RESEARCH ARTICLE OPEN ACCESS

A Comparative study of Classification Techniques for Detecting Suspicious Criminal Activities on Emails

Hitesh Mahanand[1], Mr. Deepak Kumar Xaxa (Assistant Professor) [2]

Computer Science & Engineering, School of Engineering & IT
MATS University, Aarang, Raipur (C.G), India

Abstract:

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to predict future trends and behaviors, allowing system to make proactive, knowledge-driven decisions. One of the most popular data mining techniques is Classification. Classification is a data mining approach that relegates things in a gathering to target classifications or classes. The primary goal of classification is to accurately predict the target class for each one case in the data.Decision Trees are widely used in Classification. Decision Trees with enhanced feature selection may be applied to identify the Suspicious Criminal Activities, since Criminal Activities conducted via the internet are traditional crimes that are committed through the use of an electronic communication device. The main objective of this paper is to observe, analyze and compare the different Classification techniques.

Keywords:- **Classification techniques; Criminal Activities;Decision Trees;feature selection;Secuirity;Suspiciou.**

I. INTRODUCTION

Today, security is one of the major concerns for us and it is given top priority by the National Crime Records Bureau, India. But still it continues to experience terrorist and insurgent activities such as many recent incidents like bomb attack on an Israeli diplomatic vehicle in New Delhi on 13th February, 2012, bomb blast at New Delhi's High Court on 7th September, 2011 that killed 12 people, bomb attacks in crowded areas in Mumbai on 13th July, 2011, where three separate explosions killed many people. The crime graphof India through internet is constantly increasing.

E-media such as E-mail communication and social media can provide a space for many criminal activities, and reproduce and amplify criminal activities conducted offline. Online way of communication such as internet can increase their exposure to criminal activity, or opportunities to undertake their own, as well as publicize their activities to others. Preferably crime should be prevented; if it cannot be prevented then it should at least be detected. Criminology is one of the most important fields where data mining applications can produce important results. In crime analysis, we explore and detect the crimes and their relationships with the criminals [1].

With increasing number of crimes, there is a high volume of criminal database and also the complexity of these kinds of data have made criminology a suitable field for applying data mining techniques. Identifying the suspicious criminal activity is the first step for developing further analysis. The knowledge that is obtained from data mining approaches is very usefulwhich can facilitate and support police forces [2]. According to many researchers such as Nath[3], solving crimes is a difficult and time consuming task that requires human intelligence and experience and data mining is one technique that can help us with crime detection problems. When the criminals and their activities are identified and restricted properly, then it is possible to significantly reduce the crime rate.

Criminals these days are becoming technologically intelligent in planning and committing crimes [4]. E-mail has become one of the popular means of communication. The growing electronic data is also creating the need for automated analysis. Therefore,

we need such a crime detection and analysis approach to remain ahead of the criminals and catch them. The police should also use the current and new technologies [5] to give themselves the much-needed edge.

Email is the most popular way of communication of this era. It provides an easy and reliable method of communication. Email messages can be sent to an individual or groups. A single email can spread among millions of people within few moments. Nowadays, most individuals even cannot imagine the life without email. For those reasons, email has become a widely used medium for communication of terrorists as well. A great number of researchers [6],[7], focused in the area of counter terrorism after the disastrous events of 9/11 trying to predict terrorist plans from suspicious communication. This also motivated us to contribute in this area.

Classification is one of the classic data mining techniques, which is used to classify each item in a set of data into one of predefined set of classes or groups. The idea is to define the criteria use for the segmentation of the whole database, once this is done, individual dataset can then fall into one or more groups naturally. With the help of classification, existing dataset can easily be understood and it also helps to predict how new individual dataset will behave based on the classification criteria. Data mining creates classification models by observing already classified data and finding a predictive pattern among those data [8].

The paper is organized as follows: Section II presents the problem definition that describes the challenges to classifying the E-mails; Section III presents the different methodology for classifying the emails to detecting suspicious mails. Section IV will demonstrate the comparative study that will help us to analyze the performance of the classification. Finally paper is concluded with section V along with Conclusion.

II. PROBLEM DEFINITION

Email has been an efficient and popular communication mechanism as the number of Internet users increase. Therefore, email management is an important and growing problem for individuals and organizations because it is prone to misuse.

The main objective here is to identify e-mails that contain suspicious contents indication future criminal activities.For example, if domain specific suspicious keywords (kill, attack, bomb, etc.) are found in an email message, it is classified as suspicious whereas if non-suspicious indicators are present in an email, it is further classified as non-suspicious or may-be-suspicious.

The following specific objectives are developed for this paper:

- ✓ To evaluate the potential of data mining techniques in crime detection.
- ✓ To explore the data mining methods that supports decision tree technique to experiment with crime records.
- ✓ To come up with a method that is capable to detect the suspicious criminal activities.
- ✓ To interpret and analyze the performance of the different approaches.

III. METHODOLOGY

1. E-Mail Preprocessing

The preprocessing technique that is used to reduce the complexity of the E-mails and make them easier to handle, the E-mails have to be converted from the full text version to a document vector. The categorization of E-mail contents has many similarities with themore general field of text categorization.

The steps include (shown in Figure 1):

- ✓ Deciding which information to use (feature extraction)
- ✓ Representation of the text (weighting of features)
- ✓ Removal of nonessential information (feature reduction)
- ✓ Identification of important information (feature selection)
- ✓ Constructing the classifier

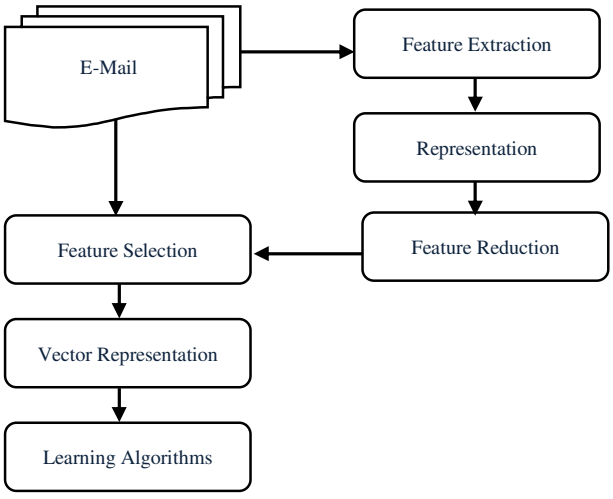

Figure 1 Email Pre-Processing

A. Feature Extraction

Avery most important concern when dealing with such abstract data as E-mail messages, is what features of the email should be chosen as its most descriptive features. The features of an E-mail will typically mean the words that it contains. Humans typically write E-mail using any haphazard combination of words that are on the top of their head; computers, on the other hand, can only derive meaning from more structure derangements of known symbols. There is a clear gap here, and it must be bridge difwe are going to get the computer to intelligently classifying the E-mail messages. To confound us further, some features of a message will be more fruitfultous than others; there will bealot of redundant data that will not assist us with learning, which should therefore not included in the machine learning process.

The procedure of preprocessing wipe out however much as could be expected the language dependent factors, tokenization, stop words elimination, and stemming [9].Proper selection of features can let a machine learning algorithm performatits best; improper selection can decreeitto imprecision and sluggishness.

Feature Extraction is the first step of preprocessing which is utilized to change over the E-mails into clear word form. Pre-processing is preferably based on a statistical process which is the dismissal of a large number of keywords. Some features will be very useful in distinguishing the type of email that the machine is dealing with, and to which folder it belongs. Other features will be almost completely useless.

But which features to include? First we must consider which features are possibly available to us. At first glance, the number of features available in an email message seems trivial. Almost all emails have a To:, From: and Subject: line in the header part of the message. All messages have a body section for the actual content. There are many other fields and features to consider, even though they may not be useful from a machine learning perspective. We must be aware of these features so that we can safely ignore them. What follows is a shortlist of some of the more common features; not ethatit cannot hope to be complete, for each new mail client or server can introduce many new types Of header information, special encoding, or special formatting of messages. This is of course another confounding factor. The extraction of these three header features is usually done by placing the words encountered the rein to the same ‘bag of words ‘as the rest of the body .It might therefore be beneficial to study whether keeping them separate from he rest of the data gives us an improvement in accuracy.

B. Representation

There presentation used by the mail filtering system is also a very important are a worth looking at. In this context, the "representation" used by a system means how the features are collected and stored within the agent’s memory. Once the features of a particular E-mail are collected in to the chosen representation, they can beefed into whatever machine learning algorithm is to be applied.Therefore, choosing a good way of representing the features is essential. If they are in a format that makes no sense to the Machine Learning algorithm, it is unlikely that the filtering will work.

Thebag–of–words approach can be viewed as feature representation which is a statisticalmethod. This collects a basic frequency count of the words present in a mail message. The other end of the scale focuses much more on trying to get the semantics of the email, an dis highly related to the domain of Natural Language Processing. The aim of these approaches is to somehow extract the meaning behind words and sentences in a message, in hopes that his will prove a more effective way of discerning between email content. Some of the existing systems using the bag–of–words approach are discussed by Rennie [10]

C. Feature Reduction

With such a wide variety of features to choose from, it is clear that the size of the feature vector that is extractedcanget very-verylarge. If we are usinga simple bagof words approach, we have one entryinthe feature vector for eachwordpresent inany folder. This canbe inthe order of tens of thousands. Suchhighdimensional data poses a severe problem for the machine learningalgorithms used. Boone [11]

notes that manyalgorithms for machine learningare not able tocope withlarge feature vectors, insteadpreferring medium tolowdimensional data.

Thus it has beennecessaryinthe past development of intelligent mail filters, todevise ways of reducingthe dimensionalityof the feature vector, by reducingthe number of features extracted. There are a number of ways this has beendone, andthey focus on removingwords foundinthe bodyof the email.

The process knownas stemming reduces the number of unique words bygroupingsimilar words together. It does this by reducingeachwordtoits 'stem'; for example, take the words 'learning' 'learner' and 'learned': these all have a commontheme, andthey maptotheir commonstem word 'learn'. Stemmingis almost widelydone as a preprocessingstep for text categorizationalgorithms, andassists insimplifyingthe problem.

D. Feature Selection

Next important step inthe pre-processing of E-mail classification is a feature selectionprocedure to convert vector space to enhance the adaptability, effectiveness and exactness of a content classifier.The purpose of feature selection is to more carefully choose the features that will be used as input to the learning algorithm (as shown in Figure 1).

From every text domain most of domain has a lot of features, but for text classification task, most of these features are not relevant and favorable, and also classification accuracy has been reduced by some noise features[12]. Hence this benefit comes not only from the further reduced size of the feature vector but also improve the vector space.

2. Classification Techniques

There are various classification techniques for e-mails.

A. Decision Tree

The decision tree [13] restructures the manual categorization oftraining documents by discovering well-definednodes in a tree structure. In a decisiontree structure it consists of two types of nodes one is internal node another one is external node. Internal nodes correspond to attributes selected by decision tree algorithm for making decision at specific level of hierarchy. The branches coming out from these internal nodes are the values of that attribute. The attribute at the top level of hierarchy in the tree has more power of classifying the instances of different classes.

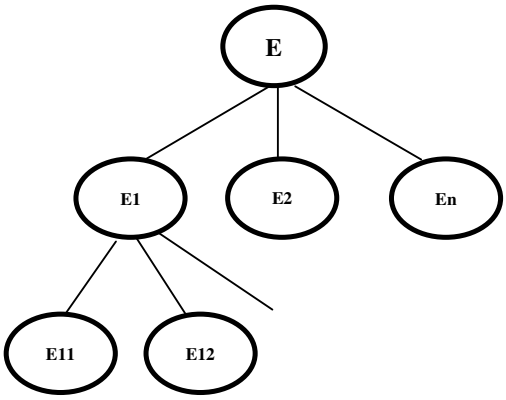

Figure 2 Decision Tree

The decision tree classification has several advantages over other classification are it is simple to build, it is an inductive algorithm, generated rules are easily understandable by humans,and provide a consolidated perspective of the classification logic, which is anadvantageous information of classification.

One of the major drawbacks of a decision tree is, with the occurrence of an alternative tree, it over-fits the training data that further categorizes the training data poorly but would classify the documents to be classified superior[14].

B. Decision Rules Classification

Decision rules classification methods uses the rule-based assumption to classify E-mail to their annotatedcategories [15].The algorithms composed a set of rules that describes the profile for each category. Rules are composed in the format of "IF condition THEN conclusion", where the condition section is filled by features of the category, and the conclusion section is represented with two options i.e.either the category's name or another rule to be tested. The rule set for a particular category is then composed by combining every different rule from the same category with the help of logical operator, typically use "and" and "or" operator. During the classification process, it is not necessary that each rule existing in the rule set needs to be satisfied. In the case of handling a dataset for each class with a large number of features, heuristics implementation is prescribed to reduce the size of rules set without influencing the performance of the classification.

During the feature extraction phase [15], it is more advantageous for classification task to construct local

dictionary for each individual category in the implementation of decision rules method. However, the drawback of the decision rule process is the inconveniences to assign a document toexclusively because of the rules from different rule sets are relevant to each other. Besides, the learning and updating of decision rule methods need extensive involvement of human experts to discover or update the rule sets. However when the number of distinguishing features are large, the decision rules method does not work properly similar to the decision trees classification method.

C. Naïve Bayes Algorithm

A simple probabilistic classifier named as Naïve Bayes classifier is based on applying Bayes' Theorem with strong independence assumptions.Naïve Bayes is a statistical analysis algorithm which works on numeric data [10].

An advantage of the naïve Bayes classifier is that it requiresa small amount of training data to estimate the parametersnecessary for classification. Naïve Bayes is a simple and fast classifier. It works well with statistical representations such as bag-of-words.

D. Support Vector Machine (SVM)

Support vector machines (SVMs) are selective classification methods which can be recognized as more accurate. The author [16] describes the axiom of SVM classification method based on the Structural Risk Minimization from computational learning theory. The principle is to find a hypothesis to assurance the lowest true error. SVM has some impressive features for it has been considered as the state-of-art in the text classification tasks. SVM has been used as different classification tasks such as text classification, hand written digit detection etc. Unique features of SVM are: it can work well in a very high dimensional feature space, it uses only a subset of original training to make decision boundary called support vectors and it is also suitable for non-linear separable data.

E. Artificial Neural Network (ANN)

A neural network machine learning algorithm is an attempt to mimic the way real neural networks in the human brain work. Each neural network has an input layer and an output layer, with one or more hidden layers between them. Depending on the input to each neuron, they will pass their output on to the next layer of neurons. Propagating values forwards thorough the network like this makes values eventually reach the output layer, where predictions can be made. ANN's are made from a largenumber of elements with an input fan order of magnitudeslargerthan in computational elements of traditional architectures [17]. These elements namely artificialneuron are interconnected into group using a mathematicalmodelfor information processing based on a connectionistapproach to computation in ANN. The neural networksmake their neuron sensitive to store item that can be used for distortion tolerant storing a large number of cases which are represented by high dimensional vectors.

For document classification tasks, different kinds of ANN approaches have beenintroduced.Due to simplicity in implementation some researchers used Single layerperceptron, which containsonly an input layer and an output layer [18]. Inputs are fed directlyto the outputs via weights in series. This approach istreatedas the simplest kind of feed-forward network.The multi-layer perceptron which is more sophisticated consists of an input layer, one or more hidden layers,and an output layer in its structure, also widely implemented for classification tasks [17].

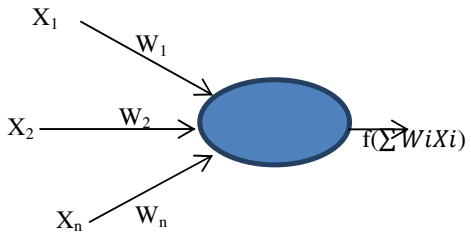

Figure 3Artificial Neural Network

The ability in handlingdocuments with high-dimensional features is the main advantage of the implementation of artificial neural network in classification tasks, and another advantage is handlingdocuments with noisy and contradictory data.In recent years, neural network has been applied in documentclassification systems to improve efficiency. Textclassification models using back-propagation neuralnetwork (BPNN) and modified back-propagation neuralnetwork (MBPNN) are proposed [18] fordocumentsclassification.

The ANN get its Inputs xi through pre-synaptic connections, Synaptic efficiency is modeled using real weights wi and the response of the neuron is a nonlinear function f of its weighted inputs.

The output from neuron j for pattern p is Opj where

$$O_{pj}(net_j) = \frac{1}{1+e^{-\lambda.netj}} \quad \text{(Eq.1)}$$

and

$$net_{tj} = bias * W_{bias} + \sum_k OpkWjk$$

Neural Network for document classification producesgood results in complex domains and is suitable for both discrete and continuous data(especially better for the continuous domain).

F. Genetic Algorithm based classification

Genetic Algorithms can identify and exploit regularities in the environment, and converges on solutions (can also be regarded as locating the local maxima) that were globally optimal [19]. This method is very effective and widely used to find-out optimal or near optimal solutions to a wide variety of problems. Genetic algorithm does not impose any limitations required by traditional methods such as gradient descent search, random search etc. The Genetic Algorithm technique has many advantages over traditional non- linear solution techniques. However, both of these techniques do not always achieve an optimal solution. However, GA provides near optimal solution easily in comparison to other methods.

The GA is very different from ―classical‖ optimization algorithms-It does the encoding of the parameters, not the parameters itself, the search is more elaborative in a given amount of time.As GA is probabilistic in nature, it may yields different solutions on different set of simulations‖. To get an optimal solution Monte Carlo methods can be adopted. The solutions space in multiple directions instead of in single direction.

Limitations:

Although because of its simplicity and classiness, Genetic Algorithm has proven themselves as efficient problem solving strategy. However, GA cannot be considered as universal remedy. Some limitations of GA are:

- ✓ The method chosen for representing any problem must be strong and firm, it must withstand random changes or otherwise we may not obtain the desired solution.
- ✓ In Genetic Algorithm, the Fitness function must be chosen very carefully. It should be able to evaluate correct fitness level for each set of values. If the fitness function is chosen poorly, then Genetic Algorithm may not be able to find an optimal solution to the problem, or may end up solving the wrong results.
- ✓ Genetic Algorithms uses random parameter selection, hence it will not work well when the population size is small and the rate of change is too high.
- ✓ In Genetic Algorithm solution is comparably better with, presently known solutions; it cannot make out ―the optimum solution‖ of its own.
- ✓ Sometimes over-fit of the fitness function abruptly decreases the size of population, and leads the algorithm to converge on the local optimum without examining the rest of the search space. This problem is also known as ―Premature Convergence‖.

IV. RESULT

The detection of criminal activities on E-mail needs text mining, machine learning and NLP techniques and methodologies to form and select pattern and knowledge from the E-mails. The main aim of this survey is to explore the existing literature, E-mail representation and classification techniques. Classification of E-mails is a crucialissue. Statistical of syntactic solution for the E-mail classification is given by most of the literature. However the classification techniques dependon the informational that we require. The execution performance of a classification algorithm in data mining is dependent upon quality of the data source.Insignificant and redundant features of data not only increasethe cost of mining process, but also affect the quality of the result in some cases [20]. Each algorithm has its own benefits and drawbacks as described insection III.

V. CONCLUSION

This paper provides a review of machine learning approaches and classification. An analysis of feature selection methods and classification algorithms were presented.E-Mail classification required more works and efforts are required toimprove the performance and accuracy of the process. New methods and solutions are required for usefulknowledge from the increasing volume of electronics documents.

The followings are the opportunities of the unstructured data classification and knowledge discovery:

- ✓ Exploration the feature selection improvement methodsfor better classification process.
- ✓ To reduce the training and testing time of classifier andimprove the classification accuracy, precision and recall.
- ✓ To detect suspicious criminal activities is a new active area of text mining.Automatic classification and analysis techniques are needed for detection ofsuspicious criminal activities on E-mails.
- ✓ Classification and clustering of semi-structured or unstructured E-mailshave some challenges and new opportunities.

- ✓ An implementation of sense-based text classificationprocedure is needed for recovering the senses from thewords used in a specific context.

.**REFERENCES:**

[1]. Hussain,Durairaj et al., "Criminal behavior analysis by using data mining techniques", International Conference on Advances in Engineering, Science and Management (ICAESM), IEEE, March 2012.

[2]. Keyvanpour, Javideh, et al., "Detecting and investigating crime by means of data mining: a general crime matching framework", World Conference on Information Technology, Elsvier B.V., 2010.

[3]. S. Nath, "Crime data mining, Advances and innovations in systems", K.Elleithy (ed.), Computing Sciences and Software Engineering, 2007.

[4]. S. Appavvu, MuthuPandian, et al., "Association Rule Mining for Suspicious Email Detection: A Data Mining Approach", Intelligence and Security Informatics, IEEE, 2007.

[5]. M. Mansourvar, et al., "A computer- Based System to Support Intelligent Forensic Study", Computational Intelligence, Modelling and Simulation (CIMSiM), Fourth International Conference, IEEE,September 2012.

[6]. S. Appavu, R. Rajaram, "Suspicious email detection via decision tree: a data mining approach," Journal of Computing and Information Technology -CIT 15 , 2007, pp. 161-169.

[7]. S. Appavu and R. Rajaram, "Learning to classify threatening e-mail", Internation Journal of Artificial Intelligence and Soft Computing. vol. 1, no. 1, 2008, pp. 39-51.

[8]. T. Abraham and de Vel, O., "Investigative profiling with computer forensic log data and association rules", Data Mining, Proceedings, IEEE International Conference, 2002.

[9]. Wang, Y., and Wang X.J., " A New Approach to feature selection in Text Classification", Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.

[10]. Jason D. M. Rennie. ifile: An Application of Machine Learning to E–Mail Filtering.

[11]. Gary Boone. Concept Features in Re:Agent, anIntelligent Email Agent. The Second International Conference on Autonomous Agents, 1998.

[12]. Jingnian Chen a,b,, Houkuan Huang a, ShengfengTiana, Youli Qua Feature selection for textclassification with Naïve Bayes" Expert Systems with Applications 36, pp.5432–5435, 2009.

[13]. J. R.Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, 1986, pp. 81-106, Kluwer Academic Publishers, Boston.

[14]. Russell Greiner, Jonathan Schaffer; AIxploratorium – Decision Trees, Department of Computing Science,University of Alberta,Edmonton,ABT6G2H1,Canada.2001.

[15]. URL:http://www.cs.ualberta.ca/~aixplore/ learning/ DecisionTrees.

[16]. ChidanandApte, Fred Damerau, Sholom M. Weiss.; "TowardsLanguage Independent Automated Learning ofText Categorization Models", In Proceedings of the 17thAnnual International ACM-SIGIR Conference on Researchand Development in Information Retrieval, pp. 23-30.1994.

[17]. Vladimir N. Vapnik, "The Nature of StatisticalLearning Theory" , Springer, NewYork. 1995.

[18]. Miguel E. Ruiz, PadminiSrinivasan; "Automatic Text Categorization Using Neural Network",In Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research, pp. 59-72. 1998.

[19]. Bo Yu, Zong-ben Xu, Cheng-huaLi ,"Latent semantic analysis for text categorization using neural network", Knowledge-Based Systems 21- pp. 900–904, 2008.

[20]. L. Zhang, J. Zhu, and T. Yao, ―An evaluation of statistical spam filtering techniques‖ ACM Transactions on Asian Language Information Processing (TALIP), v3, 2004, pp.243–269.

[21]. Wu W, Gao Q, Wang M "An efficient feature selection method for classification data mining" WSEAS Transactions on Information Science and Applications,3: pp 2034-2040. 2006.