International Journal Of Mechanical Engineering And Information Technology
An Official Publication Of IGM Publication

Crime Classification and Criminal Psychology Analysis using Data Mining

Authors
Arnab Samanta[1], Amol Joglekar[2]
[1]MSc. Computer Science Mithibai College Mumbai, India
[2]Guide Mithibai College Mumbai, India
Email- *arnab.samanta94@gmail.com*

ABSTRACT—
In today's world, Crime is growing at an exponential rate and it is an anti-social and evil element in the society which needs to be dealt with. Crime analysis is one of the most important activities in crime solving. The process of crime analysis is becoming more difficult due to the increasing amount of crimes. Therefore to make the process of crime analysis easier the use of Data Mining is proposed. Data Mining allows to scan through and analyze large amounts of data in less time and provide important results for the same. The paper first reviews literature on various data mining applications related to the field of crime prediction, detection and analysis. The paper then proposes a model which is used to classify crimes based on their level of seriousness and will provide visualization for the same for easier understanding and analysis by the crime experts. The model will also provide functionality for analyzing psychology of murders using clustering.

Key words— *Crime Analysis, Data Mining, Classification, Classification Rules, Clustering.*

1. INTRODUCTION

Crimes are defined by criminal law, which refers to a body of federal and state rules that prohibit behavior the government deems harmful to society. If one engages in such behavior, they may be guilty of a crime and prosecuted in criminal court[1]. It can also be defined as deviant behavior that violates prevailing norms – cultural standards prescribing how humans ought to behave normally.

Crimes can be divided into four major categories: -

- Personal crimes
- Property crimes
- Inchoate crimes
- Statutory Crimes

Personal crimes are crimes that result in physical or mental harm to another person and these include assault, kidnapping, homicide, rape, false imprisonment, kidnapping. Property crimes are crimes which involve interference with another person's right to use or enjoy their property and these include robbery, burglary, arson, embezzlement, forgery, false pretenses. Inchoate crimes are crimes which were started but not completed and these include attempt to any crime, solicitation, conspiracy. Statutory Crimes are violations of a specific state or federal law and these include drunk driving, selling alcohol to a minor. Crimes can be categorized based on their level of seriousness in descending order. They are Felonies, Misdemeanors and Violations.

Crime rate has been increasing at an alarming rate and the number of crimes committed today is greater than it has ever been before. The frequent and repeated thefts, murders, rapes, shoplifting, pick pocketing, drug- abuse, illegal trafficking, smuggling, theft of vehicles etc., have made the common citizens to have sleepless nights and restless days. They feel very insecure and vulnerable in the presence of anti-social and evil elements. The criminals have been operating in an organized way and sometimes even have nationwide and international connections and links. Thus to help police officers, crime experts and to ensure safety to common citizens, the use of data mining for crime pattern detection and analysis is proposed.

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets and is used widely in the field of artificial intelligence, machine

learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating[2]. Data mining process works on the assumption that more useful knowledge lies beneath the surface. It helps us to answer questions that were traditionally very time consuming to solve manually.

Data mining consists of five major elements:-

- Extract, transform, and load the data into the data warehouse.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table[3].
- Data mining consists of techniques such as association rule mining, classification, and clustering. This paper focuses on Classification and Clustering to analyze crimes.

A. Classification

Classification is a supervised learning method that used to assign objects to one of many pre-determined categories. Classification is a two-step process; the first step is the learning step where a classification model is constructed from a training set made up of database tuples and their associated class labels. Test data is used to estimate the accuracy of the classifier. Accuracy is defined as the percentage of test set tuples that are correctly classified by the classifier from the given test set. In the second step, if the accuracy of the classifier is acceptable then it is used to classify future data tuples for which the class label is not known. The ordering of the labels in classification has no meaning. Algorithms for classification include decision tree, naive bayes, generalized linear models, support vector machine, neural networks etc.

B. Clustering

Clustering is an unsupervised learning method and is the process of partitioning a set of data objects into subsets. Each subset is called a Cluster and objects within a cluster are similar to each other but dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Each data object does not have a class label associated with it as compared to the data objects Classification. Clustering learns by observation rather than by example. The main advantage of clustering compared to classification is its ability to adapt and helps to single out useful features that distinguish different groups. Clustering can be used for outlier detection. Clustering learns by observation rather than by example. Clustering methods can be classified as partitioning method, hierarchical method, density-based method, grid-based method, model-based method, constraint-based method.

2. LITERATURE REVIEW

The proposed research paper is about analyzing crime based on different parameters. Therefore, we need to study different types of crimes and techniques along with what kind of parameters or attributes are to be taken into consideration. Data collection can be done and based on that attributes can be obtained.

Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri[6] propose a pattern detection algorithm called Series Finder which is designed to detect patterns of crime committed by the same individual(s). Series Finder is a supervised learning method and the algorithm grows a pattern of discovered crimes from within a database, starting from a “seed” of a few crimes. It takes into consideration the common characteristics of all patterns called pattern-general coefficients and also the unique aspects of each specific pattern called pattern-specific coefficients. The attributes of the crime are either categorical or numerical.

For pattern building, starting from the seed each candidate crime with the highest pattern-crime

similarity to pattern P is tentatively added to P and its cohesion is evaluated. If the cohesion is large enough, then the pattern will continue to grow. If the cohesion is below threshold, then the pattern will stop growing. The algorithm stops when there are no more crimes present in the database that are closely related to any pattern. The algorithm compares with hierarchical agglomerative clustering and an iterative nearest neighbor approach as competing baseline methods using three different categories: Single Linkage, Complete Linkage, and Group Average. The authors trained the model on 2/3rd of the patterns from Cambridge Police Department and tested on the remaining 1/3rd data. Series Finder was able to discover some crimes that analysts had not previously matched to a pattern and exclude few crimes that analysts agreed should be excluded from a pattern. Advantages of Series Finder compared to other models: each seed grows a pattern independently; second takes into consideration pattern-specific weights and finally the weights of Series Finder are learned from data itself and not detectives.

Arthisree K.S and Jaganraj A [7] put forward a model for Resilient Identity Crime Detection which is used to facilitate real-time search for patterns in a multilayered and principled fashion and also to safeguard credit applications at the first stage of the credit life cycle. The model proposes to achieve resilience by adding two new real time data mining-based layers called Communal Detection and Spike Detection. Communal Detection is a whitelist-oriented approach based on a fixed set of attributes, it is used to find communal relationships and reduce their link score. Spike Detection is an attribute-oriented approach on a variable-size set of attributes; it is used to find spikes in duplicates to increase the suspicion score. The model detects a credit card fraud application pattern by a sudden and sharp rise in spike within short duration with respect to the established baseline. The addition of the new layers help to detect more types of fraudulent attacks, remove redundant attributes, and they also better account for changing behavior. Experimentation with real time data verifies their hypothesis.

P.Charanya, T.Dhivyabharathi, B.Kothai, J.Nivetha [8] also follow a similar approach as by the authors of the above paper. The system also implements Communal Detection and Spike Detection in addition to the non-data mining approaches and it is based on the unsupervised algorithm. Communal Detection algorithm takes as input: current application, moving window, threshold, and state-of-alert and produces: suspicion score, same or new parameter value, and new whitelist as output. Spike Detection algorithm takes as input: current application, moving window, and threshold and produces: suspicion score, attribute weight. The validity of user is determined based on the suspicion score. The new system combines the spike detection and communal detection algorithms to make the system more efficient and secure and it also addresses the issues of scalability, imbalanced class and changing behavior. The authors suggest future enhancements of the system by taking biometric attributes of the customer.

Shyam Varan Nath [9] presents a model to formulate crime pattern detection as machine learning task using data mining. The model uses k-means clustering with some enhancements to support the process of identification of crime patterns and the model uses semi-supervised learning technique. The model first converts operational data into deformalized data using extraction and transform-ation. Next, checks are run to verify the quality of the data. The model assigns weights to all the attributes. The significant attributes for the clustering are identified by crime detectives and crime data analysts and then the attributes assigned weights accordingly. The resulting clusters which have the possible crime patterns are visually represented using a geo-spatial plot of the crime overlaid on the map. The appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. The limitations of the model are that it can only help the detective, not replace them and also it is sensitive to quality of input data. The author suggests future enhancements by extending the model to predict the crime hot-spots.

Ubon Thongsatapornwatana[10] analyses different data mining techniques used for crime detection. The author first talks about association rule mining which is an unsupervised learning method that used to find the hidden knowledge in unlabeled data and also used to discover the interesting co-occurrences of objects in large data sets. Next the author analyses clustering techniques which is used to identify groups of records that are similar between themselves but different from the rest of the data. The different techniques used in clustering such as K-means, Hierarchical and Expectation-Maximization are discussed in detail. The author then talks about Classification which is a supervised learning method that used to assign objects to one of many pre-determined categories. The different approaches for classification such as decision tree, nearest neighbor and neural network are discussed in detail. The author follows by talking about the use of data mining techniques in the field of traffic violation and border control, violent crime, narcotics, cyber-crime and provides a summary of the same. The author then proceeds to explain the issues and challenges faced in crime detection with respect to data collection and integration, crime pattern, performance and visualization.

Isuru Jayaweera, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera, Adeesha Wijayasiri [11] propose web based crime analysis system which includes crime analysis techniques such as crime pattern visualization, hotspot detection and crime comparison. The system is composed of the following modules: crawler, classifier, entity extractor, duplicate detector, data base handler, analyzer and graphical user interface. The crawler finds newspaper articles from different newspapers. The document classifier classifies crawled newspaper articles as crime and non-crime articles and stores them in the database. Entity extractor extracts important entities from crime related newspaper articles. Duplicate Detector finds exact/near duplicates of newspaper articles and removes them from the database. Database handler handles all the database transactions of the system. Analyzer module is used to perform crime analysis operations on processed crime articles. The web based GUI is used to visualize crime statistical details of the previous years. The limitation of the system is that since it is based on newspaper articles hence it includes only a subset of total crime incidents. Crime prediction and addition of more rules to entity extraction module is proposed as future enhancement.

Author Mugdha Sharma[12] puts forward a tool which applies an enhanced Decision Tree Algorithm to detect the suspicious e-mails about the criminal activities using an improved ID3 Algorithm. ID3 is a supervised learning algorithm based on information entropy and the tool classifies the data based on the importance of each attribute. The tool extracts “suspicious keywords” and “non-suspicious indicators” from the e-mail message and the combination of the keywords and indicators are analyzed to detect criminal activities. The construction of the decision tree starts at the root which contains all the suspicious keywords. The keywords are partitioned recursively based on selected attributes and the attribute with highest importance becomes the root. The process stops when all the attributes are mapped in to the tree based on the sorted attribute- importance. The tool learns patterns from the training sample set and it is able to classify a new e-mail as suspicious, non-suspicious or may-be-suspicious. The problem faced while implementing the tool was the lack of available data. The author proposes to enhance the tool in the future by incorporating other methods of feature selection, and different classification techniques.

Li Ding, Dana Steil, Matthew Hudnall, Brandon Dixon, Randy Smith, David Brown, Allen Parrish [13] present an integrated system called PrepSearch which takes all the details of the crime as input and uses a combination of geographic profiling, social network analysis, crime patterns, and physical matching to detect suspects. In the first step, the system applies a geospatial search based on selected criteria to get all matching addresses. The system then retrieves lists of all persons related to the address or has a close relation to the address. In the next step the system refines the list of perpetrators by using biometric filtering techniques. The final

step uses a crime pattern component to rank suspects based on their criminal history. The system provides two types of visualization tools, one to provide the geographic view of crimes and the other to provide visualization ability for social networks. The limitation of the system is that it lacks the integration between the visualization tools. For future enhancements the authors propose to expand the crime description details.

Devendra Kumar Tayal, Arti Jain, Surbhi Arora , Surbhi Agarwal , Tushar Gupta , Nikhil Tyagi [14] propose an approach for the design and implementation of crime detection and criminal identification for Indian cities using data mining techniques. The proposed model consists of six modules: data extraction, data preprocessing, clustering, Google map representation, classification and weka implementation. Data extraction is used to extract raw crime data from various crime web sources. Data preprocessing cleans, integrates and reduces the extracted crime data and also provides structure to the same. K-means clustering is then used to group crime instances iteratively into two clusters with similar attributes for crime detection. The Google map representation helps to improve visualization by representing crime clusters, which are used to identify the hot spots of crime locations. Classification is done using K-NN which is used to discover similarities among different crimes and then place them into predefined classes for criminal identification and prediction. The verification of the output of the model is done through weka. In the future, the authors propose to enhance data privacy, reliability, accuracy and other security measures of the proposed model.

Mohammad Reza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia[15] put forward a multi-purpose framework for intelligent crime investigation using data mining techniques. It mainly focuses on using crime matching with behavioral burglary crime variables. The framework consists of three basic components: Entity Extraction, Data clustering and Neural Network. Entity extraction is used to extract important entities from police narrative reports and a lexical lookup approach has been used for the same. Data clustering is done in two steps, first a self-organizing neural network is used to extract the feature map. In the second step, AGNES algorithm is used for clustering which is a hierarchical bottom-up algorithm. The classification process is done by means of a Multi-Layer Perceptron (MLP) neural network with back-propagation training method. A MLP neural network is devised for each category of burglary crime variables and a MLP classifier is used because of its high tolerance to noisy data. The author proposes to incorporate spatio-temporal data along with behavioral crime variables in the future and also to implement this framework as integrated enterprise software.

Shiju Sathyadevan, Devan M.S and Surya Gangadharan. S[16] present a system which can predict regions which have high probability for crime occurrence and can visualize crime prone areas. The system performs the following: data collection, classification, pattern identification, prediction and visualization. Data collection consists of collecting data from different web sites and storing them in Mongo database. Classification is done using Naive Bayes which is a supervised learning method as well as a statistical method for classification based on probability. Crime pattern is found using apriori algorithm which is used to determine association rules which highlight general trends in the database. Crime prediction is done using a supervised machine learning technique which builds a decision tree from a set of class labeled training samples uses a set of binary rules to calculate the class value. Visualization is used to depict crime prone areas graphically using a heat map. The limitation of the system is that it takes into consideration only a limited set of factors for crime detection. Also the system predicts crime prone regions for a particular day and not the time at which the crime is happening.

3. PROPOSED MODEL

The proposed model provides two functions; first to classify the crimes based on their level of seriousness and second to analyze the psychology of murders using clustering. The model will be able to classify crimes of the following type: - Murder,

Rape, Theft, Traffic Violation, Kidnapping, Cyber-crime, Assault, Trespassing and Vandalism. The model will classify the crime into one of the following classes based on their level of seriousness, which are Felonies, Misdemeanors, and Violations. Felony is the most serious level of crime and violation is the least serious level of crime. Each crime will be classified into their appropriate classes based on rules. One crime of one particular crime type can belong to more than one class. For example, if the crime type is Traffic Violation and the crime committed is going over the speed limit then the crime will be classified as a Violation. If the crime type is Traffic Violation and the crime committed is driving under the influence of alcohol/drugs then the crime will be classified as a Misdemeanor. If the crime type is Traffic Violation and the crime committed is Hit and Run then the crime will be classified as a Felony. A few rules for classification have been stated as example in "Table 1" below. The rules generated by the model will be of similar nature

Rule Set

Rule No.	Condition	Class
1	type=Murder	Felony
2	type=Rape	Felony
3	type=Theft and amount<=10000	Violation
4	type=Theft and amount>10000 and amount<=100000	Misdemeanor
5	type=Theft and amount>100000	Felony
6	type=Traffic Violation and sub_type=Over Speeding	Violation
7	type=Traffic Violation and sub_type=DUI	Misdemeanor
8	type=Traffic Violation and sub_type=Hit and Run	Felony
9	type=Kidnapping and sub_type=Victim unharmed	Misdemeanor
10	type=Kidnapping and sub_type=Victim harmed	Felony
11	type=Cyber Crime and sub_type=Spamming and Phishing	Violation
12	type=Cyber Crime and sub_type=Hacking	Misdemeanor
13	type=Cyber Crime and sub_type=Identity Theft	Felony
14	type=Assault and sub_type=Simple	Violation
15	type=Assault and sub_type=With weapon	Misdemeanor
16	type=Assault and sub_type=Aggravated	Felony
17	type=Trespassing	Violation
18	type=Vandalism	Violation

The rules for classifying the crimes will be determined a sequential covering algorithm called PRISM. PRISM algorithm is used for learning rules for classification directly from the training data set. It takes as input a training dataset and produces a set of rules as output. The PRISM algorithm uses a depth-first search to construct the next rule for a given class C. Since the consequent of the rule is known, only the antecedent needs to be constructed. This is done by starting with an empty antecedent and iteratively adding the most promising attribute=value constraint next from the training set. This depth-first search continues until the resulting rule is specific enough that it makes no classification errors over the available data instances. The number of instances correctly classified by the rule is removed and the whole process is repeated again until all instances of the dataset can be classified by some rule. The set of rules which is the output of the prism algorithm is verified for its accuracy and correctness by a crime expert. After which any new crime instance is classified using the rules. To enhance the results of classification, the model will also provide visualization in the form of charts and graphs. Charts will indicate the increase or decrease of number and type of crimes within a specified time period. Charts would also allow comparing the percentage of crimes committed based on type of crime or level of seriousness. The model would also to provide visualization on a map to indicate hotspots corresponding to the number of crimes committed in that particular area. Thus the visualization tools would allow the police officers and crime experts to understand what type of crime and which areas require more attention. It would also allow them to understand the trends involved in different types of crime. Based on this information, the crime experts and police officers could take the necessary steps to curb the crime rate.

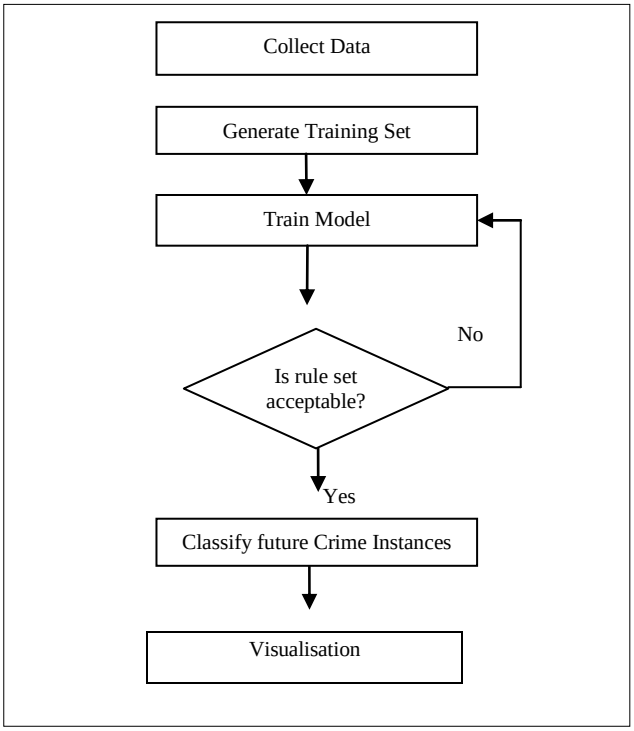

Fig-1 Flow Chart for Classification Function

The second function performed is clustering for analyzing the psychology of murders. The various reasons for committing murder are recorded based on interviews conducted with the murderers. These reasons include Revenge, Money, Love, Childhood Trauma, Substance Abuse, Medical Condition, Fantasy and Others. These reasons are encoded to symmetric binary variables by the model, where 1 indicates a positive response and 0 indicates a negative response. For example, if the reason identified for committing murder is Revenge and Substance Abuse, then the binary string generated by the model will be of the form "10001000" where each bit of the string represents one reason for committing the murder. After the binary encoding is complete, a dissimilarity matrix will be constructed using the dissimilarity measure between two murderers calculated using Symmetric Binary Dissimilarity also known as Simple Matching Distance. Symmetric Binary Dissimilarity for two binary strings i and j is calculated using the formula:-

$$d(i,j) = (r+s)/(q+r+s+t)$$

where q is the number of variables that equal 1 for both objects i and j

r is the number of variables that equal 1 for object i but that are 0 for object j

s is the number of variables that equal 0 for object i but equal 1 for object j

t is the number of variables that equal 0 for both objects i and j.

After the dissimilarity matrix is generated, clustering is applied using CLARANS algorithm which is a variation k-medoids algorithm. The clustering process is similar to that of searching a graph where every node is a potential solution. Two nodes are said to be neighbors if their sets differ by only one object. Each node is assigned a cost that is defined to be the total dissimilarity between every object and the medoid of its cluster. At each step, all neighbors of current node are searched and the neighbor which corresponds to the deepest descent in cost is chosen as the next solution. The set of neighbors is randomly chosen by the algorithm at each step. If the local optimum is found then the algorithm starts with new randomly selected node in search for a new local optimum and the number of local optimums to be searched for is a parameter. CLARANS is chosen because it is more efficient and scalable than both PAM and CLARA and returns higher quality clusters. The resultant clusters after clustering would allow crime experts to gain more in depth information regarding the minds of the murderers.

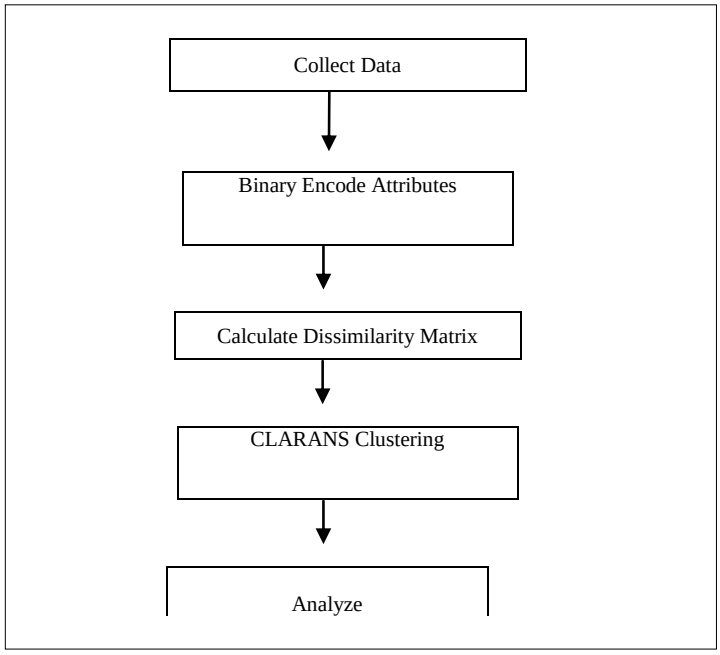

Fig.2 Flow Chart for Clustering Function

CONCLUSION

Thus the model will be able to classify a crime based on its level of seriousness as well as provide insight into the minds of murderers using clustering. The model is only proposed but not implemented. In the future the model can be implemented and tests can be conducted to check the accuracy and correctness of the results of the model. If the results are acceptable then the model can be used to classify crimes in real time and also to gain deeper insight into the minds of the murderers, thus making the jobs for the crime experts easier. In the future, the model could be extended to classify more types of crime and also clustering process could be performed to analyze the phycology of other criminals in addition to murderers.

REFERENCES

1. http://www.legalmatch.com/law-library/article/what-are-the-different-types-of-crimes.html
2. http://www.kdd.org/curriculum/index.html
3. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
4. Data Clustering: Algorithms and Applications by Charu C. Aggarwal, CRC Press Publications
5. Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann Publishers
6. Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri,"Learning to Detect Patterns of Crime", available at:http://web.mit.edu/rudin/www/docs/WangRuWaSeECML13.pdf
7. Arthisree K.S and Jaganraj A,"Identify Crime Detection Using Data Mining Techniques", ISSN: 2277 128X,© 2013 IJARCSSE
8. P.Charanya, T.Dhivyabharathi, B.Kothai, J.Nivetha,"Forge Detection in Credit Application", ISSN: 2320-2106,International Journal of Advanced Computational Engineering and Networking
9. Shyam Varan Nath,"Crime Pattern Detection Using Data Mining",2006 IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology
10. Ubon Thongsatapornwatana,"A Survey of Data Mining Techniques for Analyzing Crime Patterns", ISSN: 978-1-5090-2258-8/16,©2016 IEEE
11. Isuru Jayaweera, Chamath Sajeewa, Sampath Liyanage, Tharindu Wijewardane, Indika Perera, Adeesha Wijayasiri,"Crime Analytics: Analysis of Crimes through Newspaper Articles",ISSN: 978-1-4799-1740-2/15,©2015 IEEE
12. Mugdha Sharma,"Z - CRIME: A Data Mining Tool for the Detection of Suspicious Criminal Activities Based on Decision", ISSN: 978–1–4799–4674–7/14,©2014 IEEE
13. Li Ding, Dana Steil, Matthew Hudnall, Brandon Dixon, Randy Smith, David Brown, Allen Parrish,"PerpSearch: An Integrated Crime Detection System", ISSN:978-1-4244-4173-0/09,©2009 IEEE
14. Devendra Kumar Tayal, Arti Jain, Surbhi Arora , Surbhi Agarwal , Tushar Gupta , Nikhil Tyagi,"Crime detection and criminal identification in India using data mining techniques", AI & Soc (2015) 30:117–127,©Springer-Verlag London 2014
15. MohammadReza Keyvanpoura, Mostafa Javidehb, Mohammad Reza Ebrahimia,"Detecting and investigating crime by means of data mining: a general crime matching framework", Procedia Computer Science 3 (2011) 872–880,WCIT 2010, ©2010 Published by Elsevier Ltd
16. Shiju Sathyadevan, Devan M.S and Surya Gangadharan. S,"Crime Analysis and Prediction Using Data Mining", ISSN:978-1-4799-3486-7/14 ©2014 IEEE
17. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach and Vipin Kumar, Pearson publishers http://legaldictionary.net/