K-Means Clustering Method for Predicting and Analyzing Crime Data

Mr. Aniket Dikshit, Dr. M.M. Bokare, Dr S.B.Thorat

ABSTRAT:

In India, the crime rate is increasing each day. In the current situation, recent technological influence, effects of social media and modern approaches help the offenders to achieve their crimes. Both analysis and prediction of crime is a systematized method that classifies and examines the crime patterns. There exist various clustering algorithms for crime analysis and pattern prediction but they do not reveal all the requirements. Among these, K means algorithm provides a better way for predicting the results. The proposed research work mainly focused on predicting the region with higher crime rates and age groups with more or less criminal tendencies. We propose an optimized K means algorithm to lower the time complexity and improve efficiency in the result.

Keywords: K means, Crime datat, Machine learning,

INTRODUCTION:

In present scenario criminals are becoming technologically sophisticated in committing crime and one challenge faced by intelligence and law enforcement agencies is difficulty in analyzing large volume of data involved in crime and terrorist activities therefore agencies need to know technique to catch criminal and remain ahead in the eternal race between the criminals and the law enforcement. So appropriate field need to chosen to perform crime analysis and as data mining refers to extracting or mining knowledge from large amounts of data, data mining is used here on high volume crime dataset and knowledge gained from data mining approaches is useful and support police forces.

Research Scholar MCA, Mphil[1] ,Assistant Professor[2], Director[3]
Department of Computer Science
SSBES ITM College Nanded
bokaremadhav@yahoo.com[2], suryakant_thorat@yahoo.com[3]

To perform crime analysis appropriate data mining approach need to be chosen and as clustering is an approach of data mining which groups a set of objects in such a way that object in the same group are more similar than those in other groups and involved various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. In this paper k means clustering technique of data mining used to extract useful information from the high volume crime dataset and to interpret the data which

assist police in identify and analyze crime patterns to reduce further occurrences of similar incidence and provide information to reduce the crime. In this paper k mean clustering is implemented using open source data mining tool which are analytical tools used for analyzing data .Among the available open source data mining suite such as R, Tanagra ,WEKA ,KNIME ,ORANGE ,Rapid miner's means clustering is done with the help of rapid miner tool which is an open source

statistical and data mining package written in Java with flexible data mining support options. Also for crime analysis dataset used is Crime dataset an offences recorded by the police in England and Wales by offence and police force area from 1990 to 2011-12 .In this paper homicide which is crime committed by human by killing another human is being analyzed .

LITERATURE SUREVY:

De Bruin ,et al describes the information about the Narrative reports and criminal records are stored digitally across individual police departments, enabling the collection of this data to compile a nation-wide database of criminals and the crimes they committed. The compilation of this data through the last years presents new possibilities of analyzing criminal activity through time. Augmenting the traditional, more socially oriented, approach of behavioral study of these criminals and traditional statistics, data mining methods like clustering and prediction enable police forces to get a clearer picture of criminal careers. This allows officers to recognize crucial spots in changing criminal behaviour

and deploy resources to prevent these careers from unfolding. Four important factors play a role in the analysis of criminal careers: crime nature, frequency, duration and severity. We describe a tool that extracts these from the database and creates digital profiles for all offenders. It compares all individuals on these profiles by a new distance measure and clusters them accordingly. This method yields a visual clustering of these criminal careers and enables the identification of classes of criminals. The proposed method allows for several user-defined parameters.

RELATED WORKS:

Manish Gupta1*, et al describes the information about has been an enormous increase in the crime in the recent past. Crime deterrence has become an upheaval task. The cops in their role to catch criminals are required to remain convincingly ahead in the eternal race between law breakers and law enforcers. One of the key concerns of the law enforcers is how to enhance investigative effectiveness of the police. There is need for user interactive interfaces based on current

technologies to give them the much needed edge and fulfil the new emerging responsibilities of the police. The paper 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. The proposed interface 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. An interactive interface as crime analysis tool has been designed for this purpose.Sutapat Thirprungsri et al describes the information about this study examines the application of cluster analysis in the accounting domain, particularly discrepancy detection in audit. Cluster analysis groups data so that points within a single group or cluster are similar to one another and distinct from points in other clusters. Clustering has been shown to be a good candidate for anomaly detection. The purpose of this study is to examine the use of clustering technology to automate fraud filtering during an audit. We use cluster analysis to help auditors focus their efforts when evaluating group life insurance claims. Claims with similar characteristics have been grouped together and small-population clusters have been flagged for further investigation. Some dominant characteristics of those clusters which have been flagged are large beneficiary payment, large interest payment amounts, and long lag between submission and payment.A.Malathi ,Dr.S.Santhosh Baboo. et al describes the information about the Crime is a behavior disorder that is an integrated result of social, economical and environmental factors. Crimes are a social nuisance and cost our society dearly in several ways. Any research

that can help in solving crimes faster will pay for itself. In this paper we look at use of missing value and clustering algorithm for crime data using data mining. We will look at MV algorithm and Apriori algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. We applied these techniques to real crime data from a city police department. We also use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy.Malathi, A; Santhosh Baboo et al describes the information about about national security has increased after the 26/11 Mumbai attack. In this paper we 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. We will concentrate on MV algorithm and Apriori algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. We applied these techniques to real crime data. We also use semi- supervised learning technique in this paper for knowledge discovery from the crime records and to help increase the predictive accuracy.

Problem definition

Crime analysis tool is developed using various distinct data mining methods. It supports the police officers for investigating crimes [6]. Implementing a clustering algorithm on crime datasets enables analysis of crimes [7]. It makes identification and analysis of various criminality trends over the years through their conclusion. The random initial starting points produced by K-means which gives results in the form of cluster that helps in reaching the local optima [8]. So to overcome this problem, the partitioned data along with the data axis with the highest variance for assigning the

initial centroid for K-Means clustering was applied. So it is observed that the proposed technique uses a lesser number of iteration thereby reducing the clustering time. Using merge sort, K-means algorithm can be improved for clustering the Hidden Markov Model (HMM

DISADVANTAGES:

1. LESS ACCURACY

2. LOW EFFICIENCY

Methodology

We are working on Spyder for implementation. Here we use a Spyder 3.7 version. Spyder is an integrated development environment for systematic programming in Python. Here we implemented different packages like matplotlib,numpy,sklearn, pandas,etc. Which helps to plot elbow graph and data frame table using a K-means clustering algorithm? Dataset is collected from Kaggle datasets and import datasets into Spyder in CSV format as shown in Fig 1. We perform normalization for finding the accurate number of clusters (k) using the elbow method. The elbow method performs k-means clustering on the obtained dataset for a range of values of k (2-15) and calculates the SSE. A line chart of the SSE is plotted for each value of k

SYSTEM ARCHITECTURE:
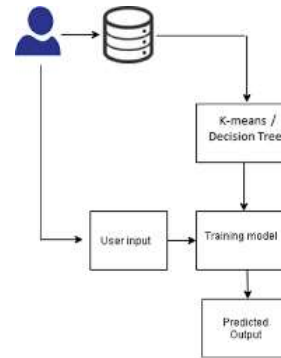
K-Means clustering

K-Means clustering is an unsupervised learning algorithm. There is no labeled data for this clustering, unlike in supervised learning. K-Means performs the division of objects into clusters that share similarities and are dissimilar to the objects belonging to another cluster.

The term 'K' is a number. You need to tell the system how many clusters you need to create. For example, K = 2 refers to two clusters. There is a way of finding out what is the best or optimum value of K for a given data.

Working of K-Means Algorithm

We can understand the working of K-Means clustering algorithm with the help of following steps −

- **Step 1** − First, we need to specify the number of clusters, K, need to be generated by this algorithm.
- **Step 2** − Next, randomly select K data points and assign each data point to a cluster. In simple words, classify the data based on the number of data points.
- **Step 3** − Now it will compute the cluster centroids.

- **Step 4** − Next, keep iterating the following until we find optimal centroid which is the assignment of data points to the clusters that are not changing any more −

First, the sum of squared distance between data points and centroids would be computed.

Now, we have to assign each data point to the cluster that is closer than other cluster (centroid).

At last compute the centroids for the clusters by taking the average of all data points of that cluster.K-means follows **Expectation-Maximization** approach to solve the problem. The Expectation-step is used for assigning the data points to the closest cluster and the Maximization-step is used for computing the centroid of each cluster.

While working with K-means algorithm we need to take care of the following things −

- While working with clustering algorithms including K-Means, it is recommended to standardize the data because such algorithms use distance-based measurement to determine the similarity between data points.
- Due to the iterative nature of K-Means and random initialization of centroids, K-Means may stick in a local optimum and may not converge to global optimum. That is why it is recommended to use different initializations of centroids.

Results

In above screen training is completed and then we got all dataset details and now click on 'Logout' link to get below screen

In above screen select state and district name and then enter details of crime and then press 'Submit' button to get below output

In above screen in blue colour we got output as 'Cuddapah is the Low Crime Area' and similarly we can test any other state

In above screen I entered some other state and crime rate and press button to get below output

In above screen in blue colour text we got output as 'Goa is High Crime area' and now click on 'Future Prediction' link to get below screen

In above screen I selected state, district and then select crime as 'Murder' and then press 'Submit' button to get below output

In above screen future predicted Murders for Gujarat state and Ahmadabad district is 62 and similarly you can select option and get future prediction and now click on 'Analysis' link to get below screen

In above screen select the type of analysis and press button to get below graphs

CONCLUSION:

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.

FUTURE SCOPE:

From the encouraging results, we believe that crime data mining has a promising future for increasing 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.

REFERENCES:

1] De Bruin ,J.S.,Cocx,T.K,Kosters,W.A.,Laros,J. and Kok,J.N(2006) Data mining approaches to criminal carrer analysis ,"in Proceedings of the Sixth International Conference on Data Mining (ICDM"06) ,Pp. 171-177

[2] Manish Gupta1*, B.Chandra1 and M. P. Gupta1,2007 Crime Data Mining for Indian Police Information System

[3] Nazlena Mohamad Ali1, Masnizah Mohd2, Hyowon Lee3, Alan F. Smeaton3, Fabio Crestani4 and Shahrul Azman Mohd Noah2 ,2010 Visual Interactive Malaysia Crime News Retrieval System

[4] Sutapat Thirprungsri Rutgers University .USA ,2011 Cluster Analysis of Anomaly Detection in Accounting Data : An Audit Approach 1

[5] A.Malathi ,Dr.S.Santhosh Baboo. D.G. Vaishnav College,Chennai ,2011 Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters.

[6] Malathi.A 1 ,Dr.S.Santhosh Baboo 2 and Anbarasi . A 31 Assistant professor ,Department of Computer Science ,Govt Arts College ,Coimbatore , India . 2 Readers , Department of Computer science , D.G. Vaishnav Collge ,Chennai , India , 2011 An intelligent Analysis of a city Crime Data Using Data Mining

[7] Malathi , A; Santhosh Baboo , S, 2011 An Enhanced Algorithm to Predict a Future Crime using Data Mining

[8] Kadhim B.Swadi al-Janabi . Department of Computer Science . Faculty of Mathematics and Computer Science .University of Kufa/Iraq , 2011 A Proposed Framework for Analyzing Crime DataSet using Decision Tree and Simple K-means Mining Algorithms.

[9] Aravindan Mahendiran, Michael Shuffett, Sathappan Muthiah, Rimy Malla, Gaoqiang Zhang,2011 Forecasting Crime Incidents using Cluster Analysis and Bayesian Belief Networks

[10] Sutapat Thiprungsri,2012 Cluster Analysis for Anomaly Detection in Accounting Data : An Audit Approach1

[11] K. Zakir Hussain, M. Durairaj and G. Rabia Jahani Farzana ,2012 Application of Data Mining Techniques for Analyzing Violent Criminal Behavior by Simulation Model

[12] https://www.gov.uk/government/publications/offences-recorded-by-the-police-in-england-and-wales-by-offence-and-police-force-area-1990-to-20

11-12