ISSN:0975-9646

Analysis of The Fraud E-Mails using Truncated Gaussian Mixture Model and Ontological Application

Sreenivasulu. V
Department of Computer Science and Engineering
Gandhiji Institute of Science and Technology,
Krishna District - Andhra Pradesh - India

Dr. Satya Prasad. R
Department of Computer Science and Engineering
Acharya Nagarjuna University - Guntur,
Andhra Pradesh – India

ABSTRACT: The advancement in the area of communication has facilitated to store huge information. At the same time these developments have given equal opportunity for the construction and destruction. Among the destructive activities that arose of late or mostly based on IT related internet developments which enforced to expand the criminal activities across the globe using email as the media. Therefore, to uphold these evils, a methodology is proposed in this paper together with data mining techniques based on mixture model. The performance evaluation is carried out using evaluation metrics like False Acceptance Ratio (FAR) and False Rejection Ratio (FRR).

Keywords: Cyber Crime, E-Mail Forensics, Data Mining, Internet Technologies, Mixture Models

I. INTRODUCTION

The technological developments have benefitted the mankind with up gradation of tools for the betterment of life. These technologies like a double edged sword can be used for constructive and destructive purposes. People with destructive nature try to carry out the technological updates for the bad purposes. The computer tools that are available with IT sector are also not exceptional. These tools can be used even for the committing offenses by sitting at the remote houses, without realizing the false sense of anonymity. These activities deal with an aim of hacking the personal details, emails, accessing into others bank accounts to commit cyber crimes. These cyber crimes can be target based or can be used for a single event or a series of events. The crimes can even are targeted at individuals where the imposter tries to destruct the financial loses, copy write violation, sexual harassment etc. Using the IT and Internet technologies the crimes even be targeted towards the property where the pin numbers of credit cards transmitted or cracked and properties are stolen. The third category involves where the crime events are targeted of organizations. In this process, a Trojan horse program is pushed across the network and thereby breaking the confidentiality and creating damage to the programs.

Some of the cyber crimes are single event cyber crimes when a user opens an attachment, the file containing the virus may damage the PC. In the series of events attack the imposter tries to interact with the victim and tries to establish an illicit relation. Among these crimes targeted towards the organizations and the country are mostly concern with a lot of literature driven in this regard. These crimes are committed through the internet where the information and the type of attack to be undergone will be posted to the law breaker by the miscreants sitting at the remote ends. These crimes can also be projected where the confidentiality of the nation is transmitted to the terrorist organizations enabling to enter into the crucial sectors of the nation and create panic to the scenario. Among the crime attacks, email based attack is the most destructive one.

Hence a methodology is presented in this paper where the emails are scrutinized and thereby helping to identify the imposters identity. In this paper the truncated Gaussian Distribution is concerned for analyzing the emails. The main advantage by using this model is that the criminal who involves into the criminal activity through email has the tendency of executing a particular behavior which can be noted from the series of emails transmitted. This tendency will be known to interpret the emails where the miscreants' behavior can be showcased against high peaks. Therefore, the lower dimensional data can be discarded. To sort out such behaviors and patterns, a model is needed. Truncated Gaussian Mixture models are well suited. The rest of the paper is presented as follows.

In Section-2 of the paper the truncated Gaussian Mixture model is presented, Section-3 of the paper highlights about the data-set considered, Section-4 of the paper focuses the preprocessing and identifying the adjectives for effective fraud analysis using word-net. Section-5 demonstrates the experimentation together with results. Section-6 summarizes and concludes the whole work.

II. RELATED WORK

Having gone through the significant proposals made by some researchers are presented in this section. A methodology for E-Mail identification based on K-Nearest Neighbor method (K-NN) is used for the calculation of TFIDF. The E-Mails are considered and the TFIDF are considered for further process. Data Mining techniques such as classification and clustering are given robust results in email forensic investigation as explained in [1] [2] [3]. Methodologies based on K-NN for classification of new emails are broadly discussed in [4]. Combination of Naïve Bayes and K-NN Technique is presented on the TiMBL dataset to identify the email patterns based on the keyword-based spam filter [5] [6]. The algorithm presumes the classification of new emails by categorizing it as spam. It attains the features via a training set that has previously pre-classified suitably and then verifies the email. Collection of classifiers is considered for the effective analysis of spam filtering [7] [8] [9]. The researches have planned a tool that helps for studying the emails at ease to reveal the associated information that is essential for investigative trails during the forensic investigations. The disadvantages of the current forensic mechanism are overcome by the proposal tool by grouping inappropriate filtering. The authors have also highlighted concurrent methodologies by indexing the email contents so as to minimize the related search time. Combination of K-NN and Naïve Bayes classifiers considered and aimed at analyzing the emails and the researchers found that the models have good classification accuracy [10] [11]. Once a user presents a small query, in turn the recovery events fail to retrieve the required information exactly. The usage of semantic tools is demonsrated in [12]. In [13] authors presented a proposal using the concepts of Data Mining to reveal the underlying concepts from the email data.

III. Truncated Gaussian Mixture Model

In this methodology a truncated Gaussian Mixture Model is proposed. The main advantage of choosing this model is that it narrows down the dataset and helps to identify the possible imposter. The model is much more advantageous when compared to Gaussian Mixture Model where the limits are from – ∞ to + ∞. The probability density function (pdf) of the truncated Gaussian Mixture Model is given below

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{A} \int_{B}^{\infty} \left(e^{\frac{-(x-\mu)^2}{\sigma}} \right) dx$$

Where A and B are called truncating limits. In this method the values of A and B are chosen such that A is the minimum probability obtained from Term Frequency Independent Data Frequency (TFIDF) and B is the maximum TFIDF for evaluating in the emails under investigation. The term frequencies are evaluated and from which the minimal and maximum values of the frequency of terms indicated by A and B Values.

IV. Dataset

To present the methodology, the emails from the set of users are obtained from a bench mark dataset Enron. The dataset (https://www.cs.cmu.edu/~enron/) contains 5,17,431 E-Mails. Among these E-Mails available for testing purpose, we have considered around 128 E-Mails. This dataset contains the emails together with the receivers and senders information.

V. Preprocessing Of Data

The emails of the dataset are proposed such that the missing data are overcome using the concepts of Data Mining the tokenization process is carried out in order to overcome these spoofing attacks where the senders name and the receivers name can be spoofed. Therefore, in the preprocessing state the senders and receivers information is discarded. Also, in order to overcome the possibility of IP Spoofing, the IP address of the email delivered and received are also not controlled. To overcome the attacks, only the body of the email is considered. The wordnet tool is used for stemming the words, identifying the grammar comprising of articles, nouns and adjectives. These adjectives help to convey the inherent messages. Therefore adjectives play a vital role and are considered for the evaluation process. The semantic interpretation of each of these adjectives is underlined and the relevant semantics are pooled on groups. These groups are coded.

Every email corpus is analyzed and the term frequencies i.e., the number of times each term is occurring are identified. This is called TFIDF and the terms that are highest are ranked. These rankings are given as input to the truncated Gaussian Mixture Model in Section-2 of the paper and the reference templates are generated. This process is called training. In the testing phase, the process is repeated and the likelihood of the testing data against the templates is carried out basing on the maximizing of the likelihood estimate.

The K.L. Divergence and Chi-Square distribution can be used to identify the consistency of the data. In this paper the K.L Divergence is used and the evaluation formulae is given as

$$KL(A,B) = \int A(x) \log\left(\frac{A(x)}{B(x)}\right) dx$$

The Step wise Algorithm is presented for methodology given below.

Step 1: Considered the database and identify the emails for study

Step 2: Preprocess the emails to overcome the missing data if any

Step 3: Using the word net undergo the tokenization, elimination of stop words

Step 4: Identify the articles, prepositions, nouns and adjectives

Step 5: Process adjectives for semantic action and identify the synonyms

Step 6: Group the data and assign the ranking

Step 7: The TFIDF from the above step is given as input to the truncated Gaussian Mixture Model to evaluate the pdf

Step 8: Repeat the Steps 1 to 7 for every template

Step 9: Consider the test data and follow the steps 1 to 7 and obtain pdf which is to be compared with the pdf of Step8

Step 10: Using KL Divergence the likelihood estimate is obtained

Step 11: The maximum likelihood estimate (MLE) the email containing is identified

VI. Experimentation

In order to present the model, the email is considered from the Enron dataset and the processed email using the word net is presented in the following sample Email

Another reason is some companies are eager to copy Mr.Welch, Long viewed as one of the most successful managers in America. Defenders of these systems say any one who gets a low grade is likely to view the process as unfair. " 'A' Students love grades: 'F' students hate grades, "said John Sullivan, a human resources Professor at San Francisco State University. But the techniques, which some employees able with terms like" rank and yank," have come under sharp criticism. While they appear to offer an objective way to judge employees, they can be vulnerable to bias, Mr. Thomas said. Managers may stereo type employees when evaluating them on vague criteria like career potential= 01* deciding that older workers, for example, may have a harder time keeping up with new technology. In some cases managers can view these systems " as a tool to be used to weed out the ones you don't want," said Thomas .S. McLeod, a lawyer in Canton, Mich., who represents employees suing Ford in another case. According to the law suit, employees are rated on a five-point scale, with only a certain percentage permitted to receive each score. Employees doing the same job in the same unit are also given a "Stack ranking," from most to least valuable. Managers decide those rankings largely using what are called " like boat discussions," where they choose which employees they would want with them if stuck in a life boat. Managers had no other clear criteria, according to Christine webber, a lawyer at cohen, Milstein, Hausfeld & tool who is representing the employees. Grading is highly subjective at Micro Soft, according to Peter M. Browne, a former executive who is also suine the company, charging discrimination. Mr.Browne, who is blake, said managers were forced to use a curve in evaluating even small groups. He said he had to rate a group of five on a curve, for example, in deciding which ones would not receive stock options.

VII. Process

The Formulas for the evaluation of Acceptance Rate and False Acceptance Rate. The FAR is given below

$$\frac{[(\text{Dataset under consideration}) - (\text{Total No.of Relatedwords})]}{\text{Size of Total Dataset}}\ X\ 100$$

The False Acceptance Rate (FAR) is calculated as

$$\left[\frac{\text{Total number of related items}}{\text{Number of items}} X100\right]$$

Each E-Mail under study is to perform the steps presented in the algorithm and basing on the most likelihood, the source of the email having generated is to be identified. This methodology is having twofold advantages where in the test time can be optimized since the data is truncated and only the frequent repeated data is considered. It facilitates to understand the patterns and the emails more accurately since pdfs are estimated (S.K.Pal, N.R.Pal(1993)). The performance of the developed model is carried out using FAR and FRR and the evaluating table is presented below.

Trait	Algorithm	Acceptance Rate (%)	FAR (%)	FRR (%)
Email-Set 1	GMM	92.70	0.09	0.56
Email-Set 2	GMM	90.43	0.56	0.76

Table 1 : The Accuracy, FAR and FRR of Emails

VIII. Conclusion

In this paper highlights the contribution in the area of identifying the fraud Emails using Mixture Models bundled with Data Mining Concepts. The results obtained are compared with that of the model based on Gaussian Mixture Model against the evaluation metrics FAR, FRR and accuracy rate. The results exhibited are based on the dataset obtained from the bench mark email dataset Enron. The proposed methodology out performs the model based on Gaussian Mixture Model. This methodology can be well adopted for identifying the fraud emails.

References

[1] Nizamani S, Memon N, Wiil UK, Karampelas P. Modelling suspicious email detection using enhanced feature selection. Int J Model Optim 2012;2(4):371-377

[2] Nizamani S, Memon N, Wiil UK, Karampelas P. CCM: A text classification model by clustering, in: 2011 International Conference on advances in social networks analysis and mining (ASONAM) ,IEEE;2011.P.461-467.

[3] Nizamani S, Memon N, Wiil UK. Detection of illegitimate emails using boosting algorithm.

[4] Sakshi M, Shinnou H. Spam detection using text clustering. In:2005 International conference on cyber worlds, IEEE; 2005.p.4 pp-316-324.

[5] Appavu S, Pandian M, Rajaram R. Association rule mining for suspicious email detection: a data mining approach. In: Intelligence and security informatics. IEEE; 2007.p.316-323

[6] S.Appavu, M. Pandian, R. Rajaram Association rule mining for suspicious email detection: a data mining approach Intelligence and security informatics. IEEE; 2007.p.316-323

[7] Dharmija R, Tygar JD. The battle against phishing: dynamic security skins. In: Proceedings of the 2005 symposium on usable privacy and security, ACM; 2005.p. 77-88.

[8] Fetter I, Sadeh N, Tomasic A. Learning to detect phishing emails. In: Proceedings of the 16th international conference on World Wide Web, ACM; 2007.p.649-56.

[9] V.Chandra Sekhar and S. Sagar Imambi. "Classifying and Identifying of Threats in Email Using Data Mining Techniques", Proceedings of the International MultiConference of Engineers and Computer Scientist Vol.1,I IMECS,19-21 March 2008, Hong Kong.

[10] Sahami et al " A Basian Approach to Filtering Junk Email In Learning for Text Categorization" – papers fromf AAAI Workshop,pp,55-62.Madison Wisconsin AAAI Technical Report WS – 98 – 05, 1998.

[11] Chandrasekhar M, Narayanan K, Upadyaya S Phishi9g email detection based on structural properties. In: NYS cyber security conference;2006.p.1-7.

[12] McCallum A, Nigam K.A comparison of event models for Naïve Bayestect classification. IN: AAAI-98 workshop on learning for text categorization, vol.752;1998.p.41-48

[13] Joachims T. A statistical learning model of text classification for support vector machines. In: Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval, ACM;2001.p.128-136.