International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 08 | Aug-2016 www.irjet.net p-ISSN: 2395-0072

Data Mining Techniques used in Crime Analysis:- A Review

Navjot Kaur[1]

[1]*Student (M.Tech), Department of Computer Science, GNDU Regional Campus Gurdaspur, Punjab, India*

Abstract - *With the advent of the technology it is possible to analyze the crime which takes place in the various regions of the country. The survey of the various technologies used for this purpose is done in this paper. The technology will help in careful investigation of the crime and group them in the form of a cluster. By analyzing the crime models forecasting can be done and the impact of the crime in various regions can be stopped. There are various data development and acquisition tools available for this purpose. Various papers use various techniques for the data mining. The crimes can be partitioned into several parts. Some crimes are soft in nature and some crimes are hard in nature. In this paper all types of crime data mining techniques are analyzed and comprehensive comparison is presented.*

Key Words: Crime, cluster, Data Mining, Data Collection.

1.INTRODUCTION

Crime is the prime concern of this paper. Crime can take place at any time and at any region of the country. Most of the law enforcement agencies are focused on creating a tool through which future crime location can be detected. These tools are based on the large collection of data. This collection of data will be accomplished with the help of software tools. These tools will include Fuzzy System, WEKA etc.

1.1 Fuzzy System

This is one of the most commonly used tools for the data collection and filtering. The filtering mechanism will be accomplished by the use of rules of fuzzy. These rules are based upon the IF-THEN rules. The conditions are specified within the IF condition. The conditions if satisfied then result will be obtained. If the condition is false then the result will not be obtained. Fuzzy sets are created in this case. Membership functions are also needed to be defined. The membership of the given data will be done by determining whether membership function result in 0 or 1. Fuzzy system consist of input stage, processing stage and then output stage. The fuzzy rules will be represented as
If temperature is "low" THEN heater is "High"
Various truth values are used within the fuzzy system. These truth values are used in combination with AND, OR etc. if we require that both the conditions must be true then AND will be used. If either of the conditions can be true to produce the result then OR can be used.

1.2 WEKA

This tool is very common in order to gather the information about the particular topic. The advantage of using this tool is that clustering will be performed automatically. The information will be fed into this software and then all the relevant information will be produced efficiently using GUI. It is a collection of machine learning algorithm which will produce efficient result based upon the gathered information. It is open source software which is used for gathering and displaying information. The following table shows the advancement in the WEKA tool. It is based upon the java platform

2. LITERATURE SURVEY

There are number of papers which we have analyzed in order to determine the technology used and data mining techniques. The review of literature will involve efficient and usable techniques such as fuzzy system and weka tool. [1] In the studied paper tool for forecasting the crime was created. The discrete choice model will be used in this case. The discrete choice model takes into account the choice of the criminal and location in order to forecast the crime. [2] In this paper a specific location Malaysia is considered. The crime location and mindset of the criminal will be considered in this case. Crime Forecasting is rarely used globally by police including Malaysia. In practice usually the police would target persons with their criminality and studying their strategy of implementing crime. The police will also monitor the current crime situation and will take necessary action when the crime index increases. Both of these scenarios require action taken after crime incurred. Therefore if crime forecasting can be adopted perhaps early crime prevention can be enforced. The aim of this study is to identify crime patterns in Kedah using univariate forecasting technique. Seventy six recent monthly data (January 2006 – April 2012) were obtained from IPK Alor Star with the permission from PDRM Bukit Aman. Exploratory Data Analysis (EDA) and adjusted decomposition technique were conducted in order to fulfill the objective of the study. The findings revealed that total crimes in Kedah were mainly contributed by type of property crime (80-85%) while violent crime has a small proportion only. Fortunately due to the productiveness of the police the property crime trend indicated curve declining pattern. [3]Special section of the crime forecasting is considered in this case. The crime forecasting is used so that crime can be controlled. No specific location is considered in this case. The

GIS(Geographical Information System) is used inorder to detect the location where crime is happened. [4]This is book in which crime analysis is conducted. The mentally of criminal is considered in this case. The help of crime forecasting is also considered in this case.

3. COMPARISON OF TECHNOLOGY

In the analyzed papers number of different technologies is used. The comparison of the technologies used is as follows:

LDA:

Linear Discriminant Analysis (LDA) is a supervised learning algorithm. LDA methods are used in statistics, pattern recognition and machine learning to find a linear combination of features. The idea behind LDA is simple, for each class to be identified, calculate linear function of the attributes. The class function having highest score is treated as the predicted class. It is a statistical classification algorithm which is used to classify the values based on the linear combination among values. Linear Discriminant Analysis perfectly handles the data within class frequencies are unequal. LDA also evaluates the performances for randomly generated test data. The LDA Algorithm maximizes the ratio of between-class variance to the within-class variance in any particular dataset there by guaranteeing maximal separability. The use of Linear Discriminant Analysis for data classification is widely used to classify many biological datasets such cancer, colon cancer, HIV analysis etc. In LDA based classification the datasets can be transformed and test vectors can be classified in two different approaches.

Transformation with class dependency: This type of approach involves maximizing the ratio of between class variance to within class variance. The main objective is to maximize this ratios of that adequate class separability is obtained. The class-specificty approach involves using two optimizing criteria for transforming the datasets independently. Transformation with class independency: This approach involves maximizing the ratio of overall variance to within class variance. This approach uses only one optimizing criterion to transform the datasets and hence all data points irrespective of their class identity are transformed using this transform. In this type of LDA, each class is considered as a separate class against all other classes.

SVM:

Support Vector Machine is a type of classification method, which estimates the classification function. SVM is a set of related supervised learning methods that analyze data and recognize patterns, used for classification. Support Vector Machine(SVM) is a non-linear classifier method which is often reported as producing better classification results compared to other methods. The main idea of SVM is to construct a hyperplane as a decision surface in such a way that the margin of separation between positive and negative examples is maximized. This process non-linearly map the input sample data to some high dimensional space, where the data can be linearly separated, thus providing higher classification(or regression)accuracy. SVMs are rather interesting in that they enjoy both a sound theoretical basis as well as state-of-the-art success in real-world applications, especially in Bioinformatics.

k-NN:

It is the nearest neighbour algorithm. The k-nearest neighbour's algorithm is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms. The algorithm operates on a set of d-dimensional vectors, D={xi|i=1...N}, where xi∈kd denotes the ith data point. The algorithm is initialized by selection k points in kdas the initial k cluster representatives or "centroids". Techniques for select these primary seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data k times. Then the algorithm iterates between two steps till junction:

Step1: Data Assignment. Each data point is assign to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step2: Relocation of "means". Each group representative is relocating to the center (mean) of all data points assign to it. If the data points come with a possibility measure(Weights), then there location is to the expectations(weighted mean)of the data partitions. "Kernelize" k-means though margins between clusters are still linear in the embedded high-dimensional space, they can become non-linear when projected back to the original space, thus allowing kernel k-means to deal with more complex clusters. The k-medoid algorithm is similar to k-means except that the centroids have to belong to the dataset being clustered. Fuzzy c-means is also similar, except that it computes fuzzy membership functions for each clusters rather than a hard one.

Bayesian Belief network:

Bayesian belief networks are often used to model domains that are characterized by inherent uncertainty[7]. Formally, a Bayesian belief network can be defined as a directed acyclic graph with the following properties:

•The nodes represent random variables, and the edges represent probabilistic dependencies between variables.
•These dependences are quantified through a set of conditional probability tables.
•Each variable is assigned a conditional probability table of the variable given its
An important concept for Bayesian belief networks is conditional independence. Two sets of variables, A and B; are

said to be (conditionally) independent given a third set C of variables if when the values of the variables C are known, knowledge about the values of the variables B provides no further information about the values of the variables A : By testing this technique on a crime dataset of a serial crime happened in Gansu, China, can successfully capture the offender's intentions and locate the neighbourhood of the next crime scene parents. This technique is also used in prediction of crime like burglary prediction.

Artificial Neural network:

A neural network is a circuit composed of a very large number of processing elements that are called Neuron. Each element works only on local information. Furthermore each element operates non parellerliy, thus there is no system clock. Predicting the geo-temporal variations of crime and disorder. This technique introduces for crime incident prediction by concentrate on geographical areas of concern that outshine traditional policing boundaries. The computerized procedure employ a geographical crime incidence-scanning algorithm to identify clusters with relative high levels of crime hot spots. This collection gives sufficient data for training artificial neural networks (ANNs) capable of modeling trends within them.

Fuzzy clustering based technique:

In fuzzy clustering data elements can belong to more than one cluster. With each element there is an associated membership level. These indicate the strength of association between data elements and cluster[8]. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters. Most widely used is fuzzy c-means clustering. The fuzzy c-means algorithm is very similar to the k-means algorithm. With fuzzy c-means, the centroid of cluster is computed by:

•Choose a number of clusters.
•Assign randomly to each point coefficients for being in the clusters.
•Repeat until the algorithm has converged (that is, the coefficients change between two iterations is no more than threshold values, the given sensitivity threshold) :
•Compute the centroid for each cluster.
•For each point, compute its coefficients of being in the clusters.

The overall procedure consists of three main steps:
1. clustering the raw data.
2. extract the membership functions from the data.
3. create the fuzzy inference [8] system.

Table -1: Comparison of techniques

S.no.	Approach	Concept	Predictive accuracy	Performance
1.	BBN	Based on Bayesian theorem	Accuracy is good	Provide good performance. Combine prior knowledge with observed data.
2.	Artificial Neural network	Information processing occurs at processing elements	Prediction accuracy is high	Fast evaluation of learned target function
3.	Fuzzy based system	Based on fuzzy logic	prediction accuracy is very high	Perform better in time space condition.

4. CONCLUSION

Crime is a problem which must be tackled. Large number of peoples is affected by crime. Crime forecasting and extracting useful information from large amount of crime data hence is very important. If the advanced prediction about the problem can be made than crime may be stopped. If not stopped then can be reduced. Large amount of work is being done toward this area. But still some work can be done to improve the forecasting system. A survey is conducted so that Crime forecasting can be improved by the use of efficient data collection and data mining strategies.

REFERENCES

[1] M. B. Mitchell, D. E. Brown, and J. H. Conklin, "A Crime Forecasting Tool for the Web-Based Crime Analysis Toolkit," 2007 IEEE Syst. Inf. Eng. Des. Symp.,pp.1–5,2007.
[2] S. Ismail and N. Ramli,"Short-Term Crime Forecast ting In Keedah," Procedia-Soc. Behav. Sci., vol.91,pp.654–660,2013.
[3] W. Gorr and R. Harries, "Introduction to crime forecasting," vol. 19, pp. 551–555, 2003.
[4] K. H. Vellani, "Crime Analysis: for Problem Solving Security Professionals in 25 Small Steps," Facilities,pp.1–56,2010.
[5] W. Gorr, A. Olligschlaeger, and Y. Thompson, "Short-term forecasting of crime," Int. J. Forecast., vol. 19, no. 4, pp. 579–594, 2003.
[6] E. R. Groff and N. G. La Vigne, "Forecasting the Future of Predictive Crime Mapping," Crime Prev. Stud., vol. 13, pp. 29–57, 2002.
[7] J. V Pepper, "Forecasting Crime: A City Level Analysis," Underst. Crime Trends Work. Rep., pp. 177–210,2007.
[8] Giles C. Oatley, Brian W. Ewart, "Crimes analysis software: 'pins in maps', clustering and Bayes net prediction", Expert Systems with Applications, Vol. 25, 2003,pp. 569–588.

[9] Tony H. Grubesic, "On The Application of Fuzzy Clustering for CrimeHot Spot Detection", Journal of Quantitative Criminology, Vol.22, No.1, 2006, pp. 229-240