Global Journal of Computer Science and Technology
Volume 11 Issue 11 Version 1.0 July 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters

By A. Malathi, Dr. S. Santhosh Baboo

D. G. Vaishnav College, Chennai

Abstracts - 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.

Keywords : *Crime-patterns, clustering, data mining, law-enforcement, Apriori.*

GJCST Classification : *H.2.8*

Strictly as per the compliance and regulations of:

Algorithmic Crime Prediction Model Based on the Analysis of Crime Clusters

A. Malathi[α], Dr. S. Santhosh Baboo[Ω]

Abstract - 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.

Index Terms : *Crime-patterns, clustering, data mining, law-enforcement, Apriori.*

I. Introduction

Security is one of the major concerns and the issue is continuing to grow in intensity and complexity. Security is an aspect that is given top priority by all political and government worldwide and are aiming to reduce crime incidence (David, 2006). Reflecting to many serious situations like September 11, 2001 attack, Indian Parliament Attack, 2001, Taj Hotel Attack, 2006 and amid growing concerns about theft, arms trafficking, murders, the importance for crime analysis from previous history is growing. The law enforcement agencies are actively collecting domestic and foreign intelligence to prevent future attacks.

Criminology is an area that focuses the scientific study of crime and criminal behavior and law enforcement and is a process that aims to identify crime characteristics. It is one of the most important fields where the application of data mining techniques can produce important results. Crime analysis, a part of criminology, is a task that includes exploring and detecting crimes and their relationships with criminals.

The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Identifying crime characteristics is the first step for developing further analysis. The knowledge that is gained from data mining approaches is a very useful tool which can help and support police forces (Keyvanpour et al., 2010). According to Nath (2007), solving crimes is a complex task that requires human intelligence and experience and data mining is a technique that can assist them with crime detection problems. The idea here is to try to capture years of human experience into computer models via data mining.

In the present scenario, the criminals are becoming technologically sophisticated in committing crimes (Amarnathan, 2003). Therefore, police needs such a crime analysis tool to catch criminals and to remain ahead in the eternal race between the criminals and the law enforcement. The police should use the current technologies (Corcoran *et al.*, 2003; Ozkan, 2004) to give themselves the much-needed edge. Availability of relevant and timely information is of utmost necessity in conducting of daily business and activities by the police, particularly in crime investigation and detection of criminals. Police organizations everywhere have been handling a large amount of such information and huge volume of records. There is an urgent need to analyzing the increasing number of crimes as approximately 17 lakhs Indian Penal Code (IPC) crime, and 38 lakhs local and Special Law crimes per year.

An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. However, in the present scenario, the following major challenges are encountered.

- Increase in the size of crime information that has to be stored and analyzed.
- Problem of identifying techniques that can accurately and efficiently analyze this growing volumes of crime data
- Different methods and structures used for recording crime data.
- The data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult.
- Investigation of the crime takes longer duration due to complexity of issues

All the above challenges motivated this research work to focus on providing solutions that can enhance the process of crime analysis for identifying and reducing crime in India. The main aim of this research work consist of developing analytical data mining methods that can systematically address the

Author α : Assistant Professor, Post Graduate and Research Department,Coimbatore. E-mail : malathi.arunachalam@yahoo.com.
Author Ω : Reader, Post Graduate and Research Department, D. G. Vaishnav College, Chennai. E-mail : santhos2001@sify.com.

complex problem related to various form of crime. Thus, the main focus is to develop a crime analysis tool that assists the police in

- To analyse the crime
- Provide information to predict the crime
- Identify and analyze common crime patterns to reduce further occurrences of similar incidence
- Provide the information to reduce the crime.

The present research work proposes the use of an amalgamation of data mining techniques that are linked with a common aim of developing such a crime analysis tool. For this purpose, the following specific objectives were formulated.

- To develop a data cleaning algorithm that to clean the dataset
- To explore and enhance clustering algorithms to identify crime patterns from historical data
- To explore and enhance classification algorithms to predict future crime behavior based on previous crime trends
- To develop anomalies detection algorithms to identify change in crime patterns

II. Preprocessing

Real world data are usually incomplete, noisy, inconsistent and missing. And such data may cause confusion for the knowledge discovery process. Data cleaning is mandatory to improve the quality of the data. Preprocessing is a routine task that usually consumes much of the efforts exerted in the entire data mining process.

Data preprocessing is a process that consists of data cleaning, data integration and data transformation which is usually processed by a computer program. It intends to reduce some noises, incomplete and inconsistent data. The results from preprocessing step can be later proceeding by data mining algorithm.

The dataset used in experiment contains various items like year, state code, status of administrative unit, name of the administrative unit, number of crimes with respect to murder, dacoity, riots and Arson, area in sq. meters of the administrative unit, Estimated Mid-Year Population of the Administrative Unit in 1000s (begins in 1964), Actual Civil Police Strength (numbers of personnel), Actual Armed Police Strength (numbers of personnel) and Total Police Strength (Civil and Armed Police). . The cleaning of data is performed in two steps.

Step 1: Implement a Iterative Match_or_Delete routine that removes data records that are not important for analysis

Step 2: Implement a missing handling procedure that fills in missing data items or records in the crime dataset.

The crime dataset used in experiment contains various items like year, state code, status of administrative unit, name of the administrative unit, number of crimes with respect to murder, dacoity, riots and Arson, Estimated Mid-Year Population of the Administrative Unit in 1000s (begins in 1964), Actual Civil Police Strength (numbers of personnel), Actual Armed Police Strength (numbers of personnel) and Total Police Strength (Civil and Armed Police).

The experiment concentrates on only those attributes that are related to crime data that is year, state, administrative name, number of crimes. The quality of the results of the mining process is directly proportional to the quality of the preprocessed data. Careful scrutiny revealed that the dataset have missing data in state and number of crimes attributes. There are a number of methods for treating records that contain missing values.

1. Omit the incorrect fields(s)
2. Omit the entire record that contains the incorrect field(s)
3. Automatically enter / correct the data with default values (e.g.) select the mean from the range
4. Derive a model to enter/correct the data
5. Replace all values with a global constant
6. Use imputation method to predict missing values.

During preprocessing, focus was only on those attributes that are related to crime data. On careful scrutiny, it was found that the following three attributes have missing data which will affect the crime analysis process.

(i) State Attribute
(ii) Four fields reporting the number of crimes occurred
(iii) Size of population of the city Attribute

An 'Iterative Match_or_Delete' approach that uses of an external pre-build codebook for handling missing values is proposed. While considering filling missing number of crimes related murder, dacoity, riots and arson, an enhanced KNN-based imputation method combined with Learning Vector Quantization (LVQ) is proposed. The empty values in size of population of a city, a novel hybrid model that combines EM algorithm and naïve bayes classification algorithm is used to predict the missing values.

III. Identification of Crime Zones Using Clustering Techniques

Given a set of objects, clustering is the process of class discovery, where the objects are grouped into clusters and the classes are unknown beforehand. Two clustering techniques, K-means and DBScan (Density-Based Spatial Clustering Application with Noise)

algorithm are considered for this purpose. The algorithm for k-means is given below.

The HYB algorithm is given below.

The HYB algorithm clusters the data m groups where m is predefined

Input – Crime type, Number of Clusters, Number of Iteration.

Initial seeds might produce an important role in the final result.

Step 1 : Randomly Choose cluster centers;
Step 2 : Assign instances to clusters based on their distance to the cluster centers
Step 3 : centers of clusters are adjusted
Step 4 : go to Step 1 until convergence
Step 5 : Output C0, C1, C2, C3

From the clustering result, the city crime trend for each type of crime was identified for each year. Further, by slightly modifying the clustering seed, the various states were grouped as high crime zone, medium crime zone and low crime zone. From these homogeneous groups, the efficiencies of police administration units i.e. states can be measured and the method used is given below.

Output Function of Crime Rate = 1/Crime Rate

Here, crime rate is obtained by dividing total crime density of the state with total population of that state since the police of a state are called efficient if its crime rate is low i.e. the output function of crime rate is high.

Thus the two clustering techniques were analyzed in their efficiency in forming accurate clusters, speed of creating clusters, efficiency in identifying crime trend, identifying crime zones, crime density of a state and efficiency of a state in controlling crime rate. Experimental results showed that HYB algorithm show improved results when compared with k-means algorithm and therefore was used in further investigations.

Prediction of Crime Trend

The next task is the prediction of future crime trends. This involves tracking crime rate changes from one year to the next and used data mining to project those changes into the future. The basic method involves cluster the states having the same crime trend and then using "next year" cluster information to classify records. This is combined with the state poverty data to create a classifier that will predict future crime trends.

The Major crimes under property crime are discussed here. There are many categories of crimes like Crime against women, property crime, Road Accident.

- Murder
- Murder for Gain Dacoity
- Robbery
- Burglary
- Theft

To the clustered results, a classification algorithm was applied to predict the future crime pattern. The classification was performed to find in which category a cluster would be in the next year. This allows us to build a predictive model on predicting next year's records using this year's data. The C4.5 decision tree algorithm was used for this purpose. The generalized tree was used to predict the unknown crime trend for the next year. Experimental results proved that the technique used for prediction is accurate and fast. The following are four different clusters produced depends upon the crime nature

- C0 : Crime is steady or dropping. Theft is the primary crime little increased and dropping.
- C1 : Crime is rising or in flux. Dacoity is the primary crime rates changing..
- C2 : Crime is generally increasing. Robbery, Murder, Murder for gain, and Burglery are the primary crime on the rise.
- C3 : Few crimes are in flux. Dacoity is in flux. It has gone down and increased then once again gone down.

IV. Implementation

Major two crimes Theft and Robbery were taken to analyse the existing crime. Crime theft was in flux, In the year 2007 it got decreased, in 2008, 2009 it got increased and 2010 once again it decreased. Crime robbery kept increasing from 2007 to 20010.

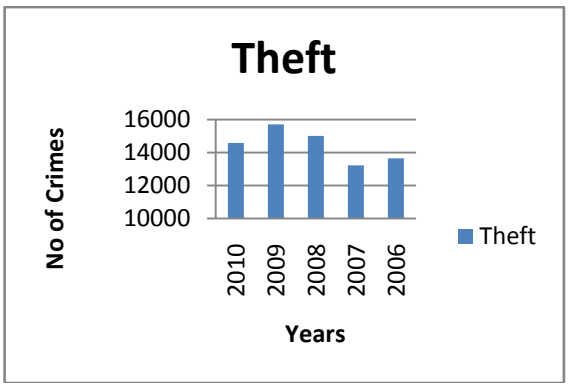

Fig. 1 : Crime Theft Analysis

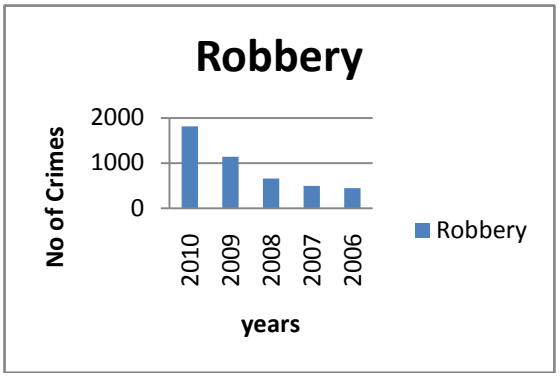

Fig. 2 : Crime Robbery Analysis

The robbery crime was taken to analyse the future crime prediction. This crime was analysed for period 2006 to 2009. Both existing algorithm and the new algorithm are executed for the same data set. The existing algorithm predicted the crime as 83%. The new algorithm predicted the crime as 89%.

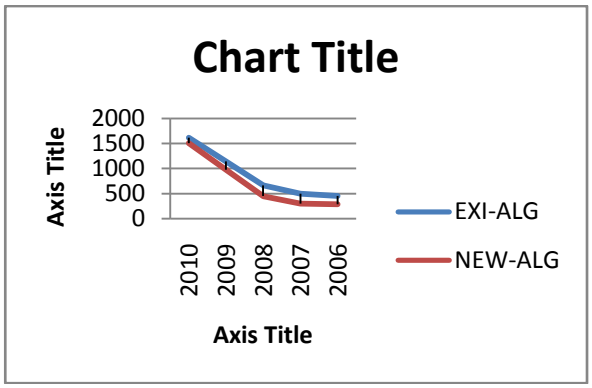

Fig. 3 : Crime prediction

V. Conlusion

A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. As information science and technology progress, sophisticated data mining and artificial intelligence tools are increasingly accessible to the law enforcement community. These techniques combined with state-of-the-art Computers can process thousands of instructions in seconds, saving precious time. In addition, installing and running software often costs less than hiring and training personnel. Computers are also less prone to errors than human investigators, especially those who work long hours.

This research work focus 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. The proposed tool, applied to crime data, can be used as a knowledge discovery tool that can be used to review extremely large datasets and incorporate a vast array of methods for accurate handling of security issues.

The development of the crime analysis tool has four steps, namely, data cleaning, clustering, classification and outlier detection. The data cleaning stage removed unwanted records and predicted missing values. The clustering technique is used to group data according to the different type of crime. From the clustered results it is easy to identify crime trend over years and can be used to design precaution methods for future. The classification of data is mainly used predict future crime trend. The last step is mainly used to identify future crimes that are emerging newly by using outlier detection on crime data.

Experimental results prove that the tool is effective in terms of analysis speed, identifying common crime patterns and future prediction. The developed tool has promising value in the current changing crime scenario and can be used as an effective tool by Indian police and enforcement of law organizations for crime detection and prevention.

References References Referencias

1. Amarnathan, L.C. (2003) Technological Advancement: Implications for Crime, The Indian Police Journal, April June.
2. Abraham, T. and de Vel, O. (2006) Investigative profiling with computer forensic log data and association rules," in Proceedings of the IEEE International Conference on Data Mining (ICDM'02), Pp. 11 – 18.
3. Brown, D.E. (1998) The regional crime analysis program (RECAP): A frame work for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, Pp. 2848-2853.
4. Corcoran J.J., Wilson I.D. AND Ware J.A. (2003) Predicting the geo-temporal variations of crime and disorder, International Journal of Forecasting, Vol. 19, Pp.623–634.
5. David, G. (2006) Globalization and International Security: Have the Rules of the Game Changed?, Annual meeting of the International Studies Association, California, USA, http://www.allacademic.com/meta/p98627_index.html.
6. de Vel, O., Anderson, A., Corney, M. and Mohay, G. (2001) Mining E-Mail Content for Author Identification Forensics, ACM SIGMOD Record, Vol. 30, No. 4, Pp. 55-64.
7. de Bruin, J.S. , Cocx, T.K. , Kosters, W.A. , Laros, J. and Kok, J.N. (2006) Data mining approaches to criminal career analysis," in Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Pp. 171-177.
8. Hauck, R.V.Atabakhsh, H., Ongvasith, P., Gupta, H. and Chen, H. (2002) Using Coplink to Analyze Criminal-Justice Data, Computer, Volume 35 Issue 3, Pp. 30-37.
9. Krishnamorthy, S. (2003) Preparing the Indian Police for 21st Century, Puliani and Puliani, Bangalore, India.
10. Keyvanpour, M.R., Javideh, M. and Ebrahimi, M.R. (2010) Detecting and investigating crime by means of data mining: a general crime matching framework, Procedia Computer Science, World Conference on Information Technology, Elsvier B.V., Vol. 3, Pp. 872-830.

11. Michelson, M. and Knoblock, C.A. (2006) Phoebus: A System for Extracting and Integrating Data from Unstructured and Ungrammatical Sources, Proceedings of AAAI.
12. Marshall, G.M. and Marshall, D.R. (2008) CRIME IN INDIA, Annual Series, 1954-2006, Published by the Government of India, Ministry of Home Affairs, National Crime Records Bureau , Electronic Dataset and Codebook, Published by Center for Systemic Peace.http://www.systemicpeace.org/inscr/inscr.htm
13. Nath, S. (2007) Crime data mining, Advances and innovations in systems, K. Elleithy (ed.), Computing Sciences and Software Engineering, Pp. 405-409.
14. Ozkan, K. (2004) Managing data mining at digital crime investigation, Forensic Science International, Vol. 146, Pp.S37-S38.
15. [SGW95] Senator, T.E., Goldberg, H.G., Wooton, J., Cottini, M.A., Khan, A.F.U., Klinger, C.D., Llamas, W.M., Marrone, M.P. and Wong, R.W.H. (1995) The FinCEN Artificial Intelligence System: Identifying Potential Money Laundering from Reports of Large Cash Transactions, AI Magazine, Vol.16, No. 4, Pp. 21-39.

This page is intentionally left blank