Crime Detection Technique Using Data Mining and K-Means

[1]Khushabu A. Bokde, [2]Tiksha P. Kakade,
[3]Dnyaneshwari S. Tumsare ,[4]Chetan G. Wadhai
B. E. Student
Department of CSE, Ballarpur Institute of Technology,
Ballarpur,Chandrapur District, Maharashtra, India.

Prof. Deepa Bhattacharya
Assistant Professor
Department of CSE, Ballarpur Institute of Technology,
Ballarpur, Chandrapur District, Maharashtra, India.

***Abstract*—Crimes will somehow influence organizations and institutions when occurred frequently in a society. Thus, it seems necessary to study reasons, factors and relations between occurrence of different crimes and finding the most appropriate ways to control and avoid more crimes. The main objective of this paper is to classify clustered crimes based on occurrence frequency during different years. Data mining is used extensively in terms of analysis, investigation and discovery of patterns for occurrence of different crimes. We applied a theoretical model based on data mining techniques such as clustering and classification to real crime dataset recorded by police in England and Wales within 1990 to 2011. We assigned weights to the features in order to improve the quality of the model and remove low value of them. The Genetic Algorithm (GA) is used for optimizing of Outlier Detection operator parameters using RapidMiner tool.**

Keywords— Crime; Clustering; K-Means Algorithm;

I. INTRODUCTION

A. Crime Analysis

Today, collection and analysis of crime-related data are imperative to security agencies. The use of a coherent method to classify these data based on the rate and location of occurrence, detection of the hidden pattern among the committed crimes at different times, and prediction of their future relationship are the most important aspects that have to be addressed.

In this regard, the use of real datasets and presentation of a suitable framework that does not be affected by outliers should be considered. Preprocessing is an important phase in data mining in which the results are significantly affected by outliers. Thus, the outlier data should be detected and eliminated though a suitable method. Optimization of Outlier Detection operator parameters through the GA and definition of a Fitness function are both based on Accuracy and Classification error. The weighting method was used to eliminate low-value features because such data reduce the quality of data clustering and classification and, consequently, reduce the prediction accuracy and increase the classification error.

The main purposes of crime analysis are mentioned below [1]:

- Extraction of crime patterns by crime analysis and based on available criminal information,
- Prediction of crimes based on spatial distribution of existing data and prediction of crime frequency using various data mining techniques,
- Crime recognition.

B. Clustering

Division of a set of data or objects to a number of clusters is called clustering. Thereby, a cluster is composed of a set of similar data which behave same as a group. It can be said that the clustering is equal to the classification, with only difference that the classes are not defined and determined in advance, and grouping of the data is done without supervision [2].

C. Clustering by K-means Algorithm

K-means is the simplest and most commonly used partitioning algorithm among the clustering algorithms in scientific and industrial software [3] [4] [5]. Acceptance of the K-means is mainly due to its being simple. This algorithm is also suitable for clustering of the large datasets since it has much less computational complexity, though this complexity grows linearly by increasing of the data points [5]. Beside simplicity of this technique, it however suffers from some disadvantages such as determination of the number of clusters by user, affectability from outlier data, high-dimensional data, and sensitivity toward centers for initial clusters and thus possibility of being trapped into local minimum may reduce efficiency of the K-means algorithm [6].

II. LETRATURE REVIEW

J. Agarwal, R. Nagpal and R. Sehgal in [1] have analyzed crime and considered homicide crime taking into account the corresponding year and that the trend is descending from 1990 to 2011. They have used the k-means clustering technique for extracting useful information from the crime dataset using RapidMiner tool because it is solid and complete package with flexible support options. Figure1 shows the proposed system architecture.

Priyanka Gera and Dr. Rajan Vohra in [11] have used a linear regression for prediction the occurrence of crimes in Delhi (India). They review a dataset of the last 59 years to predict occurrence of some crimes including murder, burglary, robbery and etc. Their work will be helpful for the local police stations in decision making and crime supervision.

―After training systems will predict data values for next coming fifteen years. The system is trained by applying linear regression over previous year data. This will produce a formula and squared correlation( ).

The formula is used to predict values for comong future years. The coefficent of determination, , is useful because is gives the proportion of variance of one variable that is predictable from other variable.‖ Figure 2 shows the proposed system architecture.

In [12] an integrated system called PrepSearch have proposed by L. Ding et al. It has been combined using two separate categories of visualization tools: providing the geographic view of crimes and visualization ability for social networks. ―It will take a given description of a crime, including its location, type, and the physical description of suspects (personal characteristics) as input.

To detect suspects, the system will process these inputs through four integrated components: geographic profiling, social network analysis, crime patterns and physical matching.‖ Figure 3 shows the system design and process of PrepSearch.

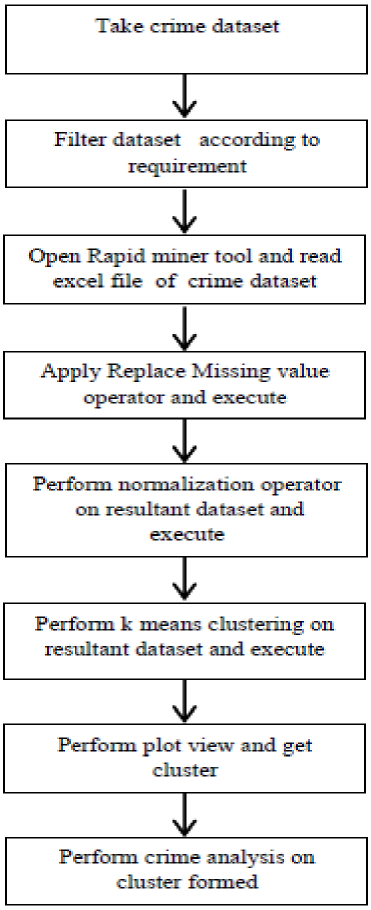
Take crime dataset
Filter dataset according to requirement
Open Rapid miner tool and read excel file of crime dataset
Apply Replace Missing value operator and execute
Perform normalization operator on resultant dataset and execute
Perform k means clustering on resultant dataset and execute
Perform plot view and get cluster
Perform crime analysis on cluster formed

Fig. 1. Flow chart of crime analysis [1]

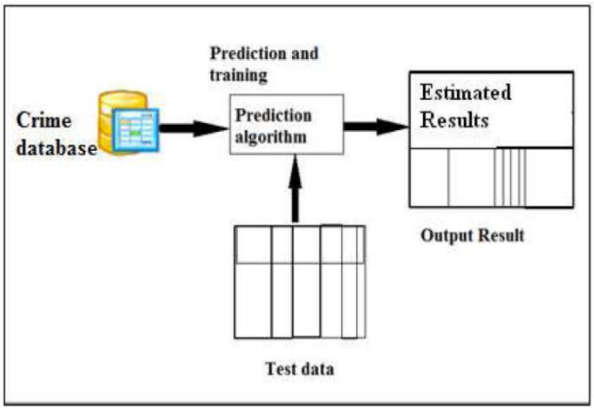

Fig. 2. Predicting future crime trends [11]

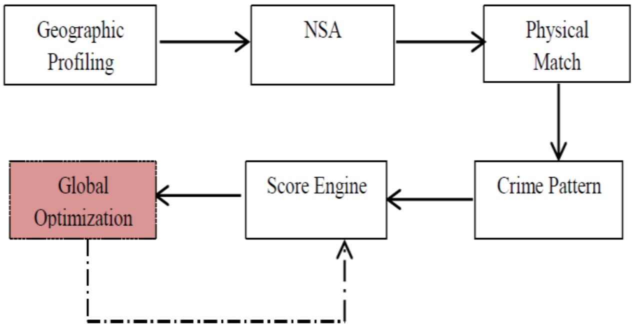

Fig. 3. System design and process of PrepSearch [12]

In [13] researches have introduced intelligent criminal identification system called ICIS which can potentially distinguish a criminal in accordance with the observations collected from the crime location for a certain class of crimes. The system uses existing evidences in situations for identifying a criminal by clustering mechanism to segment crime data in to subsets, and the Nave Bayesian classification has used for identifying possible suspect of crime incidents. ICIS has been used the communication power of multi agent system for increasing the efficiency in identifying possible suspects. In order to describe the system ICIS is divided to user interface, managed bean, multi agent system and database. Oracle Database is used for implementing of database, and identification of crime patterns has been implemented using Java platform.

In [14] an improved method of classification algorithms for crime prediction has proposed by A. Babakura, N. Sulaiman and M. Yusuf. They have compared Naïve Bayesian and Back Propagation (BP) classification algorithms for predicting crime category for distinctive state in USA. In the first step phase, the model is built on the training and in the second phase the model is applied. The performance measurements such as Accuracy, Precision and Recall are used for comparing of the classification algorithms. The precision and recall remain the same when BP is used as a classifier.

In [15] researches have introduced crime analysis and prediction using data mining. They have proposed an approach between computer science and criminal justice to develop a data mining procedure that can help solve crimes faster. Also they have focused on causes of crime occurrence like criminal background of offender, political, enmity and crime factors of each day. Their method steps are data collection, classification, pattern identification, prediction and visualizatiztion.

III. EXISTING SYSTEM

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. De Bruin et. al. [1] introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. Manish Gupta et. al. [2]. highlights the existing systems used by Indian police as e-governance initiatives and also proposes an interactive query based interface as crime analysis tool to assist police in their activities. He proposed interface which is used to extract useful information from the vast crime database maintained by National Crime Record Bureau (NCRB) and find crime hot spots using crime data mining techniques such as clustering etc. The effectiveness of the proposed interface has been illustrated on Indian crime records. Nazlena Mohamad Ali et al.[3] discuss on a development of Visual Interactive Malaysia Crime News Retrieval System (i-JEN) and describe the approach, user studies and planned, the system architecture and future plan. Their main objectives were to construct crime-based event; investigate the use of crime based event in improving the classification and clustering; develop an interactive crime news retrieval system; visualize crime news in an effective and interactive way; integrate them into a usable and robust system and evaluate the usability and system performance and the study will contribute to the better understanding of the crime data consumption in the Malaysian context as well as the developed system with the visualization features to address crime data and the eventual goal of combating the crimes .Sutapat Thiprungsri [4] examines the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. The purpose of his study is to examine the use of clustering technology to automate fraud filtering during an audit. He used cluster analysis to help auditors focus their efforts when evaluating group life insurance claims. A. Malathi et al.[5] look at the use of missing value and clustering algorithm for a data mining approach to help predict the crimes patterns and fast up the process of solving crime. Malathi. A et. al.[6] used a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years.Dr. S. Santhosh Baboo and Malathi. A [7] research work focused on developing a crime analysis tool for Indian scenario using different data mining techniques that can help law enforcement department to efficiently handle crime investigation.

The proposed tool enables agencies to easily and economically clean, characterize and analyze crime data to identify actionable patterns and trends .Kadhim B. Swadi Al-Janabi [8] presents a proposed framework for the crime and criminal data analysis and detection using Decision tree Algorithms for data classification and Simple K Means algorithm for data clustering. The paper tends to help specialists in discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identifying possible suspects. Aravindan Mahendiran et al. [9] apply myriad of tools on crime data sets to mine for information that is hidden from human perception. With the help of state of the art visualization techniques we present the patterns discovered through our algorithms in a neat and intuitive way that enables law enforcement departments to channelize their resources accordingly. Sutapat Thiprungsri[10] examine the possibility of using clustering technology for auditing. Automating fraud filtering can be of great value to continuous audits. The objective of their study is to examine the use of cluster analysis as an alternative and innovative anomaly detection technique in the wire transfer system. K. Zakir Hussain et al. [11] tried try to capture years of human experience into computer models via data mining and by designing a simulation model.

IV. PROPOSED SYSTEM

Architecture

After literature review there is need to used an open source data mining tool which can be implemented easily and analysis can be done easily. So here crime analysis is done on crime dataset by applying k means clustering algorithm using rapid miner tool.

The procedure is given below:

1. First we take crime dataset
2. Filter dataset according to requirement and create new dataset which has attribute according to analysis to be done
3. Open rapid miner tool and read excel file of crime dataset and apply “Replace Missing value operator” on it and execute operation
4. Perform “Normalize operator” on resultant dataset and execute operation
5. Perform k means clustering on resultant dataset formed after normalization and execute operation
6. From plot view of result plot data between crimes and get required cluster
7. Analysis can be done on cluster formed.

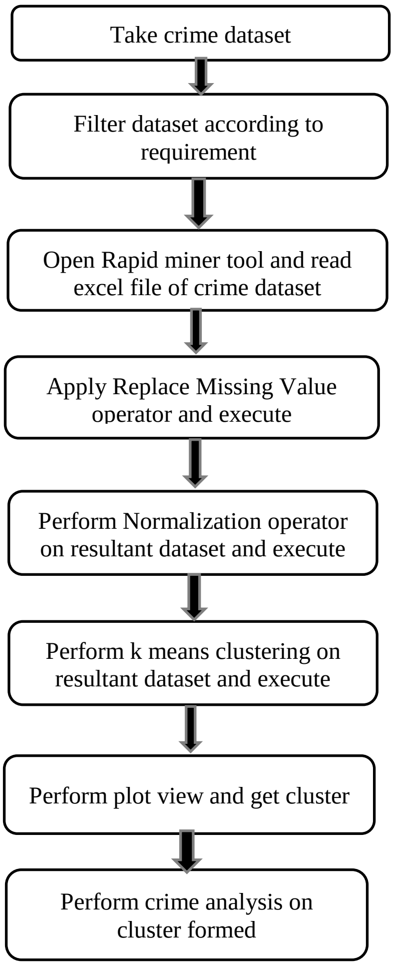
Take crime dataset
Filter dataset according to requirement
Open Rapid miner tool and read excel file of crime dataset
Apply Replace Missing Value operator and execute
Perform Normalization operator on resultant dataset and execute
Perform k means clustering on resultant dataset and execute
Perform plot view and get cluster
Perform crime analysis on cluster formed

Fig 1: Flow chart of crime analysis

V. CONCLUSION AND FUTURE SCOPE

This project focuses on crime analysis by implementing clustering algorithm on crime dataset using rapid miner tool and here we do crime analysis by considering crime homicide and plotting it with respect to year and got into conclusion that homicide is decreasing from 1990 to 2011 .From the clustered results it is easy to identify crime trend over years and can be used to design precaution methods for future

From the encouraging results, we believe that crime data mining has a promising future for increasin the effectiveness and efficiency of criminal and intelligence analysis. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern. As we have applied clustering technique of data mining for crime analysis we can also perform other techniques of data mining such as classification. Also we can perform analysis on various dataset such as enterprise survey dataset, poverty dataset, aid effectiveness dataset, etc.

VI.REFERENCES

[1] J. Agarwal, R. Nagpal, and R. Sehgal, ―Crime analysis using k-means clustering,‖ International Journal of Computer Applications, Vol. 83 – No4, December 2013.

[2] J. Han, and M. Kamber, ―Data mining: concepts and techniques,‖ Jim Gray, Series Editor Morgan Kaufmann Publishers, August 2000.

[3] P. Berkhin, ―Survey of clustering data mining techniques,‖ In: Accrue Software, 2003.

[4] W. Li, ―Modified k-means clustering algorithm,‖ IEEE Congress on Image and Signal Processing, pp. 616- 621, 2006.

[5] D.T Pham, S. Otri, A. Afifty, M. Mahmuddin, and H. Al-Jabbouli, ―Data clustering using the Bees algorithm,‖ proceedings of 40th CRIP International Manufacturing Systems Seminar, 2006.

[6] J. Han, and M. Kamber, ―Data mining: concepts and techniques,‖ 2nd Edition, Morgan Kaufmann Publisher, 2001.

[7] S. Joshi, and B. Nigam, ―Categorizing the document using multi class classification in data mining,‖ International Conference on Computational Intelligence and Communication Systems, 2011.

[8] T. Phyu, ―Survey of classification techniques in data mining,‖ Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol. IIMECS 2009, March 18 - 20, 2009, Hong Kong.

[9] S.B. Kim, H.C. Rim, D.S. Yook, and H.S. Lim, ―Effective Methods for Improving Naïve Bayes Text Classifiers,‖ In Proceeding of the 7th Pacific Rim International Conference on Artificial Intelligence, Vol.2417, 2002.

[10] S. Sindhiya, and S. Gunasundari, ―A survey on Genetic algorithm based feature selection for disease diagnosis system,‖ IEEE International Conference on Computer Communication and Systems(ICCCS), Feb 20-21, 2014, Chermai, INDIA.

[11] P. Gera, and R. Vohra, ―Predicting Future Trends in City Crime Using Linear Regression,‖ IJCSMS (International Journal of Computer Science & Management Studies) Vol. 14, Issue 07Publishing Month: July 2014.

[12] L. Ding et al., ―PerpSearch: an integrated crime detection system,‖ 2009 IEEE 161-163 ISI 2009, June 8-11, 2009, Richardson, TX, USA.

[13] K. Bogahawatte, and S. Adikari, ―Intelligent criminal identification system,‖ IEEE 2013 The 8th International Conference on Computer Science & Education (ICCSE 2013) April 26-28, 2013. Colombo, Sri Lanka.

[14] A. Babakura, N. Sulaiman, and M. Yusuf, ―Improved method of calssification algorithms for crime prediction,‖ International Symposium on Biometrics and Security Technologies (ISBAST) IEEE 2014. [15] S. Sathyadevan, and S. Gangadharan, ―Crime analysis and prediction using data mining,‖ IEEE 2014.