Tree

A decision tree is a classification method that results in a flow-chart like tree structure where each node denotes a test on attribute value and each branch represents an outcome of the test. The tree leaves represents the classes. Decision tree is model that is both predictive and descriptive; it represents relationships found in training data. The tree consists of zero or more internal nodes and one or more leaf nodes with each internal node being a decision node having two or more child nodes. Decision tree use divide and conquer method to split the problem search space into subsets. Decision tree is constructed to model the classification process. Once the tree is built it is applied to each tuple in the database and results in a classification for that tuple. There are two basic steps in this technique: building the tree and applying the tree to the dataset. The decision tree approach to classification is to divide the search space into rectangular regions. A tuple is classified based on the region into which it falls.

Given a database D = {$t_1$, t2,….., tn} where $t_i$ = { $t_{i1}$ ,……, $t_{ih}$} and the database schema contains the following attributes {$A_1$, $A_2$, ……, $A_h$}. Also given is a set of classes C = {$C_1$,……, $C_m$ }. A Decision tree is a tree associated with D that has the following properties:

1) Each internal node is labeled with an attribute $A_i$.

2) Each edge is labeled with a predicate that can be applied to the attribute associated with the parent.

3) Each leaf node is labeled with a class $C_j$.

### 2.2.4 Feature Selection

Feature Selection also known as feature reduction, attribute selection is the technique of selecting a subset of relevant features for building the learning models. Feature selection is very important step in analyzing the data, by removing irrelevant and redundant features from the data. Feature selection overall improves the performance of learning model by:

1) Alleviating the effect of curse of dimensionality.
2) Enhancing generalization capability.
3) Speeding up learning process.
4) Improving model interpretability.

Feature Selection helps in gaining the better understanding of the data by telling which are the important attributes or features and how they are related with each other. It is the process of selecting a subset of the terms occurring in the training set and using only this subset as features in classification. It serves two main purposes: First, it makes training and applying a classifier more efficient by decreasing the size of data set. Second, feature selection enhances accuracy of classifier by eliminating extra features from the data set. A Feature selection algorithm is a computational solution which is motivated by certain rules of relevance. An irrelevant feature is not useful for induction, but it also not essential that all relevant features are used for induction [5]. Feature Selection algorithm can be classified according to the kind of output they produce: (1) algorithms that produce a linear order of features and (2) algorithms that produce a subset of original features. In the study [6, 7, 8] characterization of Feature selection algorithm is described. In this context it is possible to describe this characterization as a search problem as follows:

1) Search Organization. This technique is related to the portion of hypothesis investigated with respect to their total number.

2) Generation of Successors. This technique defines by which possible variants of the current hypothesis are proposed.

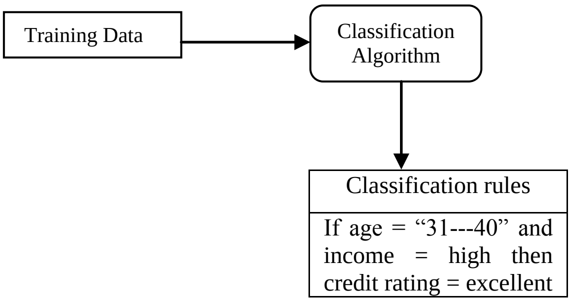
3) Evaluation Measure. Method, by which variants or successor candidates are evaluated, allowing to compare different hypothesis to supervise the search process.

In Feature Selection Algorithm we select a subset of features. Subset selection evaluates a subset of features and these algorithms can be broken into Wrappers, filters and Embedded. Wrappers use a search algorithm to find out the space of possible features. Wrappers are computationally expensive and have a risk of over fitting the model. Filters are same as Wrappers in context of search space, but instead of evaluating against a model, a simpler filter is evaluated. Embedded approach is embedded in and specific to a model.

Following are some extensively used Feature selection algorithms: (1) Best First (2) Simulated Annealing (3) Genetic algorithm (4) Scatter Search and (5) Greedy forward selection etc.

### 2.2.5 *Classification and Prediction*

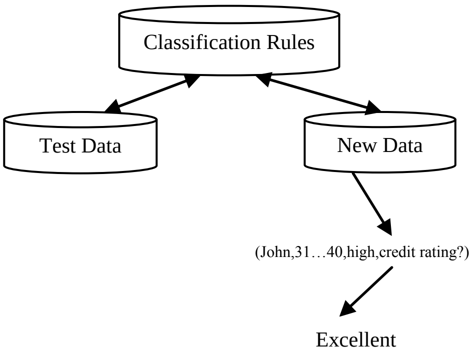
Classification is the separation of objects into classes. If the classes are created without looking at the data then the classification is known as apriori classification. If classes are created by looking at the data then the classification method is known as posterior classification. On classification it is assumed that the classes have been deemed apriori and classification then consists of training the system so that when a new object is introduced to the trained system it is able to assign the object to one of the existing classes. This approach is better known as supervised learning. Data Classification is a two step process as shown in fig. 3. In the first step, model is built describing a predetermined set of data classes. The model is constructed by analyzing database tuples described by the attributes. Each tuple is assumed to belong to one of the existing class, as determined by the class label attribute. The data tuples analyzed to build the model collectively form the training set.


Training Data
Classification Algorithm
Classification rules
If age = “31---40” and income = high then credit rating = excellent

| Name | Age | Income | Credit rating |
|---|---|---|---|
| Sandy | <=30 | low | fair |
| Bill | <=30 | low | excellent |
| Susan | >40 | medium | fair |

Figure. 3: Learning and Training of classifier

In the second step as shown in fig. 4, the model is used for classification. First the predictive accuracy of the model is estimated. The accuracy of a model on a given test data set is the percentage of test set samples that are correctly classified by the model. For each test sample the known class label is compared with the learned model’s class prediction for that sample.



| Name | Age | Income | Credit rating |
|---|---|---|---|
| Frank | > 40 | high | fair |
| Crest | <=30 | low | fair |
| Annee | 31…40 | high | excellent |

Figure. 4: Classification

Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled sample, or to assess the value ranges of an attribute that a given sample is likely to have. In this context classification and regression are the two major types of prediction problems, where classification is used to predict discrete or nominal values, while regression is used to predict continuous or ordered values.

## III. RELATED WORK

Spam Mails are one of the major problematic areas in the Internet world which can bring financial loss to organizations and also bring damage to individual users as well. Email spam also known as junk mails which are sent to a group of recipients who have not requested it. Spam is a serious problem that threatens the existence of e-mail services. As it involves no cost so it is quiet cheap to send bulk e-mail to a group of users. It consumes a lot of time to delete or sort these spam emails and also introduces a risk of deleting normal mails by mistake. In the study [9] Rambow et al. applied machine learning techniques for email summarization. In this study, RIPPER classifier is used for the determination of sentences which should be included in a summary. Learning model use features such as Linguistic feature, email features, and threading structure. This approach requires positive examples in huge quantity and it is also found that summaries are not produced for varying length based on user interest.

There are so many existing techniques for detection of these spam emails. These approaches come mainly from the area of Artificial Intelligence, Data Mining, or Machine Learning. Machine learning techniques are more varied and used extensively for spam mail classification. Decision tree classify spam mails using previous data [10]. But it is costly to calculate and recalculate as spammers change technique. In the study [11] Bayesian networks found as the very popular technique for spam mail detection. But with this approach it is quiet difficult to scale up on many features to come out with the judgement.

In [12] fuzzy clustering approach is used. In this paper author evaluated the use of fuzzy clustering and text mining for spam filtering. Fuzzy clustering is scalable and easy to update approach. This study deals with the examination of use of fuzzy clustering algorithm to build a spam filter. Classifier has been tested on different data sets and after testing Fuzzy C-Means using Heterogeneous Value Difference Metric with variable percentages of spam and used a standard model of assessment for the problem of spam mail classification. This paper makes use of text mining and fuzzy clustering as an anti-spam technique. If each email that comes in is used as part of the data pool to make decisions about future emails, spam trends will be detected. It is found that there is not large cost of calculation and recalculation that would occur with decision tree, or with some rule-based filters.

We all were aware about the fact that Spam mails create a lot of problem in today's world. So various approaches are developed to stop spam mails. The main objective in spam filtering is to rule out the unwanted emails automatically from user inbox. These unwanted are root cause for the problems like filling mailboxes, engulfing important personal mail, wasting a lot of network bandwidth also causes congestion problem, time and energy loss to the users while sorting these unwanted mails [13]. In the study [14] two methods are described for classification. First is done with some rules that are defined manually, like rule based expert system. This technique of classification is applied when classes are static, and their components are easily separated in accordance with the features. Second is done with the help of existing machine learning techniques. According to the study [15] clusters of spam emails are created with the help of criterion function. Criterion function is defined as the maximization of similarity between messages in clusters and this similarity is calculated using k-nearest neighbour algorithm.

Symbiotic Data Mining is a distributed data mining approach which unifies content based filtering with collaborative filtering is described in [16]. The main objective is to make use of local filters again in order to improve personalized filtering in context of privacy. In study [17] email classifiers based on the approach of feed forward back propagation neural network and Bayesian classifiers are evaluated. From this study it is found that feed forward back propagation neural network classifier provides very high accuracy as compared to other existing classifiers. In the paper [18] Bayesian approach is applied for the problem of classification and clustering using model based on the assumptions like: population, subject, latent variable, and sampling scheme.

According to [19] content filtering was one of the first types of anti spam filter. These types of filters make use of hard coded rules which has an associated score and is updated periodically. One main example of such type of filter is Spam Assassin [20] which works by scanning the text document of the e-mail against each rule and add score for all matching rules. According to the study [21] if total score of the e-mail exceeds some set threshold score then that message falls into spam mail category. In order to generate these score a single perceptron is used where the inputs to the perceptron indicate whether a rule was matched and the weight for the corresponding input indicates the score for each rule.

In the paper [22] 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. At each iteration an error weight adjustment value are computed by comparing the actual output value with the expected output value. Testing phase was done by subjecting the Artificial Neural Network to messages that were not used in training without adjusting the weights.

## IV. PROPOSED WORK

In this study we detect spam mails using various classifiers. The whole experiment comprised of two parts. First we will apply various classifiers for spam mail classification and check the results in terms of accuracy for each classifier. Here we use the entire data set and apply algorithm one by one without selecting any feature. In the second part we detect spam mails by not using the entire data set instead we apply feature selection algorithm first, the algorithm which we use here is Best-First Feature Selection algorithm then on the reduced data set with selected features we will apply all the classifiers one by one and check the results. It is found that classifier's accuracy improved when we embed feature selection algorithm in the process. These are some of the classifiers that we use in this study: (1) Naïve Bayes (2) Bayesian Net (3) Support Vector Machine (SVM) (4) Function Tree (FT) (5) J48 (6) Random Forest (7) Random Tree and (8) Simple Cart. We find out accuracy, Kappa statistics (KS), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE) for all the classifiers and compare the results based on all these statistics. Same characteristics are also compared for all classifiers when we use feature selection in the process.

The dataset which we use is from UCI Machine Learning Repository [23] which consists of 58 attributes where 57 continuous attribute and 1 nominal class label attribute and total number of instances is 4601. Table I presents the entire dataset with attribute description.

TABLE 1: Dataset attribute Description

| 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 e-mail was considered as spam with class label (1) and not spam with class label (0). |

The overall design of the proposed system is depicted in Fig.5 for classification of e-mail as spam without taken into consideration the feature selection approach and Fig.6 for classification of e-mail as spam with taken into consideration the Feature Selection Approach.

Below shown (Fig. 5 and Fig.6) is the overall architecture of proposed system. In this architecture first we train the spam data set which comprised of 58 attributes with total 4601 instances. Then we apply Preprocessing, as we all know that real world data contains missing values or noisy values so in order to produce good results from the data set we need to mine data. As quality decision depends on good quality data, pre-processing is crucial step before applying any classifier to the data set. Pre-Processing involves the tasks like data cleaning, data integration, data transformation, or data reduction. Before applying any data mining techniques to the data set we first normalize the entire data set in order to yield good results. Up to this step both the proposed system works similarly then after as per architecture I shown in fig.5 we apply classifiers one by one to the entire data set and evaluate the performance of classifier. Then test the data using the classifiers and classify mails as spam and non spam. As per proposed architecture II shown in Fig. 6 after the pre-processing step we first apply Feature selection algorithm, the algorithm which we deploy here is Best-First Feature Selection algorithm. Table II reflects the view of selected features after applying the algorithm to the data set.

TABLE 2: Selected Attributes after Feature Selection

| Attribute number | Attribute type | Attribute Description |
|---|---|---|
| 4,5,7,16,21,23,24,<br>25,27,42,44,46 | char_freq_CHAR | Percentage of characters in the e-mail that match CHAR |
| 52,53 | Capital_run_length_average | Average length of uninterrupted sequences of capital letters. |
| 55 | capital_run_length_longest | Length of longest uninterrupted sequence of capital letters. |

Total 15 attributes are selected out of 58 attributes. Attribute number 4, 5, 7, 16, 21, 23, 24, 25, 27, 42, 44, 46, 52, 53 and 55are the selected attributes. Classification algorithms are applied one by one on all these selected 15 attributes and results which we are getting are more promising than without applying feature selection approach in the entire process of detection of spam mails. Feature Selection has been an active and fruitful field of research in machine learning, statistics and data mining. The main aim of this approach is to select a subset of data sets by eliminating features or attributes which is of no use, or eliminating the redundant data from the data set. Feature Selection improves efficiency, and also accuracy of the classifier improved after applying feature selection algorithm. Feature Selection in supervised learning has main objective of finding a feature subset that enhances the classifier accuracy.

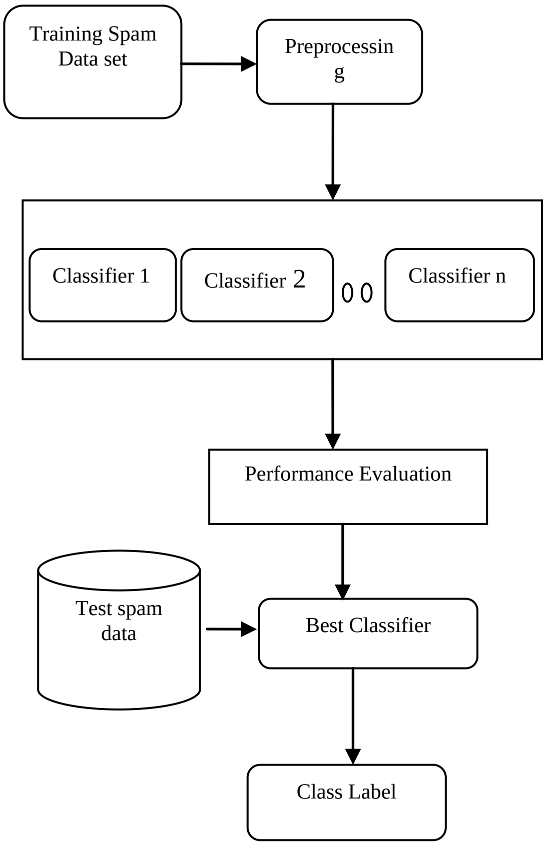


Figure. 5: Overall architecture of proposed system

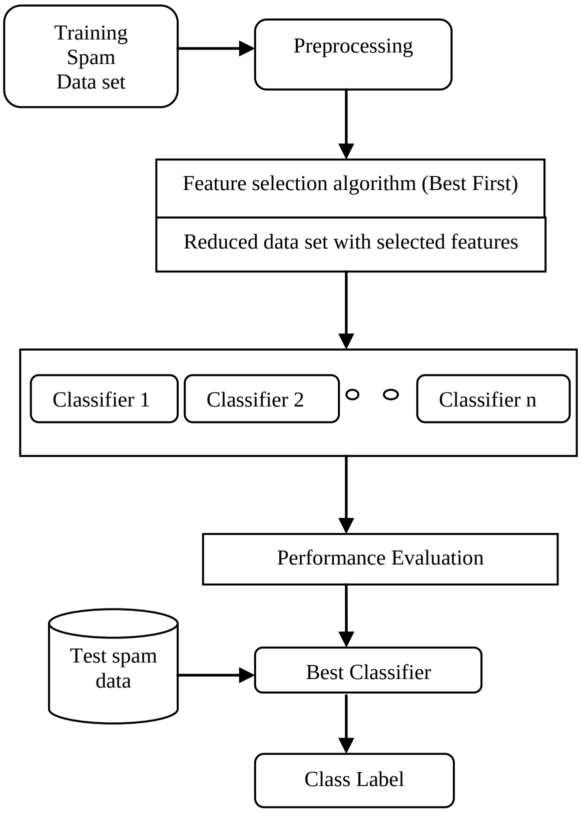


Figure. 6: architectural view of proposed system with feature selection

## V. EXPERIMENT AND ANALYSIS

### A. *Experiment I*

In order to validate the proposed scheme for spam mail detection, we conduct several experiments. The main objective is to find out the best classifier whose accuracy is better than the rest of the classifiers. The dataset which we use is Spambase dataset consisting of 57 attributes with one target attribute in discrete format. Following classification are applied one by one on the dataset: (1) Na ïve Bayes (2) Bayesian Net (3) Support Vector Machine (SVM) (4) Function Tree (FT) (5) J48 (6) Random Forest (7) Random Tree and (8) Simple Cart. And it is found form this study that out of all classifiers investigated on the given data set Random forest achieves highest accuracy that is 94.82%. Table III presented the result of entire classifiers in terms of accuracy, Kappa statistics (KS), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE). We computed all these listed statistics and prepare this comparative table from the result. After analyzing the data presented in table III , Random Forest is found to be the best classifier for spam mail classification with accuracy= 94.82%, then second highest accuracy is achieved by FT whose accuracy is 93.34% and so on. So from this study it is found that tree like classifier performs well in case of classification of spam mails.

TABLE 3: Result Set (without selecting the features)

| Algorithm | Accuracy (%) | KS | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|---|---|
| Naïve Bayes | 79.2871 | 0.5965 | 0.2066 | 0.4527 | 43.2668 | 92.64 |
| Bayes Net | 89.8066 | 0.7838 | 0.1026 | 0.2995 | 21.4795 | 61.08 |
| SVM | 90.4151 | 0.7954 | 0.0958 | 0.3096 | 20.0706 | 63.35 |
| FT | 93.34 | 0.861 | 0.0742 | 0.2468 | 15.5274 | 50.49 |
| J48 | 92.97 | 0.8528 | 0.0892 | 0.2562 | 18.6861 | 52.43 |
| Random Forest | 94.82 | 0.8908 | 0.0961 | 0.2064 | 20.1225 | 42.23 |
| Random Tree | 90.93 | 0.8108 | 0.0903 | 0.3001 | 18.9101 | 61.40 |
| Simple Cart | 92.43 | 0.8410 | 0.1055 | 0.2606 | 22.0913 | 53.33 |

### B. *Experiment II*

In this experiment we first applied Best-First Feature selection algorithm for selecting a subset of features from the given data set. Initially total 58 attributes was present in the given Spambase data set, but after applying Best-First algorithm on the given data total 15 attributes are selected. Then we apply Classifiers on this reduced data set for the detection of spam mail. Table IV gives the summary of the result. Best first filtering approach produce above 95% accurate results for four classifiers (FT, J48, Random Forest, Random Tree) and above 90% accurate results for two classifiers (Simple Cart, Bayes Net). And highest accuracy is achieved by Random Tree which is equal to 99.7175% and second highest is achieved by Random Forest which is equal to 99.54%.However it is quiet difficult to achieve 100% accuracy but these two classifiers (Random Tree and Random Forest) are very nearby to that.

TABLE 4: Result Set after filtering with Best-First

| Algorithm | Accuracy (%) | KS | MAE | RMSE | RAE | RRSE |
|---|---|---|---|---|---|---|
| Naïve Bayes | 78.9394 | 0.5935 | 0.2010 | 0.4235 | 42.0966 | 86.6669 |
| Bayes Net | 92.719 | 0.8456 | 0.0825 | 0.2444 | 17.2669 | 50.0103 |
| SVM | 86.54 | 0.7091 | 0.1345 | 0.3668 | 28.1717 | 75.0630 |
| FT | 95.54 | 0.9064 | 0.0550 | 0.1987 | 11.5181 | 40.6682 |
| J48 | 95.65 | 0.9083 | 0.0772 | 0.1965 | 16.1653 | 40.20 |
| Random Forest | 99.54 | 0.9904 | 0.0252 | 0.0827 | 5.271 | 16.9271 |
| Random Tree | 99.7175 | 0.9941 | 0.0040 | 0.045 | 0.8478 | 9.2079 |
| Simple Cart | 93.9361 | 0.8721 | 0.1070 | 0.2313 | 22.4038 | 47.33 |

## VI. CONCLUSION

In Spam mail classification is major area of concern these days as it helps in the detection of unwanted e-mails and threats. So now a day’s most of the researchers are working in this area in order to find out the best classifier for detecting the spam mails. So a filter is required with high accuracy to filter the unwanted mails or spam mails. In this paper we focussed on finding the best classifier for spam mail classification using Data Mining techniques. So we applied various classification algorithms on the given input data set and check the results. From this study we analyze that classifiers works well when we embed feature selection approach in the classification process that is the accuracy improved drastically when classifiers are applied on the reduced data set instead of the entire data set. The results gained were promising Accuracy of the classifier Random Tree is 99.715% with best-first feature selection algorithm and accuracy is 90.93% only when we don’t apply this subset selection algorithm. So, here in this study we achieve highest accuracy = 99.715%. As we all know that it is very difficult to achieve 100% accuracy but Random Tree and Random Forest (accuracy>99%) is very nearby to that. Therefore it is find that tree like classifiers works well in spam mail detection and accuracy improved incredibly when we first apply feature selection algorithm into the entire process.

## REFERENCES

[1] Nie N, Simpser A, Stepanikova I, and Zheng L.Ten years after the birth of Internet, how do Americans use the internet in their daily lives[R]. Technical report, Stanford University, 2004.

[2] Almeida T, Yamakami A, Almeida J. Evaluation of approaches for dimensionality reduction applied with NaïveBayes anti-spam filters [C]. In the Proceedings of the 8th IEEE International

conference on machine learning and applications, Miami, FL, USA,2009, 517-522.

[3] Vapnik V N. Statistical learning theory [M]. John W iley &Sons, NewYork, N Y, 1998.

[4] Ian H, Witten and Eibe Frank.Data Mining: Practical machine learning tools and techniques", 2nd Edition. San Fransisco: Morgan Kaufmann; 2005.

[5] Caruana R.A. and Freitag D. How useful is Relevance? Technical Report [A]. AAAI Symposium on Relevance, New Orleans, 1994.

[6] Blum A.L. and Langley P. Selection of Relevant Features and Examples in Machine Learning [C]. In International Symposium on Artificial Intelligence on Relevance, 1997, 245-271.

[7] Doak J. An Evaluation of Feature Selection Methods and their Application to computer Security [R]. Technical Report CSE-92-18, Davis, Ca: University of California, Department of computer Science, 1992.

[8] Liu H and Motoda H, and Dash M. A Monotonic Measure for Optimal Feature Selection [C]. In Proc. Of the European Conf. on Machine Learning, Springer Verlag, 1998, 101-106.

[9] Ducheneaut N and Bellotti V. E-mail as habitat: an exploration of embedded personal information management [A]. Interactions ACM, 2001, 8: 30-38.

[10] Carreras X, and Marquez L. Boosting trees for anti spam filtering [C]. In International conference on Recent Advances in Natural Language Processing. , 2001 160-167.

[11] Sahami M, Dumasi S, Heckerman D, and Horvitz E. A Bayesian approach to filtering junk e-mail: In Learning for text categorization [A]. Papers from the 1998 Workshop, Madison, Wisconsin, 1998.

[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] Ahmed K. An overview of content-based spam filtering techniques [A]. Informatica, 2007, 31(3): 269-277.

[14] Biro I, Szabo J, Benczur A, and Siklosi D. Linked Latent 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.

[15] Perkins A. The classification of search engine spam. http://www.ebrand management.com/white papers/spam classification, 2001.

[16] Paulo C, Clotilde L, Pedro S. Symniotic data mining for personalized spam filtering [C]. In the Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology, 2009, 149-156.

[17] Rasim M A, Ramiz M A, and Saadat A N. Classification of Textual E-mail spam using Data Mining Techniques [J]. In the Journal of Applied Computational Intelligence and Soft Computing, 2011.

[18] Erosheva E A and Fienberg S E. Bayesian mixed membership models for soft clustering and classification [J]. Proceedings of National Academy of Sciences, 2004, 97(22):11885-11892.

[19] Crawford E, Kay J, McCreath E. Automatic induction of rules for e-mail classification [C]. In 6th Australian Document Computing symposium, Coffs Harbour, Australia, 2001, 13-20.

[20] Spam Assassin. The Apache Spam Assassin Project. http://spamassassin.apache.org/.2006.

[21] Stern H. Fast Spam Assassin Score Learning tool http://search.cpan.org/src/PARKER/MailSpamAssassin 3.0.3/masses/README.perceptron,2004.

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

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

## Authors' Profiles

**Megha Rathi:** She is Assistant Professor (Grade II) at Jaypee Institute of Information Technology, India. She holds a Masters of Technology and a Bachelor of Engineering degree in Computer Science and Engineering. Currently she is pursuing her PhD in Computer Science and Engineering. Her areas of interest are Database systems, Software Engineering, Software Testing and Artificial Intelligence.

**Vikas Pareek:** He is Associate Professor at Banasthali University, Rajasthan, India. He obtained his Doctorate in the area of Cryptography. He also holds a Bachelor of Engineering degree in Computer Science and Engineering. His areas of interest are Cryptography, Algorithms, Data Structures, and Electronic Commerce. He has many publications in international journals and conferences to his credit.