A Comparative Study of Analysing and Clustering Crime Patterns Using Data Mining.

A. U. Bapat[#1], Shreya Desai[*2]

[#1]*Department of Computer Science and Engineering, Goa College of Engineering, Farmagudi, Ponda-Goa,India*

[1]uab@gec.ac.in

[*2]*Department of Computer Science and Engineering, Goa College of Engineering, Farmagudi, Ponda-Goa,India*

[2]shreya3196@gmail.com

***Abstract*— Crime is the utmost potent and disquieting aspects in our day today life. Crime analysis is a sorted way of understanding and inspecting patterns and trends in crime. Analysis of crime is cardinal for providing uttermost immunity and protection society. Data mining can help us to discover and understand critical information which can further guide the law enforcement authority to detect crime and areas of importance. The main motive of the paper is to understand the crime which entails, robbery, suicides and various offences which also include suspicious activities, noise complaints. The different algorithms that can be used for clustering and analysing crime activities based on some previously defined cases are Agglomerative clustering algorithm, K-means clustering and Density based spatial clustering with noise (DBSCAN) algorithm. A correlative study of these algorithms is presented based on different crime patterns.**

***Keywords*—Crime, data mining, qualitative, quantitative, k-means, agglomerative, DBSCAN.**

I. Introduction

This In the thrive world, the daily activities of humans' communal, political and economic life makes it vital and easy to encounter the phenomenon of crime. Crime is an unnecessary evil in the society and for any economic, social and political activities to run smoothly, crime offences must be completely eliminated from the society and therefore knowledge of crime analysis is required.

The main motive of this paper is to discern the crime patterns and perform analysis of crime in order to recognize the occasionally contrastive trends over time in crime activity. It entails the use of statistical techniques, analytical methods and even the application of scientific social data collection to encapsulate the occurrence of any crime.

There are various types of crimes which are investigated in this manner such as theft, drug related offences, homicide, prostitution offences, robbery, blackmail and extortion. This paper aims to examine the process of crime analysis. This can give people foresight in to the type of crimes in their area and be aware of it before it happens to them. Identification of the crime by observing characteristics of a particular location and analysis of content is called the qualitative analysis of a crime.

Data mining is made with the addition of two words 'data' and 'mining' in which mining includes the relation between the values of data of historical and current stipulation. Data mining is being used in many different fields to better understand and visualize data while also finding patterns and trends to enhance current information.

Crime data mining is being studied by federal, state, local, commercial and academic agencies. It aims to find new trends using data collected by local law enforcement. One such system was theorized in the "Crime Data Mining: A General Framework & Some Examples" paper [8]. Some techniques like "Entity extraction, Clustering, Association, Sequential pattern mining and others" were used in order to achieve their goal.

II. Fundamentals Of Data Mining

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge [7].A knowledge discovery process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation [7].

Data mining can be deemed as an output of the evolution of information science and technology. Data mining is initiated from the tasks such as concept description, where data is associated into classes and concepts. Data characterization is a summarization of the general characteristics or features of a target class of data [7]. The data is collected by the database query.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes [7]. Frequent patterns are the patters that occur repeatedly in the data or a transaction that has been performed. Association analysis is conducted to find the rules and their relations. Classification term refers to finding a model which is par descriptive to classify the data into a particular class or concept. Prediction is a task where the data is categorized based on the concept or label. Clusters are built based on the relativeness of the data. A cluster is formed where similar items are categorized into a single knot. Clustering analysis is a very challenging assignment because only observatory research has completed by experts who adjust the initial point of the centroid location of the cluster in datasets. [4]. Data mining also peeps into the concept of outlier mining, where the data item that tends to

violet the properties as specified are eliminated. Considering the crime data set, the specific tasks such as pre-processing, analyzing, clustering and outlier mining could be performed using the following algorithms:

A. Association Rule Mining:

This is an unsupervised method of where the hidden knowledge is discovered. Association rules are used to locate the occurrence of the data association in the data set. If the LHS item-set occurs then the RHS item-set will be likely to occur [9].

B. Clustering

Clustering is an unsupervised data analyzing method. Similar type of data is divided into an atomic group and dissimilar data into different groups. There are various concepts in the clustering techniques such as distribution based, centroid based, connectivity based, density models, subspace clustering. The most customary clustering algorithms are k-means, DBSCAN, agglomerative hierarchal, optics and error maximization.

C. Classification

Classification is a process where categorization of data is tendered into number of predefined classes. This aims at identifying the category/class into which a new data item will fall. The commonly used classification algorithms are Naive Bayes Classifier, Support Vector Machines. Decision Trees, Boosted Trees, Random Forest, Neural Networks, Nearest Neighbor.

III. Literature Survey

Zhang Ying [1] proposed a method for analyzing the crime factors correlation on data mining using the association analysis to find the frequent set of items and generate valid association rules among them and find the correlation analysis. For a given dataset, data mining technologies are used to calculate the confidence and support. Analysis and prediction further helps in decision making. Clustering is done to depict the laws and new patterns from the dataset. This helps to blend all the patterns of a single type in a group. Outlier mining is done to analyze the abnormal behavior which deviates from the described task which may lead to greater significance.

Sunil Yadav et al[2] bid a method of regression model for predicting the crime rate. Supervised, semi-supervised and unsupervised learning techniques were used for the discovery of knowledge which further led to the predictive accuracy of the dataset. The data set comprises of 42 crime heads with 14 attributes. The algorithms used are association mining along with clustering. The paper aims at uncovering patterns, trends and discerns possible conjecture. K-means algorithm forms two clusters with high and low number of people involved in the crime. Weka tool is used for the applying of k-means. The outcome of k-means algorithm are used as dataset (input) for apriori as the clusters have high and low values of the crime dataset. Apriori algorithm is worn to associate the rules and perceive the frequent data items. Classification is one of the classic data mining techniques, which is used to classify each item in a set of data into one of predefined set of classes or groups [2]. With the help of classification, existing dataset can easily be understood and it also helps to predict how new individual dataset will behave based on the classification criteria. Data mining creates classification models by observing already classified data and finding a predictive pattern among those data [2] .Correlation and regression analysis is performed to the multivariate analysis. This model reduced the crime and helped in crime detection by performing all of the above measures.

Shiju Sathyadevan, Devan M.S et al[3] followed the

preliminary step of data collection from all the social media sites and RSS feeds. The gathered data was stored in Mongo DB for future processing. The data used was the unstructured data. NoSQL was used over SQL database because it allows data insertion without any predefined schema. Naïve Bayes algorithm is a supervised algorithm and statistical method which was used for the purpose of classification based on the probabilistic approach. Naive Bayes algorithm is preordained over any other algorithm because it converges quickly. The authors used the naïve Bayes algorithm to build a model related to all the crime heads. The model was tested to classify the unknown input by training on known input. Test results showed that naïve Bayes was approximately 90% accurate. Named entity Recognition was also introduced to classify the elements into predefined categories. Third phase was the pattern identification where new trends and patterns were identified using the apriori algorithm. Decision tree concepts were used for the prediction of crime. A decision tree was built using the supervised learning techniques from a set of labelled training samples. Further the crime prone area was graphically represented indicating the criminal activities.

Turki Aljrees et al [4] proposed modified k-means as an extension of k-means algorithm. K-means is a standard algorithm where the centroids to be initialized. Having set the centroids, the clustering process is performed. The distance between each point is calculated with respect to the centroid. The process is repeated until the algorithm converges. The most important disadvantage is the initial center point and the fact that there is no correct way of doing it[4]. Thus modified k-means was introduced where the frequency of each data point was calculated and the entire space was partitioned into few segments. The data point was assigned to correct the cluster's centroid; Step one involved half of the minimum distance from i'th cluster's centroid to the remaining cluster's centroid [4]. The second step intricate reviewing data point to evaluate the distance from i'th centroid and then collate it with dC(i). Then, in the third step, if it was equivalent or a

TABLE I: COMPARATIVE STUDY OF CRIME ANALYSING TECHNIQUES.

Paper Number	Title	Focus	Methods/Tool	Advantages	Future Work
1	Analysis of crime factor correlation based on data mining technology.	To determine the crime factors and find the relation between them using data mining techniques	• Association analysis. • Analysis and prediction • Cluster analysis • Outlier mining	• Highlights the hidden information due to the use of association mining.	• Focus on large scale databases and high dimensional data issues. • Focus on structured data.
2	Crime Pattern Detection, Analysis & Prediction	To discover the knowledge and increase the crime prediction accuracy	• Association Mining (Apriori) Clustering(k-Means) • Classification Techniques (Naive Bayes) • Correlation & Regression • Weka and R tool	• Highlights on the insights of the data set. • Correlation and regression is used to find the perfect relations of the attributes.	• Data acquisition and staging • Extension of crime detection and analysis will be to generate the crime hot-spots that will help in deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources.
3	Crime analysis and prediction using data mining	Extract previously unknown, useful information from an unstructured data that will help to solve crimes.	• Data collection • Classification • Pattern identification • Prediction • Visualization	• Highlights on the different algorithms that can be used for classification, pattern recognition and prediction.	• Crime profiling • Snatching • Predict the time when the crime has occurred.
4	Criminal pattern identification based on modified k-means clustering	To propose a modified k-means algorithm that leads to better ways of observing the data set and determine the similarities and dissimilarities as a specific domain.	• K-means • Modified k-means	• Helps to achieve hidden predictive patterns by observing and analyzing vast databases. • Helps in finding the number ofclusters needed to initialize the centroid point that will lead to better clustering • Applying modified k-means helps to understand the correct factor that is to be considered in the cluster.	• Implement modified k-means on vast data set and optimize the clusters formed.
5	Crime	To find the	• K-means	• Understanding different	• Enhance privacy and other

	Prediction and Forecasting in Tamilnadu using Clustering Approaches	best algorithm for clustering crime data	• Agglomerative hierarchal clustering • DBscan • KNN	types of algorithm that can be used for clustering crime data.	security measures to protect the crime data
6	Analysis and Prediction of Crimes by Clustering and Classification	To classify clustered crimes based on occurrence frequency during different years	• Genetic algorithm(GA) • Rapid miner	• Outlier detection was performed using the GA	• Improve the clustering and optimizing process

smaller amount than dC(i) , then the data point was assigned to the i'th cluster[4].In collation to k-means, the modified k- means algorithm could cluster the patterns depending on the prominent factor. This method also allowed gaining different hidden patterns from a huge dataset. Moreover it was very easy to find the right factor that could be used for clustering.

S.Sivaranjani et al [5] proposed a schematic flow of crime detection in Tamilnadu. The author aimed at reducing the crime rate in the state. The data set was extracted from National Crime Records Bureau (NCRB) of India. The dataset has 20 crime heads. The raw data was pre-processed and a crime database is prepared. Furthermore, crime identification and prediction is conducted. The KNN classification searches through the dataset to find the similar or most similar instance when an input is given to it [5].KNN helps to analyze large datasets for predicting the forthcoming crimes in various cities. K-means algorithm is used to detect the patterns that are internally present and obtain the relationship of the attributes within. K-means is further worn for clustering the dataset into similar groups. GMAPI is used to visualize the results of k-means. A bottom up approach of agglomerative hierarchal clustering is worn to obtain atomic cluster. DBScan is implemented to cluster high density areas into an arbitrary cluster. The performance of all the algorithms was evaluated on the basis precision, F-measure and recall. Among all the clustering algorithms, it was noticed that DBScan gives the best results.

Rasoul Kiani et al [6] presented a frame work for the purpose of clustering of crime data and prediction the crime data based on real data. The genetic algorithm was implemented to work upon the outlier detection. Fitness function was illustrated depending upon the accuracy and error parameters of the classification function. The features were given a weight. The features with low weight were deleted building upon the threshold that was set appropriately. The main purpose of this framework was to generate the training and the testing data. Low weighted features were deleted which lead to high dimensional data. Further the optimization of the outliers was performed using the genetic algorithm.

IV. Conclusions

It is concluded that, algorithms like k-means is most prominently used for the purpose of clustering crime data. Clustering techniques are used for crime detection and classification techniques are used for crime prediction. The K- Means clustering, Agglomerative hierarchical clustering and DBSCAN clustering are the most commonly used algorithms. Association rule mining helps in deriving the relations among the datasets. Upon relative comparison it was noticed that density based spatial clustering gave the highest accuracy and formed effective clusters. From the above study, it is understood that KNN can be used for the prediction of crime. Naïve Bayes algorithm forms the classification task with best results and hence considered to be one of the best algorithms for classification. This will help the law enforcement agencies to understand the type of crime happening and work upon the reduction of crime rates. Tools like Weka, Rapid miner and R tool can be worn to perform the above tasks.

Acknowledgement

I would like to convey my sincere gratitude towards my guide Prof. A. U. Bapat for his continuous support and guidance. Further I would also like to thank Prof. (Dr.) J.A. Laxminarayana, Head of Computer Department and Dr. Krupashankara M.S who gave me the opportunity to work on the entitled project, which helped me to understand the latest trends in technology related to crime prediction and analysis.

I would also like to thank my parents and my classmates for their constant support throughout.

References

[1] Zhang Ying, "Analysis of crime factors correlation based on data mining technology", 2016 International conference on Robots and Intelligent System,IEEE.

[2] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav, "Crime Pattern Detection, Analysis & Prediction",International Conference on Electronics, Communication and Aerospace Technology, IEEE.

[3] Shiju Sathyadevan, Devan M.S and Surya Gangadharan. S,"Crime Analysis and Prediction Using Data Mining ",2014 First International Conference on Networks & Soft Computing, IEEE.

[4] Turkialjrees, Damingshi, Davidwindridge And Williamwong, "Criminal Pattern Identification Based On Modified K-Means Clustering", Proceedings Of The 2016 International Conference On Machine Learning And Cybernetics,
Jeju, South Korea, 10-13 July, 2016, IEEE

[5] S.Sivaranjani,Dr. S. Sivakumari and Aasha M,"Crime Prediction and Forecasting in Tamilnadu using Clustering Approaches",2016 International Conference on Emerging Technological Trends [ICETT], IEEE.

[6] Rasoul Kiani, Siamak Mahadavi and Amin Keshavarzi," Analysis and Prediction of Crimes by Clustering and Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.8, 2015.

[7] Jiawei Han and Micheline Kamber,"Data Mining:Concepts and Techniques, Second edition"

[8] H. Chen ; W. Chung ; J.J. Xu ; G. Wang ; Y. Qin and M. Chau,"Crime data mining: a general framework and some examples, IEEE Computer Society.

[9] Ubon Thongsatapornwatana,"A Survey of Data Mining echniques for Analyzing Crime Patterns",2016 Second Asian Conference on Defense Technology (ACDT), IEEE.

[10] Chhaya Chauhan and Smriti Sehgal,"A Review: Crime Analysis Using Data Mining Techniques and Algorithms",International Conference on Computing, Communication and Automation (ICCCA2017), IEEE.

[11] Nurul Hazwani Mohd Shamsuddin, Nor Azizah Ali, Razana Alwee,"An Overview on Crime Prediction Methods", 2017 6th ICT International Student Project Conference (ICT-ISPC), IEEE.