International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 24 Issue 15 – DECEMBER 2017.

Intelligent System for Classification and Selection of Cybercrime Incidents

Tejhan Bharadwaj R [#1]

[#] *Easwari Engineering College, Chennai, India*

***Abstract*— Due to the happening of malicious spam over the compromised data on online social network users, the necessity of earlier detection model is of high. To overcome this problem some system needs to detect the compromised data on online social network users in early stages, for this purpose there are different methodologies are used. This paper focuses on the intelligent system for classification and selection of cybercrime incidents. The crime cases are in the text form. Hence, the collected dataset is preprocessed using porter's stemmer algorithm which removes the stopwords, conjunctions etc in the document. Feature selection is the secondary objectives of our study. Here, ensemble based improved global feature selection is adopted that selects the optimal features using highest information gain score. The selected features are then processed into Fuzzy C-mean classification. The selected features are transformed into matrix with variant patterns. The computed matrix is further analyzed for the closeness of the data objects. Based on the closeness value, the data objects are moved onto their relevant classes. Experimental analysis is done in Facebook where majority of crime cases are not solved. Media sharing is mostly done using the facebook. Thus, the proposed scheme analyzed on the removal of irrelevant web pages. Performance metrics studied are the precision, recall and accuracy for the collected text. The results prove that the proposed scheme helps the crime detection system to easily classify the crime patterns with better classification accuracy.**

***Index Terms*— Cybercrime, Incidents, Media sharing, Fuzzy c-mean classification and classification accuracy.**

I. INTRODUCTION

Due to the recent developments made in the information technologies, the growth of web accessing users is inclined. It necessitates the security of the users from the offensive environment like cyberbullying, cyberstalking etc. In order to take precautionary steps, the technical details of each user are to be studied. Generally, the users registered themselves with the web systems using their sensitive information. Thus, the crime people aim is to track and crack that sensitive information and then precede their invasive action. The laws in several countries didn't punish the crime committed people e.g United Arab. The process of doing offensive activities with the use of any hardware device is known as the cybercrime. It generally breaks the rules and regulations in the computer networks. It is divided into categories, namely, violent and non-violent cyber crimes [1]. In the present time, most of the crime is non-violent in nature.

Computer forensics is the field of computer science that specifically deals with the cyber investigations. In relevant to, network forensics assists to discover the attacks and their behaviors using the log and status information. With the ubiquity and the enhanced growth of power technology has inclined the data collection and the storage computation. As the magnitude of the data increases, the data analysis technologies can't meet the user's specifications. This is being aided by the data mining system. The process of applying clustering, classification and prediction over the collected data helps to discover the hidden knowledge [2]. The main task of the security system is to prevent or predict the fraudulent actions. One of the main challenges in web crime is the analysis on big data involved in the criminal and terrorist activities. Thus, data mining technologies performs on variant web pages to find out the fraudulent activities [3].

The data for crime often presents an interesting dilemma. While some data is kept confidential, some becomes public information. Data about the prisoners can often be viewed in the county or sheriff's sites [4]. However, data about crimes related to narcotics or juvenile cases is usually more restricted. Similarly, the information about the sex offenders is made public to warn others in the area, but the identity of the victim is often prevented. Thus as a data miner, the analyst has to deal with all these public versus private data issues so that data mining modeling process does not infringe on these legal boundaries [5].

The paper is structured as follows: Section II depicts the related work; Section III depicts the proposed work; Section IV predicts the experimental results and analysis and concludes in Section V.

II. RELATED WORK

This section depicts the existing methods suggested for predicting the behavior of cybercrime. The author in [6] developed a tool which detects the crime -creating users. They collected the data and then preprocessed it to make an efficient tool. A hybrid DBSCAN algorithm with k-mean classification was designed for classifying the variant crime based on their behaviors. The author in [7] designed an OVER project that detected the software vulnerabilities. They created a database that composed of crime details, property details, offender and victims. For those databases, neural network schema is defined for making better decision systems. Relied upon the distance and time, the crime data is splitted and then coded for the particular area. Though, it targeted to decrease the victim user's action, yet failed to support the online fraudulent transactions.

The author in [8] studied about the importance of data mining in the digital forensics systems. They depicted the variants of digital forensics, viz, file forensic and the memory forensic modules. The file forensic module analyzes the files

and their storage drive whereas the memory forensic deals with the browsing history, running process, ports, and login details. Along with that network forensic module which deals over the Source IP, Destination IP, Source MAC, Destination MAC, Method, Protocol, Captured Time, Captured Length, Frame Type, Version and Destination Host is studied. In addition to, k-mean and apriori algorithms were studied for the attack discovery. It has failed to predict the location of the user. The author in [9] studied about the sequential mining algorithms using association rules. It mainly examined the fraudulent activities perceived in the emails. The designed algorithm depicts the content modification and recipient's address modification.

The author in [10] studied about the cyberbullying system using supervised based data mining techniques. They derived contextual features of the documents to make use of support vector machines. In addition to, they have used three features such as n-gram, weighting and words frequency model. They achieved better classification accuracy. The author in [11] studied about the application of binary and multiclass classifiers. They illustrated the topic based classification model for detecting the cyber bullying attackers. They concluded that taking into account such features will be more useful on social networking websites and can lead to a better modeling of the problem. The author in [12] discussed about the language based modeling for detecting the cyberbullying models. Each data is labelled as the web service model and then analyzed using Weka tool kit. Along these lines, C4.5 classifier was used as instance based learning classifiers.

The author in [13] studied about the gender based cyberbullying detection model using Myspace system. It analyzed using the user's information. The dataset consists of more than 3, 81,000 posts in about 16,000 threats. Overall 34%of posts are written by female and 64% by male authors. The Gender Specific Approach improved the Baseline by 39% in precision, 6% in recall, and 15% in F-measure. The author in [14] studied about the text mining model for detecting the offensive contents. They introduced Lexical Syntactical Feature (LSF) model to classify the contents. Based on the name-calling harassing systems, the lexical features are assigned as the offensive content. The author in [15] studied about the naïve bayes classification and the memory based classification systems to filter the unwanted spam messages. If any spam detected, the system automatically removes/ block the messages.

III. PROPOSED METHODOLOGY

This section depicts the working procedure of an intelligent system of cybercrime incidents. This study concentrates on developing a fuzzy c-mean classification model that elegantly classifies offensive content into their respective classes based on their patterns. In order to effectively classify the similar patterns, feature selection plays an important role in the data analysis systems. Thus, an Ensemble based Improved Global Feature Selection measure is analyzed. The proposed Fuzzy C-mean classification is explained as follows:

A. Preprocessing & Feature Selection:

Text preprocessing is the foremost step in our proposed system which helps to achieve the desired solution for defined problem. The obtained data is of incessant text which composes of stopwords. The removal of stopwords from the text helps to classify the content. Vector Space Model (VSM) is the preprocessing and selection techniques which are used to eliminate the stopwords, repeated patterns and missing data and to optimally select the features. The steps as follows:

1) Removal of Stopwords:

In order to reduce the data sparsity, the removal of proposition, conjunction and pronouns in the obtained text.

2) Applying the word stemmer:

This step ensures the removal of the repeated words to make the meaningful content. Porter's stemmer technique is used to transform the English words into the meaningful content based on the set of rules.

3) Indexing:

The preprocessed data is organized into documents based on the indexing model. In the general context, each document composes of set of words. These set of words are modulated into vector space. Each word in the document is referred in the term-frequency manner. It is given as:

$$D= (W_{wd}).$$

Where D is the document with the weight W_{wd}.

4) Selecting the features:

This step is to assure that the selected optimal features helps to achieve better computational time and the classification accuracy. Since time is the motivation metric, in our study, an ensemble based improved global feature selection technique is adopted. This technique resolves the multi-class problem. It is given as follows:

$$IG(f) = -\sum_{k=1}^{NC} P(C_k)\log P(C_k) + P(f)\sum_{k=1}^{NC} P(C_k|f)\log(C_k|f) + P(f')\sum_{k=1}^{NC} P(C_k|f')\log P(C_k|f')$$

Where, NC= Aggregate No. of classes

$P(C_k)$= Probability of the class C_k

P (f) = Probability of the selected feature f.

P (f`) = Probability of the non-selected features f.

$P(C_k|f)$ = Probability of the selected feature f for the class C_k.

$P(C_k|f')$ = Probability of the non-selected feature f for the class C_k.

The steps involved in the Improved Global feature selection are:

a) Assign the label for the features.
b) Compute odd ratio score for each feature.
c) Relied upon the highest score, assign new label to the local features.
d) For the newly label assigned features, compute the Information Gain (IG) score for the features.
e) Sort the features in descending order.
f) Highest IG score is considered as the global features and then placed in the new feature list.

B. Classification model:

Though, there is several unsupervised learning techniques, Fuzzy C-mean classification are considered as the most adopted technique for our study. The objects of the class are treated in the fuzzy membership degree values between 0 and 1. The steps are as follows:

2.1) Creation of the dataset:

The dataset, P patterns with F features are composed into M- dimensional column vectors and N observations as follows:

$$X = \begin{matrix} x_{11} & x_{12} & x_{1M} \\ x_{21} & x_{22} & x_{2M} \\ x_{N1} & x_{N2} & x_{NM} \end{matrix}$$

$$x_k = [x_{1k}, x_{2k}, \ldots, x_{Nk}]^T$$

$$X = \{x_{k|k=1,2,\ldots M}\}$$

2.2) Computing objective function:

This process is to minimize the objectivity of the created dataset. It is computed as follows:

$$J_m(X; U, V) = \sum_{t=1}^{c}\sum_{k=1}^{M}(U_{tk})^m \|x_k - v_t\|^2 A$$

Where U and V are the c-partition and center computation for the matrix X;

C is the no. of classes.

M is the weighting component with A weight matrix.

2.3) Closeness of the data objects:

To find the distance between the data objects, the closeness value is computed as follows:

$$D_{tkA}^2 = \|x_k - v_i\|_A^2 = (x_k - v_i)^T A (x_k - v_i)$$

2.4) Classifying the objects into their optimal class:

From the previous step, numerous classification criteria are achieved. To pick the relevant classes, fuzzy c-partition with least-square error is used.

$$H_c(U) = -\sum_{t=1}^{C}\sum_{k=1}^{M}(U_{tk}\log(U_{tk}))/M$$

$$F(U) = \sum_{t=1}^{c}\sum_{k=1}^{M}(\frac{U_{tk})^2}{M})$$

The design pattern class with the highest membership value is considered as candidate design pattern class for a given design problem. The highest class membership value depicts more closeness between the textual similarities for problem definition part of design patterns (i.e. Members) of corresponding candidate class (i.e. Clusters) and the description of the design problem.

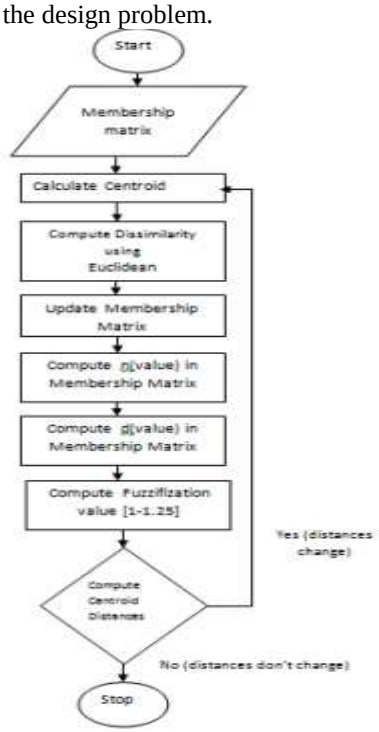

Fig.1 Working of the improved fuzzy c-mean classification

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section depicts the experimental analysis of our proposed study. Initially, the crime data is collected from well-renowned social network, Facebook. Nowadays, most of the users hold the account in facebook. It is reported in survey that 83% users are real, 5% users are partial name, and 2.35% are in fake name. In specific to majority of the cases are not sharing their personal information. Privacy setting is not correctly utilized by most of the cases. Thus, higher rate of security risks are faced by this social network. The crimes in facebook are:

a) Scams: By clicking inappropriate hyperlinks by the users, the scammers track the sensitive information.
b) Cyberbullying: The crime person targets the adults.
c) Stalking: By sending the irrelevant messages, the user's information is gathered.
d) Robbery: With the help of Google Map, the user's location is predicted and robs the properties.
e) Identity theft: By creating multiple fake accounts, the sensitive information of a user is hacked.

The below fig.2 dictates the statistical survey of facebook users. Sample of 170 cases are used for our study. From that, it's been finding that 90% of the users having the facebook account and 10% of users are not having facebook accounts.

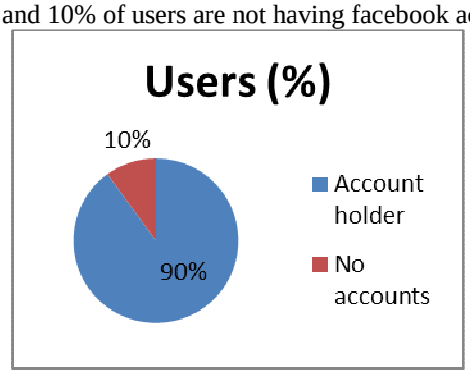

Fig.2. Facebook users

Similarly, the fig.3 presents the reason behind the usage of facebook.

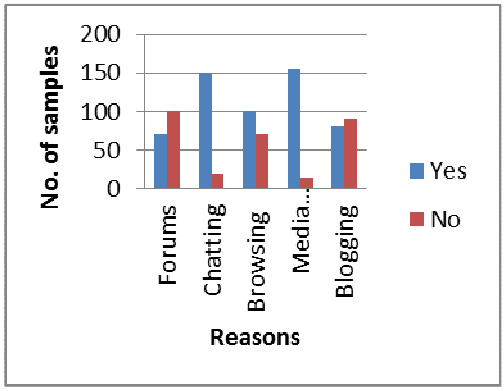

Fig.3 Reasons for using facebook

It is inferred from the above figure, chatting and media sharing are the most common purpose of using facebook. Let us consider a sample case,

Case no.111 Bala Kumar is sentenced to jail for 14 years in the palaiyankotai prison, for the violent crime of a person using knife.

Using our proposed preprocessing method, the stopwords are removed from the above text and its classified using fuzzy c-mean scheme. To take care of the different features for different crimes types, we introduced the concept of weighing

the attributes. This allows placing different weights on different attributes dynamically based on the crime types being classified. This also allows us to weigh the categorical attributes unlike just the numerical attributes that can be easily scaled for weighting them. Using the integral weights, the categorical attributes can be replicated as redundant columns to increase the effective weight of that variable or feature. However, we have introduced this weighting technique here in light of our unsupervised learning system. Performance metrics studied are the

a) Precision: Precision is defined as the clicking ratio of relevant webpage and the retrieved webpage to the aggregate ratio value of the retrieved webpages.

$$Precision = \frac{|relevant\ webpage| \cap |retrieved\ webpage|}{|retrieved\ webpages|}$$

b) Recall: Recall is defined as the ratio of relevant webpage and the retrieved webpage to the aggregate ratio value of the relevant webpages.

$$Recall = \frac{|relevant\ webpages| \cap |retrieved\ webpages|}{|relevant\ webpages|}$$

c) Accuracy: Accuracy is defined as the harmonic mean to depict the efficiency of the classification.

$$Accuracy = 2.\frac{precision.recall}{precision + recall}$$

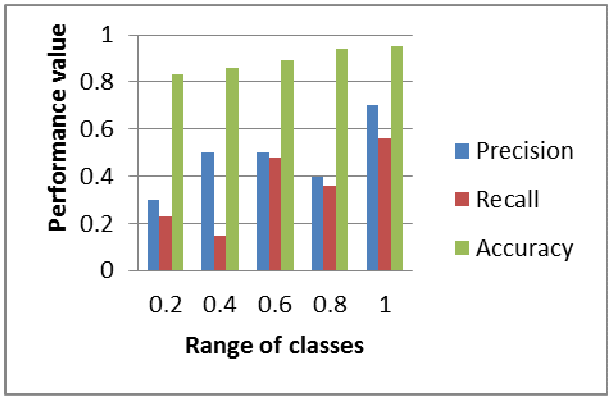

Fig. 4 Performance graph

V. CONCLUSION

Data mining can be used to model crime detection problems. Crimes are a social nuisance and cost our society in several ways. Any research that can help in solving crimes faster will pay for itself. About 10% of the criminals commit about 50% of the crimes. This paper concentrates on developing an improved Fuzzy C mean classification that elegantly classifies the crime pattern. Based on the weighting component of the attributes, the data objects are classified. The proposed scheme is experimented on the well-renowned social network, Facebook. The textual data are collected and preprocessed using stemmer algorithm. The structured data is used for selecting the optimal features using improved global feature selection scheme. Each feature is transformed in the vector space. Then, the global features are selected from the score of information gain. The features that contain highest IG score are taken as the classification input. Based on the closeness value of the selected features, the data objects are classified. Performance metrics studied are the precision, recall and accuracy. Since media sharing is the major task of the facebook users. By clicking the inappropriate links, the scammers gets the sensitive information of a users. Thus, the analysis is processed out in the clicking of relevant and irrelevant web pages. It has been verified from the results that our proposed scheme ensures better classification accuracy.

REFERENCES

[1] George Tsakalidis et al, “A Systematic Approach Toward Description and Classification of Cybercrime Incidents”, IEEE transactions on systems, man, and cybernetics: systems, 2017.
[2] Singhal, Amit,"Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering24 (4): 35– 43, 2001.
[3] B. Croft, D. Metzler, and T. Strohman, “Search Engines: Information Retrieval in Practice”. Addison Wesley, 2009.
[4] Gwizdka J, Chignell M. “Towards information retrieval measures for evaluation of web search engines”. Unpublished manuscript (1999).
[5] Ricardo Baeza-Yates, "Applications of Web Query Mining"3408. Springer Berlin / Heidelberg. pp. 7–22., 2005.
[6] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, Tefko Saracevic, "Searching the web: The public and their queries". Journal of the American Society for Information Science and Technology52 (3): 226–234, 2001
[7] Shital C. Patil*, 2 R. R. Keole, "Improving Search Result Delivery using Content and Usage Mining", International Journal of Applied Research and Studies (iJARS), Volume 3, Issue 3, 2014.
[8] "Introduction to Information Retrieval - Evaluation". Stanford University. 21 April 2013. Retrieved 23 March 2014.
[9] Mihajlovic, V., Djoerd Hiemstra, Henk Ernst Blok, and Peter MG Apers. "Exploiting Query Structure and Document Structure to Improve Document Retrieval Effectiveness.", 2006.
[10] Liu, Tie-Yan. "Learning to rank for information retrieval." Foundations and Trends in Information Retrieval 3, no. 3,pp. 225-331, 2009.
[11] D. R. Radev, H. Qi, H. Wu, W. Fan, "Evaluating webbased question answering systems". Proceedings of LREC, 2002.
[12] Nallapati, Ramesh, and Chirag Shah. “Evaluating the quality of query refinement suggestions in information retrieval”. Massachusetts Univ Amherst CenterFor Intelligent Information Retrieval, 2006.
[13] Robert M. Losee, “When Information Retrieval Measures Agree About the Relative Quality of Document Rankings”, Journal of the American Society for Information Science, 51(9), pp. 834-840, 2000.
[14] Alex Hartemink, “Clarifying various terms for evaluating classifier (or hypothesis testing) performance”, 2011.
[15] Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. "The balanced accuracy and its posterior distribution". Proceedings of the 20th International Conference on Pattern Recognition: 3121–24, 2010.