An Improved Algorithm for Data Preprocessing in Mining Crime Data Set

Kadhim B. Swadi Aljanabi

Department of Computer Science
College of Mathematics and Computer Science
University of Kufa
kadhimbs@yahoo.com

Abstract

This paper presents an improved algorithm for data preprocessing to solve the problem of missing values and smoothing the outliers in the real world data sets. Previous works in this field are based mainly on replacing the missing values with the average, class average, most common values and some other techniques in the same direction, and outliers were generally cancelled from the data set. Crime and criminal data sets have their own special characteristics and benchmark in that missing values and outliers have different meanings than in other fields, so they need to be processed in different manners. The algorithm is based mainly on using clustering techniques to group the objects according to their similarities and dissimilarities, then smoothing the outliers accordingly and the missing values are processed according to their clusters. WEKA is used as a tool to find different clusters of the criminals.

Keywords: Data Mining, Clustering, Outliers, Missing values.

1.INTRODUCTION

Data mining represents one of the emerging field that can be used in a wide disciplinary of applications including marketing, banking, city planning, health insurance, and many other fields that highly affect the communities. Crime analysis is one of these important applications of data mining. Data Mining contains many tasks and techniques including Classification, Association, Clustering, Prediction, and Link Analysis. Each of them has its on importance and applications[1,2,3,4,5,6].

Advances in technology, which allow analyses of large quantities of data, are the foundation for the relatively new field known as *crime analysis.* Crime analysis is an emerging field in law enforcement without standard definitions. This makes it difficult to determine the crime analysis focus for agencies that are new to the field. In some police departments, what is called “crime analysis” consists of mapping crimes for command staff and producing crime statistics. In other agencies, crime analysis might mean focusing on analyzing various police reports and suspect information to help investigators in major crime units identify serial robbers and sex offenders.

Crime analysis is the act of analyzing crime. More specifically, crime analysis is the breaking up of acts committed in violation of laws into their parts to find out their nature and reporting Some analysts do all this and other types of analysis[5]. The role of the crime analyst varies from agency to agency.
statements of these findings. The objective of most crime analysis is to find meaningful information in vast amounts of data and disseminate this information to officers and investigators in the field to assist in their efforts to apprehend criminals and suppress criminal activity. Assessing crime through analysis also helps in crime prevention efforts [5, 7, 8].

2. WHY ANALYZE CRIME?

Crime Analysts usually tend to justify their existence as crime analysts in what is known as law enforcement agency. It is important to articulate some of the reasons it makes sense to analyze crime. Some good reasons are listed here [5]. There may be more other reasons depending on the community culture, geographic effects, and others

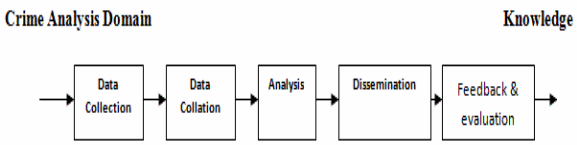
1. Analyze crime to inform law enforcers about general and specific crime trends, patterns, and series in an ongoing, timely manner.
2. Analyze crime to take advantage of the abundance of information existing in law enforcement agencies, the criminal justice system, and the public domain.
3. Analyze crime to maximize the use of limited law enforcement resources.
4. Analyze crime to have an objective means to access crime problems locally, regionally, statewide, nationally, and globally within and between law enforcement agencies.
4. Analyze crime to be proactive in detecting and preventing crime.
5. Analyze crime to meet the law enforcement needs of a changing society.

Analyze crime to understand the criminal behaviors.

In general there are four different techniques for analyzing crimes, they are

1. Linkage Analysis
2. Statistical Analysis
3. Profiling
4. Spatial Analysis

Each of the above techniques has its own advantages and drawbacks and can be used in specific cases. The four techniques use the steps shown in figure(1) in the analysis process:



Fig(1). Steps for Crime Analysis Process.

Data mining techniques can help discovery and exploitation of knowledge, which can aid in many aspects of knowledge management. Information on knowledge falls into three categories:

1. Knowledge about the past, which is stable, voluminous, and relatively accurate
2. Knowledge about the present, which is unstable, compact, and relatively inaccurate, and
3. Knowledge about the future, which is hypothetical.

- Data Mining, or Knowledge Discovery in Databases (KDD) in simple words is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1, 6, 8, 9, 10]. It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large Databases. KDD is the process of identifying a valid, potentially, useful and ultimately understandable structure in data. Spatial data mining methods are applied to extract interesting and regular knowledge from large spatial databases. In practice, data mining has two components: discovery and exploitation. During the discovery component, facts are discovered and represented as information-bearing data. During the exploitation component, these facts are applied to the solution of a specific problem. First, we discover; second, we act. The steps in the process are formulation of the problem, data evaluation, feature extraction and enhancement, prototyping and model evaluation. Crime Detection is an area of vital importance in Police Department. Crime rates are rapidly changing and improved analysis enables discerning hidden patterns of crime, if any, without any explicit prior knowledge of these patterns. In this background, a study is planned as per the following objectives:

- Extraction of crime patterns by analysis of available crime and criminal data.
- Prediction of a crime based on the spatial distribution of existing data and anticipation of crime rate using different data mining techniques.
- Detection of crimes.
- Predict the behavior of a criminal or groups of criminals according to their historical data with different attributes.

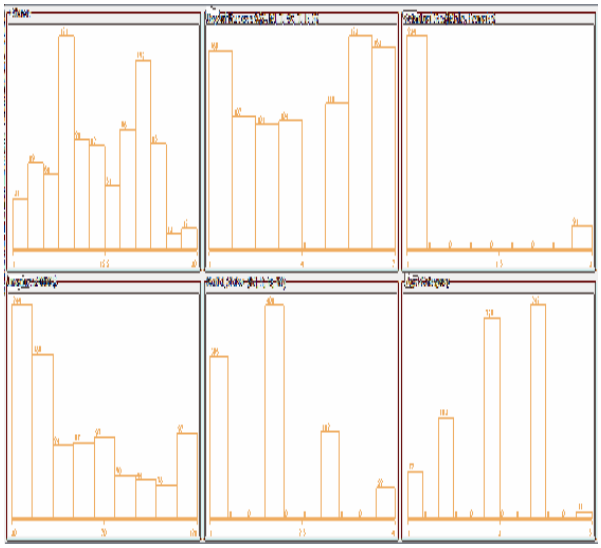
These approaches preprocess data to quickly generate relevant results, analyze patterns and co-occurrence of identified concept and develop an automated solution of crime pattern.

3. DATA COLLECTION

The dataset that was used as training and testing data were extracted from the Internet[11], these data contain data about both Crimes and Criminals with the following main attributes:

- CrimeID :Individual Crimes are designated by unique Crime IDs
- CrimeName :Disguised crime's name
- CrimeType :Indicates crime type.
- Day, date, time :indicate when a crime happened.
- CriminalID: Individual Criminals are designated by unique IDs
- Gender: Belongs to which gender.
- Age: Age of Individual criminal.
- Job: job of Individual Criminal.
- Location: Location of Individual criminal .
- Marital status of the criminal.

More than 600 criminal records and 1000 crime records were collected to test the proposed technique effectiveness. The distribution of the collected data is shown in figure(2) below.



Fig(2). Distribution of Offenses vs Criminal Attributes

4. Data preprocessing

Real world data usually have the following drawbacks: Incompleteness, Noisy and Inconsistence. So, these data need to be preprocessed to get the data suitable for analysis purposes, and the preprocessing includes the following tasks [1, 2 ,8, 9]:

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Different preprocessing techniques were used to get clean data, these include:

1. Removing outliers, some of the data is the crime and criminal datasets represent outliers and can not be included in the analysis algorithms and techniques, so these data records were deleted from the set.

2.Filling missing data, some criminal ages, jobs, and income were not mentioned in the tables, average and most commonly used values were used to substitute these missing values.

3.Data reduction using normalization and aggregation.

5. Proposed Techniques

Missing values and outliers in both crime and criminal data sets have absolutely different meanings than in many other data sets. For example unknown criminal address is different from that for an employee because crime analysis tends to discover and find out the unknowns, whereas in employment data set these unknowns can be cancelled to prepare data for mining algorithms, in addition to that outliers in crime and criminal data sets represent some types of knowledge needed to be explored and processed and needed to be focused on, for example increasing a specific crime “Theft” for example in 5% annually is expressed as a normal situation between adults whereas 0.005% increase of this type of crimes between children can be of great meaning, this percent may be expressed as outliers in applications other than crime analysis. The proposed technique depends mainly on finding out criminal clusters before

applying preprocessing algorithms for missing values and outliers, then these clusters will be used to smooth the outliers and replace the missing values. Two clustering techniques were used in this paper, Expectation Maximization (EM) and K-Means algorithms.

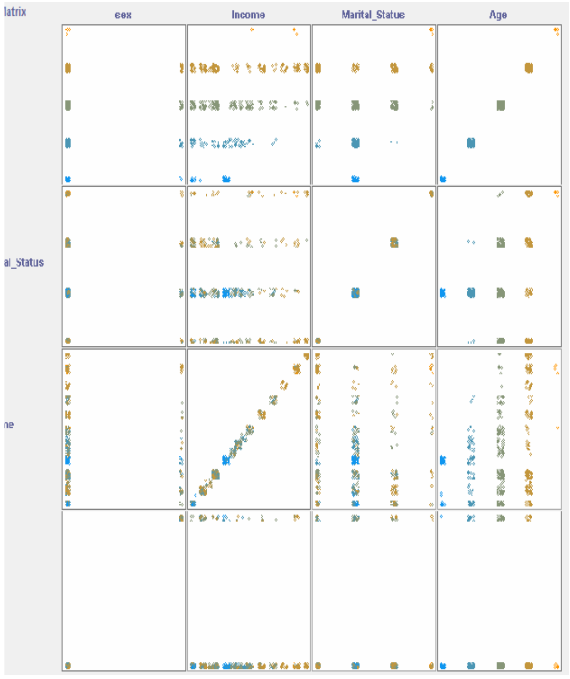
5.1. Expectation Maximization (EM) Clustering Technique

Applying this technique on the logged data set led to the results shown in table (I), and the different clusters are given in figure(3).

Relation: Crime-Criminal
Instances: 925
Attributes: 6, sex, Income, Marital_Status, Age
Test mode: evaluate on training data
Number of clusters selected by cross validation: 6

Table(I). EM Clustering for Criminal data Set

Attribute		Cluster					
		0 (0.09)	1 (0.06)	2 (0.01)	3 (0.72)	4 (0.11)	5 (0)
Sex	*Mean*	2	1	1	1	1	1
	Standard Deviation	0.298	0.298	0.298	0.0008	0.0003	0.298
Income	*Mean*	54.011	96.522	37.5591	49.5645	93.0141	32.4824
	Standard Deviation	29.6706	17.4063	13.275	21.2955	19.1181	11.9028
Marital Status	*Mean*	2.0659	1.0499	1.0196	2	3.4553	3.1133
	Standard Deviation	0.8619	0.2191	0.1795	0.8693	0.5021	0.3171
Age	*Mean*	2.8791	3.7843	3.5046	2.284	3.9252	3.4561
	Standard Deviation	0.8365	0.4113	0.5124	0.8993	0.5876	0.5606
Cluster Instances		91 (10%)	20 (2%)	0	724 (78%)	90 (10%)	0



Fig(3). EM Clustering Visualization

5.2. K-Means Clustering

K-Means represents one of the most popular clustering techniques that is used to group objects according to the similarities between objects in the same group and the dissimilarities between the objects of different groups. The results of applying this algorithm on the logged data are given in table(II).

Relation: Crime-Criminal
Instances: 925
Attributes: 6 (sex, Income, Marital_Status, Age)
Number of iterations: 2
Within cluster sum of squared errors: 211.1029

Table(II). K-Means Clustering for Criminal data Set

		Cluster 0	Cluster 1
Attributes	Full data Set (925)	(91)	(834)
Sex	1.0984	2	1
Income	55.7838	54.011	55.9772
Marital Status	1.9708	2.0659	1.9604
Age	3.0551	2.8791	3.0743
Cluster Instances		91 (10%)	834 (90%)

6. Conclusions

As real world data sets contain missing values and outliers, it is necessary to apply some suitable techniques to overcome such problems that may arise in analyzing the data and preparing them for mining algorithms. Crime and Criminal data sets have some specific characteristics in that missing values and outliers have great meaning in Knowledge Discovery using data mining techniques, so different preprocessing techniques have to be considered here. Deleting missing values and outliers do not fit here, so the proposed technique is used to solve such problem.

The most common technique that is used to replace missing values in data sets is the use of the most frequent value of the attribute, using the average of that attribute or using the class average of that attribute, however, when the missing value represents the target and the goal of the analysis process, estimation of the attribute value doesn't fit here.

As free data set for both crime and criminals collected from the free data available on the Internet, missing values and outliers were replaced and smoothed using two different clustering techniques (EM and K-Means). The mean and the standard deviation for each cluster were found first and then the missing values were replaced by the most suitable values according to the cluster to which the missing values objects belong, and the outliers were smoothed using the same technique.

The results given in tables I and II and in figures 2 and 3 give better estimation for missing values in any object of type crime or criminal, and this is done by finding out the cluster to which the new object belongs and then replacing the missing values in the object with that of the cluster.

7. References

[1] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" 2nd ed., Morgan Kaufmann, 2006.

[2] M. Steinbach, P.-N.Tan and V. Kumar, Introduction to Data Mining, Addison-Wesley, 2006. ISBN: 0-321-32136-7

[3] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2002.

[4] D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001.

[5] Deborah Osborne, MA, Susan Wernicke, MS, "Introduction to Crime Analysis:
Basic Resources for Criminal Justice Practice, The Haworth Press, New York, London, Oxford, 2003.

[6] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed., 2005, ISBN 0-12-088407-0

[7] Derek J. Paulsen, Sean Bair, and Dan Helms Tactical Crime Analysis: Research and Investigation, *2009.*

[8]. Kadhim B. Swadi AlJanabi, Haydar K. "Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures", ICIT 2010, University of Kufa, October 2010.

[9] Tianyi Wu, Yuguo Chen and Jiawei Han, "Association Mining in Large Databases: A Re-Examination of Its Measures", in Proc. 2007 Int. Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD'07), Warsaw, Poland, Sept. 2007.

[10] Hong Cheng, Xifeng Yan, Jiawei Han, and Chih-Wei Hsu, "Discriminative Frequent Pattern Analysis for Effective Classification", in Proc. 2007 Int. Conf. on Data Engineering (ICDE'07), Istanbul, Turkey, April 2007.

[11] Austin Police Department Office, http://www.ci.austin.tx.us/police/crime.htm. 2006.