Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data

Dr: Zakaria Suliman Zubi[1],Ayman Altaher Mahmmud[2]
[1]Sirte University, Faculty Of Science, Computer Science Department
Sirte, P.O Box 727, Libya,
{zszubi@yahoo.com}
[2]The Libyan Academy, Information Technology Department, Tripoli, Libya
{aaa.mahmmud@yahoo.com}

Abstract: - This paper presents a proposed model for crime and criminal data analyzes using Simple K Means Algorithm for data Clustering and Aprior Algorithm for data Association rules. The paper tends to help specialist in discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and indentifying possible suspects. Clustering is based on finding relationships between different Crime and Criminal attributes having some previously unknown common characteristics. Association rules mining is based on generate rules from crime dataset based on frequents occurrence of patterns to help the decision makers of our security society to make a prevention action. The data was collected manually from some police department in Libya. This work aims to help the Libyan government to make a strategically decision regarding prevention the increasing of the high crime rate these days. Data for both crimes and criminals were collected from police departments' dataset to create and test the proposed model, and then these data were preprocessed to get clean and accurate data using different preprocessing techniques (cleaning, missing values and removing inconsistency). The preprocessed data were used to find out different crime and criminal trends and behaviors, and crimes and criminals were grouped into clusters according to their important attributes. WEKA mining software and Microsoft Excel were used to analyze the given data.

Key-Words: - Data mining, data analysis, association rules mining, clustering, criminal data, data mining methods, visualization.

1 Introduction

Data Mining or Knowledge Discovery in Databases (KDD) in simple words is nontrivial extraction of implicit, previously unknown, and potentially useful information from data [1,2,3]. It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. KDD is the process of indentifying a valid, potentially, useful and ultimately understandable structure in data. Data mining represents of the emerging field that can be used a wide disciplinary of applications including marketing, banking, airlines and many other fields that highly affect the communities. Crime analyzes is one of these important applications of data mining. Data mining contains many tasks and techniques including Classification, Association, Clustering, Prediction each of them has its own importance and applications [1,2, 3].

Advances in technology, which allow analyzes of large quantities of data, are the foundation for the for relatively new field known as crime analyze.

Crime analyzes is an emerging field in law enforcement without standard definitions. This makes it difficult to determine the crime analyzes focus for agencies that are new to the field. In some police departments, what is called "crime analysis" consist 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.

Crime analysis is proceeding 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 analysis [4]. The role of the crime analysts varies from agency to agency. Statement of these findings is the objective of most crime analysis 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.

ISBN: 978-960-474-350-6

Assessing crime through analysis also helps in crime prevention efforts [4,7, 10].

2 Why Analyze Crime

Crime Analysts usually tend to justify their existence as crime analysts in what is known as law enforcement agency. It makes sense to analyze crime. Some good reasons are listed below [5, 9]. There may be more other reasons depending on the community culture, geographic efforts, and others, but the most value reasons could be the following:

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 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, nationally within and between law enforcement agencies.
5. Analyze crime to be proactive in detecting and preventing crime.
6. Analyze crime to meet the law enforcement needs of a changing society.
7. Analyze crime to understand the criminal behaviors.

In general there are four different techniques for analyzing crimes, theses techniques are listed as follows:

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.

3 The objectives of the paper

The objectives of this paper work are pointed out as follows:

1. To identify the nature of crime and the crime prevention process.
2. Extracting named entities from narrative reports.
3. To explore and choose among the various data mining software that support clustering and association rule mining techniques to experiment with crime records.
4. To build and train as well as test the performance of the model.
5. To interpret and analyze the results of the model that how strong is the model to extract crime data patterns.
6. To compare the clustering and association rules data mining techniques and select the one which performs the best results.
7. To compare our proposed model with some recent working model.
8. Finally to forward recommendations based on the finding of the study.

4 Problem Solution

We will use a clustering/association rules based model to analyze crime and criminals. This model could MLRC. By this model, we aim to explore the applicability of data mining technique in the efforts of crime prevention with particular emphasis to the dataset which was collected from some police departments. Applying some algorithms will demonstrate the overall results of using both algorithms to perform better results rather in association rules mining or clustering. The rules generated by association rule mining could be easily presented in human language which might be used by police officers to help them decided a crime prevention strategy.

5 Data Mining Task

We will use some data mining tasks represented in the following phases:

5.1 Data Collection Phase

In this phase, the dataset that we used as training and testing data were extracted from the police department. These data contain data about both Crimes and Criminals with the following main attributes:

1. Crime ID: Individual crimes are designated by unique crime id.
2. Crime type: indicates crime type.
3. Date: Indicate when a crime happened.
4. Gender: Male or Female.
5. Age: age of individual Criminal.
6. Crime Address: location of the crime.
7. Marital status: status of the Criminal.

ISBN: 978-960-474-350-6

More than 350 crime records were collected to test the proposed model. The distribution of the collected data is shown in figure 1 below.

CRIMEID,CRIMETYPE,CRIMEADDRESS,CRIMEDATE,GENDER,MARRIED,AGE
1,BURGLARY,TRIPOLI,30SEP12,M,YES,46
2,BURGLARY,BENGHAZI,30SEP12,M,NO,34
3,BURGLARY,BENGHAZI,30SEP12,M,NO,30
4,ARSON,BENGHAZI,30SEP12,M,YES,29
5,ROBBERY,TRIPOLI,30SEP12,M,YES,28
6,MURDER,TRIPOLI,30SEP12,M,YES,46
7,KIDNAPPING,JAFARA,30SEP12,M,NO,26
8,RAPE,JAFARA,30SEP12,M,NO,25
9,DACOITY,TRIPOLI,30SEP12,M,YES,45
10,THEFT,BENGHAZI,1OCT12,M,YES,46
11,MUGGING,BENGHAZI,1OCT12,M,NO,23
12,FRAUD,BENGHAZI,1OCT12,M,YES,28
13,HOMICIDE,BENGHAZI,1OCT12,M,NO,19
14,MUGGING,BENGHAZI,1OCT12,M,YES,43
15,THEFT,TRIPOLI,1OCT12,M,NO,20
16,MUGGING,TRIPOLI,1OCT12,M,NO,31
17,HOMICIDE,TRIPOLI,1OCT12,M,NO,29
18,ROBBERY,TRIPOLI,1OCT12,M,YES,30
19,ROBBERY,TRIPOLI,1OCT12,M,NO,29
20,ROBBERY,JAFARA,1OCT12,M,YES,30
21,ROBBERY,JAFARA,1OCT12,M,YES,31
22,ROBBERY,JAFARA,1OCT12,M,YES,29
23,ROBBERY,JAFARA,1OCT12,M,YES,33

Fig.1: Raw Data

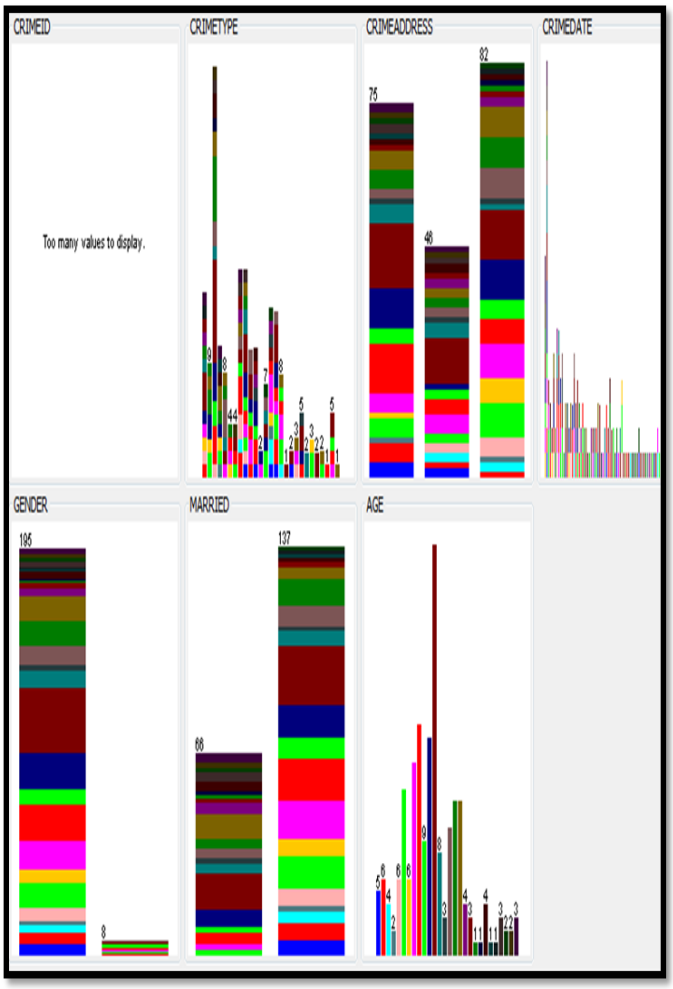

Fig.2: Attributes for crime and criminal

5.2 Data Preprocessing Phase

The real world usually has the following drawbacks: Incompleteness, Noisy, and Inconsistence [9]. Consequently, these data need to be preprocessed to get the data suitable for analysis purpose. The preprocessing includes the following tasks as it shown in [1,2,5,6]:

1. Data cleaning.
2. Data integration
3. Data transformation.
4. Data reduction.
5. Data discretization

In figure (2), we illustrate the distribution of offenses versus of different crime and criminal attributes.

6 Data Set

In this paper, we will consider crime database as a training dataset used in our model. The mentioned database contains a real data values from crime and criminal attributes. We will also consider 70 percent as training value of the proposed model and 30 percent for testing.

7 Methods and model

There are several kinds of data mining methods; Some of the major data mining methods are known as classification, generalized rule induction and summarization [8, 10].

In this work, we will use clustering and association rules methods in order to analyze crime pattern aim to reduce and prevention the crimes as much as possible.

ISBN: 978-960-474-350-6

8 Problem Solution

The proposed model will be named as Mining Libyan Criminal Records (MLCR).

The purpose of this study is to explore the applicability of data mining technique in the efforts of crime analyze and prevention. The data was collected from some police departments in Libya. Our MLCR proposed model will be able to extract crime patterns by using association rule mining and clustering to classify crime records on the basis of values of crime attributes.

The MLCR proposed system will be implemented to conduct to interact with two types of mining algorithms to overcome with two different types of results effectively. Those two approaches are considered as sub-prototypes of the proposed MLCR model. Those prototypes will be illustrated as follows:

8.1 Mining Libyan Criminal Record-using Association rules (MLCR-AR):

In this prototype we will use the Libyan national criminal record dataset gathered from many legal criminal resources.

Association rule mining is method used to generate rules from crime dataset based on frequents occurrence of patterns to help the decision makers of our security society to make a prevention action.
The process includes the following actions:

1. The process of discovering frequently occurring item sets in a database.
2. Intrusion detection: to identify patterns of program executions and user activities as association rules.

In our approach, we used the apriori algorithm in order to discover the best association rules with crimes and criminal attributes, and the results as the follow:

Apriori
=======
Minimum support: 0.3 (4 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 14
Generated sets of large itemsets:
Size of set of large itemsets L(1): 5
Size of set of large itemsets L(2): 6
Size of set of large itemsets L(3): 2
Best rules found:

1. MARRIED=NO 12 ==> GENDER=M 12<conf:(1)> lift:(1) lev:(0) [0] conv:(0)

2.CRIMEADDRESS=TRIPOLI 5 ==> GENDER=M 5 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

3. CRIMEADDRESS=TRIPOLI 5 ==> MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)

4.CRIMEADDRESS=TRIPOLI MARRIED=NO 5 ==> GENDER=M 5 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

5. CRIMEADDRESS=TRIPOLI GENDER=M 5 ==> MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)

6. CRIMEADDRESS=TRIPOLI 5 ==> GENDER=M MARRIED=NO 5 <conf:(1)> lift:(1.08) lev:(0.03) [0] conv:(0.38)

7. CRIMEADDRESS=BENGHAZI 4 ==> GENDER=M 4 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

8. CRIMEADDRESS=JAFARA 4 ==> GENDER=M 4 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

9. CRIMEADDRESS=JAFARA 4 ==> MARRIED=NO 4 <conf:(1)> lift:(1.08) lev:(0.02) [0] conv:(0.31)

10. CRIMEADDRESS=JAFARA MARRIED=NO 4 ==> GENDER=M 4 <conf:(1)> lift:(1) lev:(0) [0] conv:(0)

8.2 Mining Libyan Criminal Record-using Clustering (MLCR-C):

This prototype will use the same dataset indicated in MLCR_AR prototype.

Clustering is the technique that is used to group objects (crime and criminals) without having predefined specification for their attributes.

A cluster is collection of data objects having the following characteristics:

1. Similar to one another within the same cluster.
2. Dissimilar to the objects in other clusters.

Cluster analysis: Grouping a set of data objects into clusters.

Clustering is unsupervised classification: no predefined classes. Simple K-means clustering algorithm is used in this paper.

K-mean algorithm clusters uses the data members groups were *m* is predefined. Input-Crime type. Number of clusters, Number of Iteration Initial seeds might produce an important role in the final results.

ISBN: 978-960-474-350-6

Step1: Randomly choose cluster centers.
Step2: Assign instance to cluster based on their Distance to the cluster centers.
Step3: Centers of clusters are adjusted.
Step4: go to Step1 until convergence.
Step5: output X_0 ,X_1,X_2 ,X_3.

Output:

		Actual	
		Positive	Negative
Predicted	Positive	a	B
	Negative	c	D

Table (1): confusion matrix

All of these values are derived from information provided from the truth table, also known as a confusion matrix, provides the actual and predicted classifications from the predictor

$$TPR=a/a+b$$

$$FPR=b/b+d$$

$$Accuracy= a+d/a+d/a+d+c+d \quad precision=a/a+b$$

The mean idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different results. Consequently, the better choice is to place them as much as possible far away from each other. The next step is to take each point belong to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group is age is done. At this barycenter of the cluster resulting from the previous step and after we have selected *k* as a new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop will be generated after all to produce the results. As a result of this loop we may notice that the *k* centers change their location step by step until no more changes are done or in other words centers do not move any more.

Finally, this algorithm aims to calculate the minimizing an objective function knows as squared error function given by:

$$(1) \quad \boldsymbol{J(v)=} \sum_{i=1}^{c} \sum_{j=1}^{ci} \left(\|xi - vj\| \right)^2$$

Whereas,

'$||x_i - v_j||$' is the Euclidean distance between

x_i and v_j

'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centres.
the results shown below

=== Clustering model (full training set) ===

K-Means
======

Number of iterations: 3
Within cluster sum of squared errors: 43.0
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(13)	(9)	(4)
CRIMEID	13	13	65
CRIMETYPE	MOLESTATION	MOLESTATION	DACOITY
CRIMEADDRESS	TRIPOLI	JAFARA	TRIPOLI
CRIMEDATE	12OCT12	05NOV12	12OCT12
GENDER	M	M	M
MARRIED	NO	NO	NO
AGE	19	19	30

Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0 9 (69%)
1 4 (31%)

Figure (3), illustrates the final results of generating the criminal age and the number of crime attributes in the dataset.

ISBN: 978-960-474-350-6

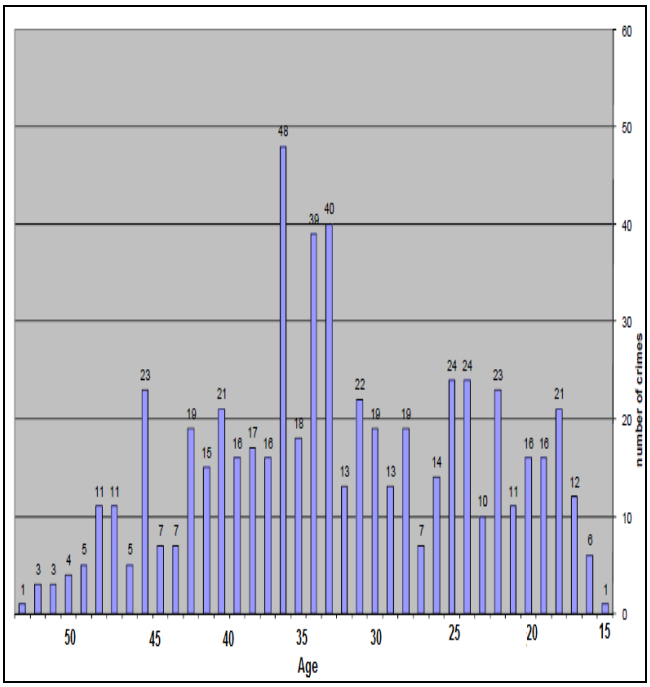

Fig.3: criminal age vs. number of crimes

9 Implementation

A software tools and frameworks that we use in our implementation uses a variety of readily-available software tools and frameworks to deal with the incidental tasks of software development and be able to concentrate on the main objectives of this work. These tools are listed as follows:

1. Google App Engine: Google provides developers with a framework to build and quickly deploy web applications under Google's infrastructure. We decided to build our application using this framework in order to have a better integration with the Google Map API are, which we use to display the crime data.

2. National datasets contents more than 350 records with 7 attributes.

3. WEKA is another prototype data mining tool solution available over the internet. This is being developed by the University of Waikato. New Zealand, through it is implemented primarily in java, recently many more computer language have been added to it.

   WEKA is a shell command based program. Therefore it cannot be directly executed on the web. The user has to create a file in Common Delimited Format (CSV) files can then be input to the WEKA program.

9 Conclusion

An acceptable model for data mining which comes up with excellent results of analyzing crime data set; it requires huge historical data that can be used for creating and testing the model.

More than 350 crime records that were used in this work can give estimation and lead to an acceptable model. WEKA and Excel software were used to preprocess and analyze the collected crime and criminal data. The collected data were preprocessed to be transformed from numeric to nominal by using Numeric-To-Nominal function in WEKA approach.

The raw data that collected from Supreme Security Committee for Tripoli, Benghazi and Al-Jafara were introduced as well. A sample of the confusion matrix was also indicated in this paper. The attributes for crime and criminal and the results of K-means algorithm shows a promising results of our proposed model. It also gives the overall statistical knowledge about the criminal age vs. crime type. This provides the input to the clustering K-means algorithm. Finally our proposed model aims to help the Libyan Security Committee to identify the criminal behavior and specifying offense types related to criminal groups in Libya.

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] Derborah Osborne, MA, Susan Wernicke, MS, "*Introduction to Crime Analysis: Basic Resources for Criminal Justice Practice, the Haworth press*, New York, London, Oxford, 2003.

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

[6] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent pattern without Candidate *Generation: A frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery*, 8(1): 53-87, 2004.

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

[8] Zakaria S. Zubi, Rema A. Saad,*"Using Some Data Techniques for Early Diagnosis of Lung cancer"*,ISBN:978-960-474-273-8.

[9] Zakaria Suliman Zubi, Marim Aboajela Emsaed. 2010. Sequence mining in DNA chips data for diagnosing cancer patients. In Proceedings of the 10th WSEAS international conference on Applied computer science (ACS'10), Hamido Fujita and Jun Sasaki (Eds.). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 139-151.

[10] Zakaria Suliman Zubi. 2009. Using some web content mining techniques for Arabic text classification. In Proceedings of the 8th WSEAS international conference on Data networks, communications, computers (DNCOCO'09), Manoj Jha, Charles Long, Nikos Mastorakis, and Cornelia Aida Bulucea (Eds.). World Scientific and Engineering Academy and Society, Stevens Point, Wisconsin, USA, 73.