International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 1, Jan-Feb-2021, ISSN (Online): 2395-566X

Criminal Activities Predictive Analysis Using Data Mining Techniques

M.Tech. Scholar Meenu Rai, Associate Prof. Bhawana Pillai
Dept. of Computer Science & Engg.
Lakshmi Narain College of Technology & Science,
Bhopal,M.P., ,India

***Abstract*- Data mining is the extraction of knowledge from large databases. One of the popular data mining techniques is Classification in which different objects are classified into different classes depending on the common properties among them. Machine learning are widely used in Classification. This paper proposes a data mining technique which applies an enhanced existing machine learning Algorithm to detect the suspicious criminal activities. An improved decision tree Algorithm with enhanced feature selection method and attribute- importance factor is applied to generate a better and faster Decision Tree. The objective is to detect the suspicious criminal activities and minimize them. This paper aims at highlighting the importance of data mining technology to design proactive application to detect the suspicious criminal activities.**

***Keywords*- Data Mining, Predictive Analysis, Criminal Activity, Machine Learning.**

I. INTRODUCTION

Although methods aimed at predicting crime have been used by police officers for a long time, it is only recently that information technology has moved these attempts from simple heuristic methods (initially based on local knowledge of police officers) to sophisticated mathe matical algorithms using a large amount of information. Criminals exploit these opportunities to perform illicit activities.Drug dealers use chat rooms for drug trafficking. Terrorists and hate groups use social media to promote their ideology. Systems have an archive feature that stores all conversations for later reference. If the investigator gets access to the archived conversations in confiscated computers or in public chat servers, those conversations can be helpful in crime investigation.

The content of online communication can reveal the life styles of the participants, their social networks, activities, and preferences. However, manually analysing a large volume of chat conversations to find evidence related to a criminal case is very tedious and time-consuming. Most investigators use traditional search methods in forensic software tools or in desktop search engines to crawl and extract relevant data. This method has three noteworthy limitations. Existing search tools may find documents and sentences related to the search terms, but they do not provide information related to a suspect's social networks and activities.

The simple string-matching method is not suitable for investigation of crime cases because drug dealers seldom use the terms''drug'' and ''cocaine'' in their conversation. The experience and prior knowledge of the person perfo rming the search are crucial for good quality search results; otherwise, a search can often result in irrelevant, incomplete, or inconsistent information. To address the limitations of existing forensic search tools. We study propose a framework that extracts communities from a given chat log and uses agglomerative clustering to find and summarize the topics of interest in the identified communities.

Crime investigators can perform a search query and see the results in the designed visualizer. The purpose of this data mining framework is to collect instinctive and interpretable evidence from a chat log to simplify and facilitate the investigation process, especially in the initial stage when there are not enough indications for the investigator to start with. An overview of the proposed framework is presented.

There are three main modules of the framework: Clique Detector, Concept Miner, and Information Visualizer. The Clique Detector identifies the cliques (communities) in the chat log. For this purpose, it first recognizes all the mentioned entities in the given chat log. In the context of investigation, the term entity can be the name of a person, an organization, a phone number, or a physical address. For simplicity, we assume that the entity refers to the person's name.After extracting entities, the Clique Detector uses the co-occurrence frequencies of the entities in chat sessions and identifies the communities, called cliques.

II. RELATED WORK

Dynamics behind a criminal's relationship can be very helpful in identifying suspects and understanding the activities and interests of criminals.For any criminal event,

© 2021 IJSRET

crime investigators must consider and retrieve criminal network information.The investigation related to social networks can reveal hidden subnetworks, actors, their roles, and communication pattern. The investigation team can use the information of evidential value from a chat log to recreate events. It can be helpful in identifying infor mation such as the physical location of suspects, identity, transactional information, and information about the victim

S.Alami et al.[1] to target and detect suspicious topics and eventual illegal events. Similarity approach is used in text analysis to detect suspicious posts in microblog public ations. The evaluation of our proposed approach is done within real posts.

K.Biron et al.[2] The findings of this study demonstrate that the integration of data science and investigation activities is feasible and can extend the investigation process beyond the scope of the police profile structured database based on manual interrogation activities. This study ends by concluding the limitations and future research directions.

M.Nijim et al[3] The drone technology has also been misused as a platform by, nefarious individuals who use the UAV technology with a unique eves dropping payload to launch network intrusion into corporations by hovering their UAV over the corporate infrastructure. According to Federal Aviation Administration(FAA),commercial airline pilots have reported over a 100 close call situations or drones in visible proximity of the aircrafts at the airports.

E. B. B. Palad et al[4] This demonstrates how text mining can assist the PNP-ACG in analyzing online scam criminal data as it also highlights the importance of employing data mining tools in the legal and criminal investigation domains in the Philippines. Further work can be carried out in the future using different and a more inclusive cybercrime datasets and other classification techniques in Weka or any other data mining tool.

A.Prakash et al[5] The advancement of competent mech anism to detect suspicious transaction is a critical problem because of financial transaction made in global market. The system we are developing targets on how monetary institutions can obtain better results in anti-money laundering initiatives.To overcome this criticality we propose an efficient approach to find out if the transactions are legal or illegal by adopting Hash-based technique and Graph Theoretic Approach.

S. K. Dash et al[6] In this paper, we analyze Chicago city crime data fused with other social information sources using network analytic techniques to predict criminal activity for the next year. We observe that as we add more layers of data which represent different aspects of the society, the quality of prediction is improved. Our prediction models not just predict total number of crimes for the whole Chicago city, rather they predict number of crimes for all types of crimes and for different regions in City of Chicago.

P. Salunkhe et al[7] Decision Tree (DT) is also one of the technique which can help for file forensic investigation purpose. So, system can adopt a way by using Decision Tree for generating, storing and analyzing data retrieved from log files which pose as evidence in file forensic analysis. This paper focuses on how Decision Tree can allow system to quickly, easily and inexpensively analysis of log data available in various file formats for file forensic analysis.

III. PROPOSED METHODOLOGY

1. Predictive Analytics:
Predictive analytics is the technique of analyzing the past or historical data in order to predict the future outcome. It is different from data mining. As explained in the Figure 1 [5], predictive analysis starts by capturing relationships between the different variables in the data. After that, hypothesis is developed based on these results. Following this, based on the outcome of the previous steps, a model is built in order to test this hypothesis

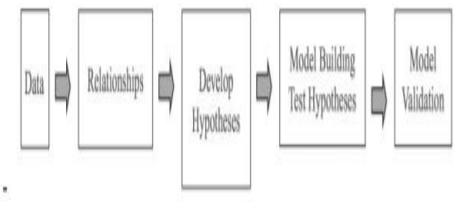

Fig 1. Data mining and predictive analytics.

There are numerous advantages of using Predictive Analytics in general. For example, an organization can study it's internal data to identify trends in profit, so that they can adopt the necessary steps to possibly replicate that in the future. It is also a useful technique for professionals in the marketing industry as it can help decide which campaign successfully generated revenue and business. For the purpose of this research, Predictive Analytics is helpful for the following reasons: It will help us identify the progression in crime throughout the years. It will help us closely observe the variables having highest correlation with the predictor or target variable.

Visualizing the data can even help map potential outliers, which can then be effectively handled during data pre processing. The analysis can bring up some interesting facts from the past which might prove to be useful to the while planning their patrol or strategies against crime

© 2021 IJSRET

2. Classification Techniques:
Classification techniques are used to segregate the data into one or more categories also known as class labels. The goal of classification is to create a certain set of rules that will either make a binary decision, or predict which of the multiple classes should the data be classified into.

Classification can mainly be divided into two types:
2.1 Binary Classification: In Binary Classification, the goal is to classify the elements into one of the two categories specified, say,X and X(not X). To determine the efficiency of the˜ Binary Classifier, we pass a set of inputs to the classifier and examine the output. There are 4 possible results-True Positive, True Negative, False Positive and False Negative [6]. Generally, 'Accuracy' is used to determine the efficiency of a binary classifier.

2.2 Multiclass Classification: Multiclass Classification in volves classifying the data into more than two categories. The most common types of Multiclass Classifiers are [6]: Pigeonhole Classifier: In a pigeonhole classifier every item is classified into only one of the many categories. Hence, for a given item, there can be only one output category assigned to it. Combination Classifier: This type of classifier can place an item into more than one output categories. Hence, unlike a pigeonhole classifier, this type of classifier does not assign a unique category to each input. · Fuzzy Classifier: These classifiers not only assign an input to more than one categories but also assign a degree to each category. This means that every input belongs to every category by a certain degree. Hence the output is an N-dimensional vector, where N is the number of categories.

Prediction using Clustering and Classification techniques The authors of have described a method to predict the type of crime which can occur based on the given location and time. Apart from using the data from the Portland Police Bureau (PPB), they have also included data such as ethnicity of the population, census data and so on, from other public sources to increase the accuracy of their results. Further, they have made sure that the data is balanced to avoid getting skewed results.

The machine learning techniques that are applied are Support Vector Machine(SVM),Random Forest, Gradient Boosting Machines, and Neural Networks. Before applying the machine learning techniques to predict the category of the crime, they have applied various preprocessing techniques such as data transformation, discretization, cleaning and reduction. Due to the large volume of data, the authors have sampled the data to less than 20,000 rows. They used two datasets to perform their experiments - one was with the demographic information used without alterations and in the second case, they used this data to predict the missing values in the original dataset. In the first case, ensemble techniques like as Random Forest or Gradient Boosting worked best, while in the second case, SVM and Neural Networks showed promising results. Since a smart city should give importance to the safety of their citizens, the authors of [2] have designed a strategy to construct a network of clusters which can assign police patrol duties, based on the informational entropy. The idea is to find patrol locations within the city, such that the entropy is maximized. The reason for the need to maximize the entropy is that the entropy in this case is mapped to the variation in the clusters, i.e. more entropy means more cluster coverage [2].

The dataset used for the research is the Los Angeles County GIS Data. The data has around different crime categories. Taking the help of a domain expert, the authors have assigned weights to these crimes based on the importance of the crime. Also, the geocode for each record is taken into consideration and the records that do not have a geocode are skipped. Because the authors in [2] are trying to maximize the entropy in this case, consider the equation (3) The probability p(c1) is defined as the ratio of weight of the centroid of the crime to the weight of the system, plus the ratio of the quickest path between two centroids, to the quickest path in the whole system.

3. Crime Data Mining Techniques:
Traditional data mining techniques such as association analysis, classification and prediction, cluster analysis, and outlier analysis identify patterns in structured data. Newer techniques identify patterns from both structured and unstructured data. As with other forms of data mining, crime data mining raises privacy concerns. Nevertheless, researchers have developed various automated data mining techniques for both local law enforcement and national security applications. Entity extraction identifies particular patterns from data such as text, images, or audio materials. It has been used to automatically identify persons, addres ses,vehicles, and personal characteristics from police narrative reports. In computer forensics, the extraction of software metrics—which includes the data structure, progr am flow, organization and quantity of comments, and use of variable names—can facilitate further investigation by, for example, grouping similar programs written by hackers and tracing their behavior.

Entity extraction provides basic information for crime analysis, but its performance depends greatly on the availa bility of extensive amounts of clean input data. Clustering techniques group data items into classes with similar characteristics to maximize or minimize intraclass similarity—for example, to identify suspects who conduct crimes in similar ways or distinguish among groups belonging to different gangs. These techniques do not have a set of predefined classes for assigning items. Some researchers use the statistics-based concept space algo rithm to automatically associate different objects such as persons, organizations, and vehicles in crime records.7

© 2021 IJSRET

Using link analysis techniques to identify similar transac tions, the Financial Crimes Enforcement Network AI System exploits Bank Secrecy Act data to support the detection and analysis of money laundering and other financial crimes. Clustering crime incidents can automate a major part of crime analysis but is limited by the high computational intensity typically required. Association rule mining discovers frequently occurring item sets in a database and presents the patterns as rules. T

his technique has been applied in network intrusion detection to derive association rules from users' interact ion history. Investigators also can apply this technique to network intruders' profiles to help detect potential future network attacks. Similar to association rule mining, sequential pattern mining finds frequently occurring sequences of items over a set of transactions that occurred at different times. In network intrusion detection, this approach can identify intrusion patterns among Deviation detection uses specific measures to study data that differs markedly from the rest of the data. Also called outlier detection, investigators can apply this technique to fraud detection, network intrusion detection, and other crime analyses.

However, such activities can sometimes appear to be normal, making it difficult to identify outliers. Classif ication finds common properties among different crime entities and organizes them into predefined classes. This technique has been used to identify the source of e-mail spamming based on the sender's linguistic patterns and structural features. Often used to predict crime trends, classification can reduce the time required to identify crime entities.

However, the technique requires a predefined classific ation scheme.Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy. String comparator techniques compare the textual fields in pairs of database records and compute the similarity between the records.These techniques can detect deceptive infor mation—such as name, address, and Social Security number—in criminal records.Investigators can use string comparators to analyze textual data, but the techniques often require intensive computation. Social network analysis describes the roles of and interactions among nodes in a conceptual network. Investigators can use this technique to construct a network that illustrates criminals' roles, the flow of tangible and intangible goods and information, and associations among these entities. Further analysis can reveal critical roles and subgroups and vulnerabilities inside the network. This approach enables visualization of criminal networks, but investigators still might not be able to discover the network's true leaders if they keep a low profile.

The type of crimes that the research project will investigate will be limited to armed robberies. However, the methodologies that are discussed in this project for identifying patterns in armed robbery records, are in general applicable to all other types of crimes, with minor (or no) modifications. The choice of armed robbery as a type of crime was made due to the fact that these kinds of crimes usually involve all issues that may be involved in other types of crimes. An armed robbery usually involves an offender, a victim or victims, and a weapon, and sometimes a car, in addition to the general characteristics of a crime. The same issues (objects), however, may not be pertinent to other crimes such as aggravated sexual abuse, as an example, where we may lack the use of a weapon or a get away vehicle. The research project is divided into three major phases, each of which will be discussed under a separate heading in this paper. the three different phases in the project design: (a) Pre-processing phase (b) Neural network phase (c) Heuristics system phase

VI. CONCLUSION

In this paper a series of state-of-the-art crime data analytics and visualization techniques were utilized to analyze crime data from three varies cities, which allowed us to identify patterns and obtain trends. By exploring the, a machine learning, we found that both the model and the algorithm perform better than data mining., in order to achieve the best prediction of trends in terms of.

Additional results explained earlier will provide new in sights into crime trends and will assist both police depart ments and law enforcement agencies in their decision making.Criminal activities predictive analysis using data mining techniques.

REFERENCE

[1] S.Alami and O.Elbeqqali, "Cybercrime profiling: Text mining techniques to detect and predict criminal activities in microblog posts," 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA), Rabat, 2015, pp.1-5, doi: 10.110 9/SITA.2015.7358435.

[2] K.Biron,W.Mansoor,S.Miniaoui,S.Atalla, H. Mukhtar and K. F. Bin Hashim, "Data Science Tools for Crime Investigation, Archival, and Analysis," 2019 IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI),Le icester, United Kingdom, 2019, pp.1263-1266, doi:10 .1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI. 2019.00235.

© 2021 IJSRET

[3] M.Nijim and N.Mantrawadi,"Drone classification and identification system by phenome analysis using data mining techniques,"2016 IEEE Symposium on Tech nologies for Homeland Security (HST), Waltham, MA, 2016, pp. 1-5, doi: 10.1109/THS.2016.7568949.
[4] E.B.B.Palad, M.S.Tangkeko, L. A. K. Magpantay and G.L.Sipin,"Document Classification of Filipino On line Scam Incident Text using Data Mining Tech niques,"2019 19th International Symposium on Communications and Information Technologies (IS CIT), Ho Chi Minh City, Vietnam, 2019, pp. 232-237, doi: 10.1109/ISCIT.2019.8905242.
[5] Prakash,A.S., A.K.H., K.T.P. and P.K.K.N., "Proposal of Expert System to Predict Financial Frauds Using Data Mining," 2019 3rd International Conference on Computing Methodologies and Communication (ICC MC),Erode,India,2019,pp.10801083,doi:10.1109/ICC MC.2019.8819709.
[6] S.K.Dash, I.Safro and R. S. Srinivasamurthy, "Spatio-temporal prediction of crimes using network analytic approach," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 1912-1917, doi: 10.1109/BigData.2018.8622041.
[7] P.Salunkhe, S.Bharne and P. Padiya, "Data analysis of file forensic investigation," 2016 International Confe rence on Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016,pp.372375,doi:10.1109/SCOPES.2016.7955854.
[8] Anand Singh Rajawat, Dr. Akhilesh R. Upadhya, "Big Web Data Mining for Predicting Usage Behaviour Using Fusion Map Reduce Model ", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCS EIT), ISSN : 2456-3307, Volume3, Issue1, pp.1641-1647, January-February-2018.
[9] Sherer J.A., Sterling N.L., Burger L., Banaschik M., Taal A.(2018) An Investigator's Christmas Carol: Past, Present, and Future Law Enforcement Agency Data Mining Practices. In: Jahankhani H. (eds) Cyber Criminology. Advanced Sciences and Technologies for Security Applications. Springer,Cham. https://doi. org/10.1007/978-3-319-97181-0_12
[10] Manuel J., Cordeiro R., Silva C. (2018) Between Data Mining and Predictive Analytics Techniques to Cybersecurity Protection on eLearning Environments. In: Silhavy R., Silhavy P., Prokopova Z.(eds) Applied Computational Intelligence and Mathematical Metho ds. CoMeSySo 2017. Advances in Intelligent Systems and Computing, vol662. Springer, Cham. https:// doi.org/10.1007/978-3-319-67621-0_17
[11] Farsi M., Daneshkhah A., Far A.H., Chatrabgoun O., Montasari R.(2018) Crime Data Mining, Threat Analysis and Prediction. In: Jahankhani H.(eds) Cyber Criminology. Advanced Sciences and Technologies for Security Applications. Springer, Cham. https://doi .org/10.1007/978-3-319-97181-0_9
[12] Pournouri S., Akhgar B.(2015) Improving Cyber Situ ational Awareness Through Data Mining and Predic tive Analytic Techniques. In: Jahankhani H., Carlile A., Akhgar B., Taal A., Hessami A., Hosseinian-Far A. (eds) Global Security, Safety and Sustainability: Tomorrow's Challenges of Cyber Security. ICGS3 2015. Communications in Computer and Information Science, vol 534. Springer, Cham. https://doi.org/ 10.1007/978-3-319-23276-8_3

© 2021 IJSRET