Malathi, A; Santhosh Baboo, S.
Enhanced Algorithms to Identify Change in Crime Patterns
International Journal of Combinatorial Optimization Problems and Informatics
Morelos, México

Available in: http://www.redalyc.org/articulo.oa?id=265219635004
Enhanced Algorithms to Identify Change in Crime Patterns

Malathi. A and Dr. S. Santhosh Baboo
Assistant Professor and Reader
Malathi.arunachalam@yahoo.com

Abstract. A major challenge facing all law-enforcement and intelligence gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. There has been an enormous increase in the crime in the recent past. The concern about national security has increased significantly since the 26/11 attacks. Crimes are a social nuisance and cost our society dearly in several ways. Here we look at use of missing value, clustering algorithm and Anomalies detection for a data mining approach to help predict the crimes patterns and speed up the process of solving crime. We will look at MV algorithm, DBScan and PAM outlier detection 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 can use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy.

Keywords: Data Mining, DBScan, Anomalies detection algorithm.

1. Introduction

Crime is a major issue where the top priority has given by our government. 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.

Law enforcement agencies like that of police today are faced with large volume of data that must be processed and transformed into useful information. 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 [9]. According to [13], 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.

Now-a-days, the criminals are becoming technologically sophisticated in committing crimes [1]. 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 [4] 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 analysing 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.

- Crime information volume has been increased.
- 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:

- To perform crime analysis to detect crime patterns.
- 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.
- 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.

2. Crime Prediction Framework

The data mining techniques used in the present research work are clustering, classification and deviation detection. Clustering techniques group data items into classes with similar characteristics to maximize or minimize intraclass similarity. For example, to identify suspects who conduct crimes in similar ways or distinguish among groups belonging to different gangs. These techniques do not have a set of predefined classes for assigning items. Some researchers use the statistics-based concept space algorithm to automatically associate different objects such as persons, organizations, and vehicles in crime records [3] [8]. Using link analysis techniques to identify similar transactions, the Financial Crimes Enforcement Network AI System [15] exploits Bank Secrecy Act data to support the detection and analysis of money laundering and other financial crimes. Clustering crime incidents can automate a major part of crime analysis but is limited by the high computational intensity typically required.

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 [2] [7]. Often used to predict crime trends, classification can reduce the time required to identify crime entities. However, the technique requires a predefined classification scheme. Classification also requires reasonably complete training and testing data because a high degree of missing data would limit prediction accuracy.

Deviation detection uses specific measures to study data that differs markedly from the rest of the data. Also called outlier detection, investigators can apply this technique to fraud detection, network intrusion detection, and other crime analyses. However, such activities can sometimes appear to be normal, making it difficult to identify outliers.

Apart from this, initially a preprocessing step that performs a cleaning process in two steps. The first step removes data records that are not important for analysis, while the second step implements a missing handling procedure to fill in missing data items or records in the crime dataset.

Normally, clustering of crime has a special meaning and refers to geographical grouping of crime. In the present work, it is used to group different kind of crime patterns. The result of clustering is then used by the classification process to predict crime trend. Further, it is also process to further identify anomalies from the crime trends. Thus cluster-based prediction crime analyzing tool is proposed in the present research work. The techniques and methods used in all these four steps are discussed in the following sections.

3. Data Preprocessing

Most of the data collection techniques like survey studies, field experiments, Crime findings, etc., produce huge amount of information, where missing values are inevitable. 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).

3.1. Missing value handling for state field

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 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.

3.2. Missing value handling for number of crimes occurred attribute

In the present research work, while considering the state attribute a novel method that uses a careful matching technique is proposed. Here, all the available unique name of the administrative unit and its corresponding status code are collected and stored as separate list (ADUNIT_STATE). The missing value process, when encountered with an empty value in the State field, performs an iterative search in the list ADUNIT_STATE and picks the corresponding state code for an administrative unit name. If no match is found, then the entire record is deleted.

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 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 $D_c$. Use only those features in the instance vectors from the complete set $D_c$, which are observed in the vector $x$.
   c) Use the $K$ closest instances vectors ($K$-nearest neighbors) 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 $K$-nearest neighborhood.

The advantages of using KNN imputation are:

1. $k$-nearest neighbor can predict both qualitative attributes (the most frequent value among the $k$ nearest neighbors) and quantitative attributes (the mean among the $k$ nearest neighbors).
2. It does not require creating a predictive model for each attribute with missing data. Actually, the $k$-nearest neighbor algorithm does not create explicit models.
3. It can easily treat instances with multiple missing values.
4. It takes in consideration the correlation structure of the data.

The challenging decisions that have to be carefully chosen are:

1. 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.
2. 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.
3. 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.

3.3. Missing value for 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.

4. 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 [17]:

$$V(D) = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - u_i)^2$$

(1)

Where $V$ is the variance, $S_i$ is a cluster, $u_i$ is its mean, $D$ is the dataset of all points $x_j$. Partition the dataset into $k$ clusters such that intracluster variance is minimized.

Time and Space Complexity

The primary part of the algorithm is the standard K-means algorithm. Let us assume that current partition of the $N$ $p$-dimensional point into $k$ clusters, Compute the distances from each and every point to every cluster centroid and reassign. So for the simplest case of squared Euclidean distance, at every iteration, there is $k$ computations of centroids, each one gets involved in $p$ arithmetic means.

- $k$ computations of centroids
  (Each of which involves $p$ arithmetic means.)

- $n*k$ distance computations,
  (Each of which involves $p$ sums of squared differences)

- $n$ minimums over $k$ distances

K-means algorithm often has hierarchical clustering using LINKAGE concepts. Hierarchical clustering needs $n \times n$ distance matrix, while K-means requires only the $n \times p$ data matrix, and $p$ is often much smaller than $n$. 

35
5. Anomalies Detection Algorithm to Identify Change in Crime Patterns

The previously used technique can be used to find the overall picture of crime trend for each state and whole country. This section presents another task, where unusual crime patterns are discovered. This can be achieved by using outlier detection algorithm, which can be used to highlight records that have an anomalous set of values. For this purpose, two kinds of data are analyzed. The first is the “per capita” data and second is the “crime rate”. The per capita data can identify states that have one or more years of unusually high crime rates for their size. The crime rate can be used to highlight increases in crime, thus identifying states with sudden increase in crime rate. When both are combined, they can be used by law enforcement people to target “root cause and prevention” projects. For this purpose, a hybrid outlier detection model that uses a partition-based clustering algorithm with a modified distance based algorithm that uses shapes for anomaly detection is used.

The basic approach to detect outliers is to find large non-globular clusters and define all the very small clusters outside of them as outliers. PAM (Partitioning Around Medoid) algorithm is used for this purpose. The PAM algorithm has the drawback of high computation complexity, which has a direct impact on the speed of anomaly detection. This complexity raised due to the huge number of computations required during anomaly detection.

In the present research work, this time complexity is reduced by considering shapes during anomaly detection. Two shapes, a circle and a square surrounding the circle, are considered. The main objective of using shapes is to reduce the number of computations and thus to increase the speed of the algorithm. This model is referred to as PDSAD (Partition and Distance Method based on Shape for Anomaly Detection) model. The working of PDSAD model is given below.

Initially, for a point ‘Q’, draw a circle with centre Q and radius D. Draw two squares that touches the interior (BoxI) and exterior (BoxE) boundaries of the circle. Count the number of points inside BoxI (SmallCount). If the number of points is greater than M, then q is not an anomaly. Else, count the number of points inside BoxE, outside circle (LargeCount) and store it in an array ‘A’. If the sum of SmallCount and LargeCount is lesser than M, then q is an anomaly. If this condition also fails, then the traditional distance calculation is performed on A to detect outliers and inliers.

Experiments were conducted to analyze the efficiency of outlier crime trend and the results proved that the proposed method is an efficient model for predicting irregular crime patterns.

6. Results and discussions

Experiments were conducted to analyze the efficiency of outlier crime trend and the results proved that the proposed method is an efficient model for predicting irregular crime patterns. 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 2010.

![Fig. 1. Crime Theft Analysis](image-url)
Fig. 2. Crime Robbery Analysis

The robbery crime was taken to analyse the future crime prediction. This crime was analysed for the 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

7. Conclusion

This paper focused on improving missing value handling procedures for efficient clustering and classification and to predict the future crime. 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.

The present work can be integrated with visualization techniques such as Geographical Information System for enhancing the understanding of the results and patterns. Further, the tool proposed was analyzed with the dataset available free for download from the Internet. In future, work is planned to include a data extraction module that can establish connection with government Crime Analysis Tool and extract records for further crime analysis.

References