An Enhanced Algorithm to Predict a Future Crime using Data Mining

Malathi. A
Assistant Professor
Post Graduate and Research department of Computer Science, Government Arts College, Coimbatore, India

Dr. S. Santhosh Baboo
Reader, Post Graduate and Research Department of Computer Science, D.G Vaishnav College Chennai, India

ABSTRACT
Concern 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.

General Terms
Crime data mining, MV Algorithm, Apriori Algorithm

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

1. INTRODUCTION
We today, security are considered to be 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 complex problem related to various form of crime. Thus, the main focus is to develop a crime analysis tool that assists the police in

  - Detecting crime patterns and perform crime analysis

- Provide information to formulate strategies for crime prevention and reduction
- Identify and analyze common crime patterns to reduce further occurrences of similar incidence

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
  - cleans the crime dataset, by removing unwanted data
  - Use techniques to fill missing values in an efficient manner
- To explore and enhance clustering algorithms to identify crime patterns from historical data
- To explore and enhance classification algorithms to predict future crime behaviour based on previous crime trends
- To develop anomalies detection algorithms to identify change in crime patterns

2. INDIAN POLICE STRUCTURE

To propose any intelligent system (Michelson *et al*., 2006; Tuchinda *et al*., 2007) as crime analysis tool for police, it is required to understand Indian Police structure, responsibilities of the police, key changes and challenges the police is forcing (Krishnamorthy, 2003).

Superintendence over the police force in the State is exercised by the State Government. The head of the police force in the State is the Director General of Police (DGP). States are divided territorially into administrative units knows as districts. A group of districts form a range, which is looked after by an officer of the rank of Deputy Inspector General of Police (DIGP). Some States have zones comprising two or more ranges, under the charge of an officer of the rank of an Inspector General of Police (IGP). A Senior Superintendent of Police (SSP)/Superintendent of Police (SP) is the head of the district police administration and is assisted by an Assistant Superintendent of Police (ASP) and few Deputy Superintendents of Police (DSP). A district may have many Police Stations that are manned by Inspectors, Sub Inspectors, Assistant Sub Inspectors, Head Constables and Constables. Police Station is the basic unit of police administration through which both crime and non crime duties are discharged. Police Stations are the places where complaints and First Information Reports (FIRs) are lodged. Police Stations also serve as the window of 'citizen interface' for the police. Common people approach Police Stations for assistance. Therefore, public expectations from police stations are more direct, pressing and at times extremely demanding. Operationally, police stations are at the nucleus of all policing activities. All important operational duties – be it duties to the State or services to other government departments or citizens- are executed and coordinated through police stations (Manish 2006).

3. REVIEW OF LITERATURE

Data mining in the study and analysis of criminology can be categorized into main areas, crime control and crime suppression. Crime control tends to use knowledge from the analyzed data to control and prevent the occurrence of crime, while the criminal suppression tries to catch a criminal by using his/her history recorded in data mining.

Brown (1998) constructed a software framework called ReCAP (Regional Crime Analysis Program) for mining data in order to catch professional criminals using data mining and data fusion techniques. Data fusion was used to manage, fuse and interprets information from multiple sources. The main purpose was to overcome confusion from conflicting reports and cluttered or noisy backgrounds. Data mining was used to automatically discover patterns and relationships in large databases.

Crime detection and prevention techniques are applied to different applications ranging from cross-border security, Internet security to household crimes. Abraham et al. (2006) proposed a method to employ computer log files as history data to search some relationships by using the frequency occurrence of incidents. Then, they analyzed the result to produce profiles, which can be used to perceive the behavior of criminal.

De Bruin et. al. (2006) introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly. This method also provided a visual clustering of criminal careers and identification of classes of criminals.

From the literature study, it could be concluded that crime data is increasing to very large quantities running into zota bytes (1024bytes). This in turn is increasing the need for advanced and efficient techniques for analysis. Data mining as an analysis and knowledge discovery tool has immense potential for crime data analysis. As is the case with any other new technology, the requirement of such tool changes, which is further augmented by the new and advanced technologies used by criminals. All these facts confirm that the field is not yet mature and needs further investigations.

4. PREPROCESSING

A 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).

4.1 Missing value handling for state field

Some researchers use the statistics-based concept space algorithm to automatically associate different objects such as persons, organizations, and vehicles in crime records (Hauck *et*

al., 2002). Using link analysis techniques to identify similar transactions, the Financial Crimes Enforcement Network AI System (Senator *et al*., 1995) exploits Bank Secrecy Act data to support the detection and analysis of money laundering and other financial crimes. Classification finds common properties among different crime entities and organizes them into predefined classes. This technique has been used to identify the source of e-mail spamming based on the sender's linguistic patterns and structural features (de Vel *et al*., 2001)

The experiment concentrate on only those attributes that are related to crime data, that is year, state, administrative name, number of crimes for the years 1971 to 2006. 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 teh 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.

4.2 Missing value handling for number of crimes occurred attribute

In the present research work, while considering filling missing number of crimes related murder, dacoity, riots and arson, two methods were used. Initially, all the four fields are analyzed for empty values. If all the four attributes have empty values for a particular record, then the entire record is considered as irrelevant information and is deleted.

- While taking individual attributes into consideration, a novel KNN-based imputation method is proposed. In this method, the missing values of an instance are imputed by considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function.

- The MV new algorithm is as follows

1. Divide the data set D into two parts. Let D_m be the set containing the instances in which at least one of the features is missing. The remaining instances will complete feature information form a set called D_c.

2. For each vector x in D_m:

    a. Divide the instance vector into observed and missing parts as $x = [x_o, x_m]$.

    b. Calculate the distance between the x_o and all the instance vectors from the set D_c.

    c. Use only those features in the instance vectors from the complete set D_c, which are observed in the vector x.

    d. Use the P closest instances vectors and perform a majority voting estimate of the missing values for categorical attributes. For continuous attributes replace the missing value using the mean value of the attribute in the P (related instances)

The challenging decisions that have to be carefully chosen are:

(i) The choice of the distance function. In the present work, four distance measures, Euclidean, Manhattan, Mahalanobis and Pearson, are considered and the one that produced best result is considered.

(ii) The KNN algorithm searches through all the dataset looking for the most similar instances. This is a very time consuming process and it can be very critical in data mining where large databases are analyzed. To speed up this process a method that combines missing value handling process with classification is proposed.

(iii) The choice of k, the number of neighbors. Experiments showed that a value of 10 produce best results in terms of accuracy and hence is used in further experimentation.

Thus, the traditional KNN Imputation method was enhanced in two manners. The first enhancement is achieved by proposing a new distance metric and the second enhancement is achieved by using LVQ (Learning Vector Quantization) methods combined with generalized relevance learning to perform the classification and missing value treatment simultaneously. Both these enhancement when combined together produces a model (E-KDD) that is efficient in terms of speed and accuracy.

4.3 Missing value handling in the prediction of the size of Population of the city

The first task is the prediction of the size of the population of a city. The calculation of per capita crime statistics helps to put crime statistics into proportion. However, some of the records were missing one or more values. Worse yet, half the time, the missing value was the "city population size", which means there was no per capita statistics for the entire record. Over some of the cities did not report any population data for any of their records. To improve the calculation of "yearly average per capita crime rates", and to ensure the detection of all "per capita outliers", it was necessary to fill in the missing values. The basic approach to do this was to cluster population sizes, create classes from the clusters, and then classify records with unknown population sizes. The

justification for using clustering is as follows: Classes from clusters are more likely to represent the actual population size of the cities. The only value needed to cluster population sizes was the population size of each record. These values were clustered using EM algorithm and initially 10 clusters were chosen because it produced clusters with mean values that would produce per capita calculations close to the actual values.

5. 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 DBScan algorithm show improved results when compared with k-means algorithm and therefore was used in further investigations.

5.1 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. 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.

- C0: Crime is steady or dropping. The Sexual Harassment rate is the primary crime in flux. There are lower incidences of: Murder for gain, Dacoity, Preparation for Dacoity, rape, Dowry Death and Culpable Homicide.

- C1: Crime is rising or in flux. Riots, cheating, Counterfeit, and Cruelty by husband and relatives are the primary crime rates changing. There are lower incidences of: murder and kidnapping and abduction of others.

- C2: Crime is generally increasing. Thefts are the primary crime on the rise with some increase in arson. There are lower incidences of the property crimes: burglary and theft.

- C3: Few crimes are in flux. Murder, rape, and arson are in flux. There is less change in the property crimes: burglary, and theft. To demonstrate at least some characteristics of the clusters,

6. IMPLEMENTATION

From the Graph the prediction can be concluded as follows

Using the existing algorithm the crime is analysed from 2001 to 2008 years and predicted the crime for the year 2009. The predicted crime for the year 2009 is 84%. The same set of crime data is analyzed using the new algorithm for the year 2001 to 2008. The predicted crime for the year 2009 by the new algorithm is 86%. The number of crime predicted by the existing algorithm is 6450. The number of crime predicted by the new algorithm is 6800. The original crime happened in 2009 is 7650.

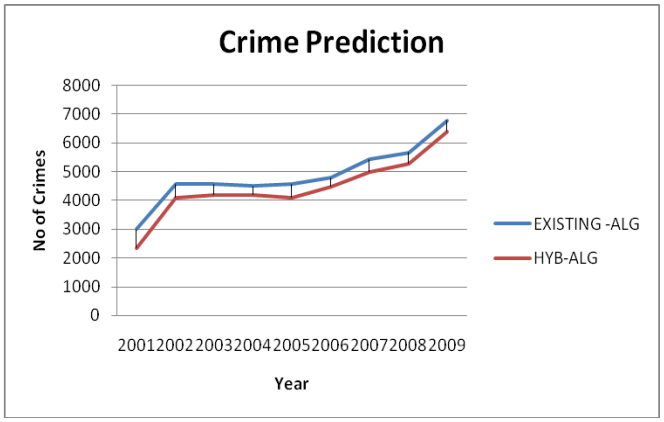

Fig 1: Crime Prediction

7. DATASET USED

The crime dataset used in the present research work was downloaded from the Integrated Network for Societal Conflict Research (INSCR) website (Marshall and Marshall, 2008). INSCR was established to coordinate and integrate information resources produced and used by the Center for Systemic Peace. The Indian data resources were prepared by researchers associated with the Center for Systemic Peace and are generated and/or compiled using open source information. These resources are made available as a service to the research community. All CSP data resources have been cross-checked with other data resources to ensure, as far as possible, that the information recorded is accurate and comprehensive.

The dataset has details regarding the crime in India and was posted and compiled by Marshall and Marshall (2008). The dataset covers the years 1954-2006 and contains records transcribed from the original materials published by the Government of India, Ministry of Home Affairs. The data is structured according to three distinct administrative levels: Federal State, State, and District (district level information begins in 1971). The levels can be separated by using the STATUS variable, which has three values namely, Federal State, State Territory and District. Records for the years 1965 and 1966 are missing; no copies of the Crime in India reports for those years have been located in US libraries. Records for the numbers of murders, dacoity, and riots in 1966 are recorded as they were reported in the 1967 edition (each annual edition, beginning in at least 1967, reports figures for the current and previous year for comparison purposes).

8. CONCLUSION

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.

9. REFERENCES

[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] 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.

[16] Tuchinda, R., Szekely P. and Knoblock, C.A (2007) Building Data Integration Queries by Demonstration, Proceedings of the 12th International Conference on Intelligent User Interfaces.

[17] Manish Gupta, B. Chandraand M. P. Gupta 2006, "Crime Data Mining for Indian Police Information System" in the 'Journal of Crime' Vol.2 Issue. 6 April 2006 pp.43-54