IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 75-79

A Study on Spam Mail Detection Techniques in data Mining

A.Deepa [1], C.Malathi [2]
[1](Department of computer science, VLB Janakiammal college of arts and science, India)
[2](Department of computer science, VLB Janakiammal college of arts and science, India)

Abstract : *Internet users spend their time on social networks than search engines and other websites. Social Media networks such as Face book, Twitter, YouTube entities set up social networking pages to enhance direct interactions with the online users. Social media networks mostly depend on users for content contribution and sharing. Information is spreaded across social networks quickly and effectively to the users. Virus from the spammers could lead to personal or business loss and damage in the content. However, at the same time social media networks become admitting of different types of unwanted users or malicious spammer or hacker actions. There is a decisive need in the society or industry for security solution in social media. This demonstration specifies a propose of Social Spam Guard, a scalable and online social media spam detection system based on data mining for social network that deals with security. The best algorithms used for the detection is Machine learning approach such as Naive Bayes classifier model and Decision Trees. The classifier model helps identified based on its accuracy to correctly classify spam and non-spam emails.*

Keywords : *Spam Mail, Data Mining, Social Network, Naive Bayes classifier model.*

I. INTRODUCTION

In recent years, internet has been created several platforms for making human life become more secure. Among these e-mail is a substantial platform for user communication. Email is nothing; simply it is called an electronic messaging framework which transmits the message from one user to another [1].We have entered the era of social media networks such as Facebook, Twitter, YouTube etc., Internet users spend more of their time on social networks. Information can be easily spreaded within the social media networks. Because of this, websites expose to various types of unwanted and malicious spammer or hacker actions. There is a need in the society and industry for a security solution in social media networks. Spam can be transmitted from any source like Web, Text messages, Fax etc., depending upon the mode of transmission spam can be categorized into various categories like email spam, web spam, text spam, social networking spam [2]. A company or brand page on social media also needs to be clean to reduce the risk of damaging its reputation. Virus links from the spams could lead to personal or business loss and damage in the content.Spam filtering is a challenging area for an mixture of reasons. For spam email, users are facing lot of problems like abuse of traffic, limit the storage space, computational power, become a barrier for finding the additional email, waste users time and also threat for user security [9, 10]. So, if the email should be more secure and effective, appropriate Email filtering is an essential process. Several types of researches have been performed on email filtering. According to researcher's overview, Email filtering is a process to sort email according to some criteria. As there are various methods exist for email filtering, among them, inbound and outbound filtering is well known. Inbound filtering is the process to read a message from internet address and outbound filtering is to read the message from the local user. Moreover, the most effective and useful email filtering is Spam filtering which performs through anti spam technique. As spammers are proactive natures and using dynamic spam structures which have been changing continuously for preventing the anti-spam procedures and thus making spam filtering is a challenging task. Spam filtering is a process to detect unsolicited massage and prevent from entering into user's inbox. There are lots of algorithms can be used in e-mail filtering in that Machine learning approach has been widely used They include Naïve Bayes, support vector machines, Neural Networks, K-nearest neighbour and so on.

II. BACKGROUND STUDY

This study presents an overview of Data Mining, the algorithms of data mining, feature selection and most of the terms.

2.1 Data Mining

Data Mining is basically the discovery of knowledge from large database. It is a technique that helps to find out new patterns in the large data sets. It is a mixture of various fields such as Artificial Intelligence, Machine Learning, statistics, and Database systems. The main objective of data mining approach is to retrieve information from a large data set and it transforms into an understandable manner for the future use. The data mining task is an automatic or semi-automatic analysis of large quantities of data to extract an interesting

patterns. Data Mining is the process of analyzing data from different perspective and summarizing it into useful information and it can be used to increase revenue, cut costs, for classification, prediction etc.

The relationships are divided into 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 obtained to identify market segments. 3) Associations: Data mining is applied to data set to find out the associations. 4) Sequential Patterns: Data is obtained to predict behavior patterns and trends.

Data mining is called as Knowledge discovery from databases (KDD), because data mining is necessary step in the process of knowledge discovery from the database. Data mining is a part of knowledge discovery as in [13]..Knowledge Discovery steps Data mining involves many different algorithms to achieve the desired tasks. All of the specified algorithms try to robust 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 is based on the purpose of the algorithm to fit a model to the data based on Preference, and all algorithms require some sort of approach for searching. Knowledge discovery is a combination of all these steps shown in figure.1.

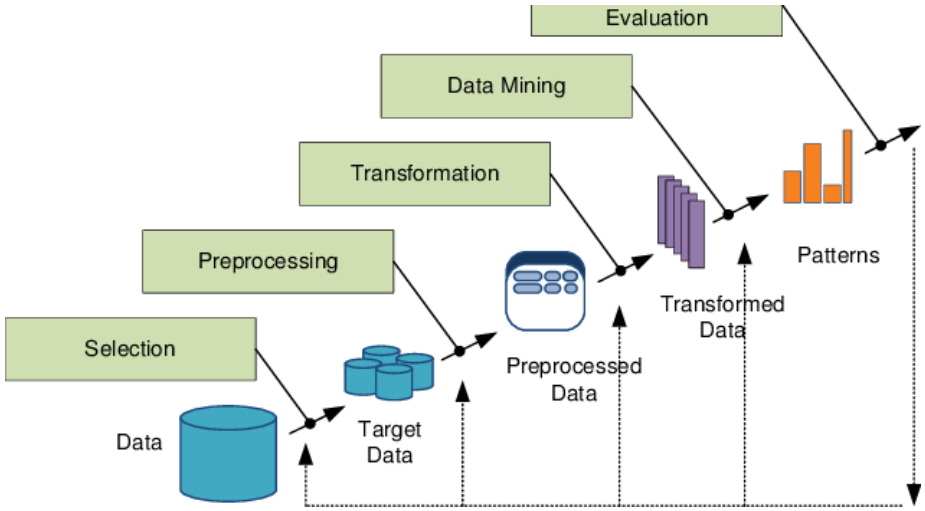


Figure.1.KDD process

2.2 Algorithms used in the study

2.2.1 Support Vector Machine

Support vector machines (SVM) are supervised learning models with associated learning models that helps to analyze data and that are mainly used for classification purpose. SVM takes a set of input data and output the prediction that data lies in one of the two categories such as it classifies the data into two possible classes. Given a set of training examples, each marked as fit in to one of the two classes .An SVM training algorithm builds a model that assign new data in one class or the other and it performs classification by constructing an N dimensional hyper plane that optimally categorize the data in two. SVM are set of connected supervised learning method worn for classification and regression [3]. SVM map input vector to a higher dimensional plane wherever a maximal separating hyper plane is constructed. Two parallel hyper planes constructs on each side of the hyper plane that separates the data. The unscrambling hyper plane is the hyper plane that maximizes the distance between the two hyper planes.

2.2.2 Naïve Bayes

A naïve Bayes classifier is a simple probabilistic classifier with strong independence supposition that is a naïve bayes classifier assumes that the presence or absence of a particular feature of a class is not related to the presence or absence of any other trait. It depends 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 requires a little amount of training data to estimate the parameters required for classification. Results show that these three machine learning algorithms gives better results without preprocessing among which Naïve Bayes algorithm is highly accurate than other algorithms as in [7].In Bayesian classification we have a hypothesis that specifies the given data belongs to a particular class then we have to calculate the probability for the hypothesis to be true. Bayesian classifiers are basically statistical i.e. it predicts the class association with probabilities such as the probability that a given sample data belongs to a particular class. The naïve Bayes method depends on Bayesian approach hence it is a easy, apparent and speedy classifier [4]. We have to first analyze some terms used in the theorem. P (X) is the probability that event X will occur. P (X/Y) is the probability that event X will occur given

that event Y has already occurred or we may define it as the conditional probability of X based on the condition that Y has already happened. Bayes theorem is defined in equation:

$$P(X/Y) = P(Y/X)\, P(X)\, P(Y) \quad (1)$$

If we consider X to be an object to be classified with the probabilities of belonging to one of the classes Z1,Z2,Z3 etc. by calculating P(Zi /A). Once these probabilities have been calculated for all the classes, we simply assign X to the class that has highest probability:

$$P(Zi/A) = [P(A/Zi)\, P(Zi)] / P(A) \quad (2)$$

Where P (Zi /A) is the probability of the object A belonging to a class Zi , P(A/Zi) is the probability of obtaining attribute values A if we know that it belongs to class Zi. P (Zi) is the probability of any object belonging to a class Zi without any other information, and P(A) is the probability of obtaining attribute values X whatever class the object belongs to. The modified Naïve Bayes showed the accurateness of 91% [6].

2.2.3 Decision Tree

A decision tree is a classification method that results a tree structure where every node denotes an experiment on attribute value and each twig represents an outcome of the experiment. The tree leaves represents the classes. Decision tree is model that represents both predictive and descriptive. The tree is based on 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 uses divide and conquer process to split the trouble search space into subsets. Decision tree is build to model the classification method. Once the decision tree is built it is then applied to each tuple in the database and that results in a classification for that tuple. There are two basic steps in this technique first to build the tree and second is to apply the tree to the dataset. The decision tree divides the search space into rectangular regions. A tuple is classified based on the state into which it fall.

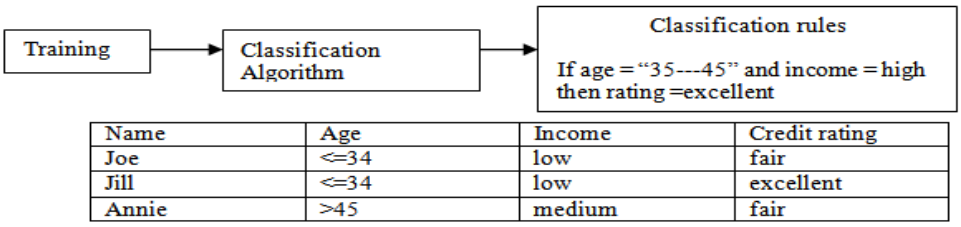
Given a database D = {t1, t2,….., tn} where ti = { ti1 ,……, tih} and the database schema contains the following attributes {A1, A2, ……, Ah}. Also given is a set of classes C = {C1,……, Cm }.
A Decision tree is a tree associated with D that has the following characteristics:
1) Each internal node is label with an attribute Ai .
2) Each edge in this is label with a predicate that can be applied to the trait associated with the parent.
3) Each leaf node is label with a class Cj .

2.2.4 Feature Selection

Feature Selection is also known as feature cutback. Attribute selection is the technique of selecting a subset of related features for building the learning models. Feature selection is an significant step in analyzing the figures, by removing irrelevant and redundant features from the data. Feature selection improves the presentation 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.
A quality selection algorithm is a computational answer which is forced by certain rules of significance. An immaterial feature is not useful for induction, but it also not essential that all relevant features are used for induction [5].

III. CLASSIFICATION AND PREDICTION

Classification is the separation of objects into classes. If the classes are formed with no looking at the data then that classification is known as apriori classification. If classes are formed by looking at the data then the classification method is known as posterior classification. When a fresh thing is introduced to the qualified system it is able to allot the object to one of the obtainable classes. This approach is known as supervised learning. Data Classification is a two step process as shown in figure. 2. In the first step, the model is built relating a predestined set of data classes. The form is constructed by analyzing database tuples described by the attributes. Each tuple is unspecified to belong to one of the existing class, as resolute by the class tag attribute. The data tuples analyzed to construct the form collectively from the training set.



Name	Age	Income	Credit rating
Joe	<=34	low	fair
Jill	<=34	low	excellent
Annie	>45	medium	fair

Figure.2: Learning and Training of classifier

IV. RELATED AND PROPOSED WORK

4.1 Related work

In study [11] 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. In study [14] 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 judgment .In study [15] 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 study [12] 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 [8] 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.

4.2 Proposed work

In this study spam mails are detected using various classifiers. The whole research comprised of two parts. First we will apply various classifiers for spam mail detection classification and check the results in provisions of correctness for each classifier. We use the complete data set and apply algorithm one by one devoid of selecting any feature. Secondly we detect spam mails by not using the complete data set instead we apply feature selection algorithm first. The algorithm which we are applying is Best-First Feature Selection algorithm. Then with selected features we apply all the classifiers one by one to check the results. It is found that classifier's accuracy is 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 (4) Random Tree and (5) Simple Cart. We find out accuracy in Mean Absolute Error ,Root Mean Squared Error etc., for all the classifiers and compare the results based on all these statistics.

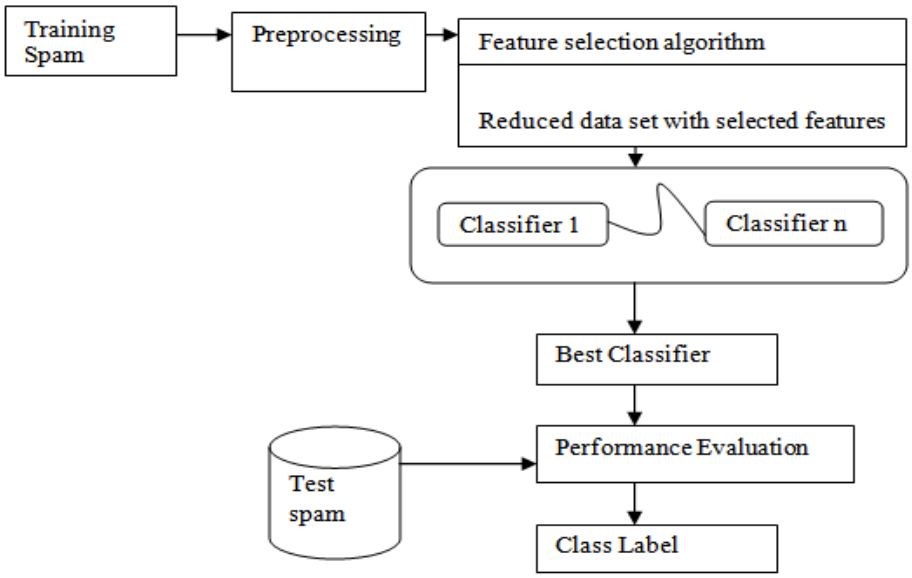


Figure.3: The architectural view of the proposed system with feature selection

V. CONCLUSION

Spam mail classification method is a major area of concern nowadays. It helps in the detection of useless e-mails and threats. Now a day's most of the researchers are working in this area to find out the best classifier model for detecting the spam mail. For that a filter is required with high precision to filter the unnecessary mails. This paper we focuses on finding the best classifier for spam mail classification using Data Mining techniques. So we have applied various classification algorithms with equations on the given input data set to check the exact results. From this learning we analyze that classifiers works well and if we embed feature selection approach in the classification process it provides the accuracy of results and it improved significantly when classifiers are applied on the condensed data set instead of the complete data set. The results gained were capable precision of the classifier Tree is 99.71% with best-first feature selection algorithm and accuracy is 90.93%. Therefore it is found that tree like classifiers works well in spam mail detection and accuracy enhanced extremely when we first apply feature selection algorithm into the entire process.

ACKNOWLEDGEMENTS

The view and conclusion enclosed in this document are those of the authors and should not be interpret as in place of social policies, either uttered or obscure, of the sponsors. We are very greatful for the owners of the publicly accessible photos used in this paper.

REFERENCES

[1]. Awad, W. A., &ELseuofi, S. M. (2011). MachineLearning methods for E-mail classification.International Journal of Computer Applications,16(1).
[2]. L. Firte, C. Lemnaru, and R. Potolea,"Spam Detection Filter using KNN Algorithm and Resampling", in 6th International Conference on Intelligent Computer Communication and Processing -IEEE, pp.27-33, 2010.
[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.
[6]. D. K. Renuka, T. Hamsapriya, M. R. Chakkaravarthi and P. L. Surya, "Spam Classification Based on Supervised Learning Using Machine Learning Techniques", in 2011 International Conference on Process Automation, Control and Computing -IEEE, pp. 1–7, 2011.
[7]. A. Harisinghaney, A. Dixit, S. Gupta, and Anuja Arora, "Text and image based spam email classification using KNN, Naïve Bayes and Reverse DBSCAN Algorithm", in proc. IEEE-International Conference on Reliability, Optimization and Information Technology (ICROIT), 2014, pp.153-155.
[8]. 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.
[9]. S., Heckerman, D., & Horvitz,E. (1998, July). A Bayesian approach to filtering junk e- mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).
[10]. Wang, Q., Guan, Y., & Wang, X. (2006). SVM-Based Spam Filter with Active and Online Learning.In TREC.
[11]. Perkins A. The classification of search engine spam. http://www.ebrand management.com/white papers/spam classification, 2001.
[12]. 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.
[13]. P. Verma and D. Kumar, "Association Rule Mining Algorithm's Variant Analysis," in International Journal of Computer Applicaation, Vol. 78, No. 14, pp. 26–34, 2013.
[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]. 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.