Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention

DR: ZAKARIA SULIMAN ZUBI, AYMAN ALTAHER MAHMMUD

***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.

.

***Keywords*—** association rules mining, clustering, criminal, visualization.

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

II. WHY ANALYZE CRIME?

Crime usually tend to justify their existence as crime analysis 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 below [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 abundance of information existing in law enforcement agencies the criminal justice system, and the public domain.

Zakaria Suliman Zubi- He is an Associate Professor since 2010. Works currently at Sirte University, Faculty Of Science, Computer Science Department Sirte, P.O Box 727, Libya,. Email : {zszubi@yahoo.com}.

Ayman Altaher Mahmmud – He is a postgraduate student at the Libyan Academy, Information Technology Department, Tripoli, Libya. Email: {aaa.mahmmud@yahoo.com}

Analyze crime to maximize the use of limited law enforcement resources [7].

III. TYPES OF CRIME ANALYSIS

A. Tactical Crime Analysis

The tactical crime analysis involves analyzing data to develop information on where, when, and how crimes happens in order to assist officers and investigators in identifying and understanding specific and immediate crime problems [4]. Tactical crime analysis units will work closely with patrol officers and investigators. The goal of tactical analysis is to promote a rapid response to a crime problem happening currently. One of the roles as a tactical crime analysis is to detect current patterns of criminal's activity to predict possible future crime events.

B. Strategic Crime Analysis

Strategic crime analysis is concerned with long-range problems and planning for long-term projects. Strategic analysis examine long term increases or decreases in crime, known as "crime trends", A crime trend is the direction of movement of crime and reflects either no change or increases/decreases in crime frequencies within a specific jurisdiction or area [4]. For instance, strategic analysts might study increased car thefts during the winter months when citizens warm up their cars, leaving them unlocked and unattended in various locations.

C. Administrative Crime Analysis

Administrative crime analysis focuses on providing summary data, statistics, and general trend information to police managers. This type of analysis involves providing descriptive information about crime to department administrators, command staff, and officers, as well as to other city government personnel and the public. Such reports provide support to administrators as they determine and allocate resources or help citizens to have a better understanding of the community crime and disorder problems.

D. Investigative Crime Analysis

Investigative crime analysis involves profiling suspect and victims for investigators based on analysis of available information. It is sometimes called "criminal investigative analysis". Generally; it focuses on hypothesizing about what type of person is committing a particular crime series.

E. Intelligence Analysis

Intelligence analysis focuses on organized crime, terrorism, and supporting specific investigations with information analysis and presentation. Analysts can support investigations by becoming the "processor" of information for officers. In a homicide investigation, the tools of analysis can be used to organize investigative information and display it in the form of time lines and association link charts.

F. Operations Analysis

Operations analysis examines how a law enforcement agency is using its resources. It focuses on such topics as deployment, use of grant funds, redistricting assignments, and budget issues. In many agencies crime analysts are asked to assist on special projects for the department that fall into the category of operations analysis.

IV. DATA MINING AND CRIMR PATTERN

We will look at how to convert crime information into a data-mining problem [11]. In this case it can help the analysts to identify crimes faster and help to make faster decisions. We have seen that in crime terminology a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in data mining terminology a cluster is group of similar data points which can be a possible crime pattern. Thus appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns. Thus clustering algorithms in data mining are equivalent to the task of identifying groups of records that are similar between themselves but different from the rest of the data. In our case some of these clusters will be useful for identifying a crime spree committed by one or same group of suspects. Given this information, the next challenge is to find the variables providing the best clustering. These clusters will then be presented to the detectives to drill down using their domain expertise.

The automated detection of crime patterns, allows the detectives to focus on crime sprees first and solving one of these crimes results in solving the whole "spree" or in some cases if the groups of incidents are suspected to be one spree, the complete evidence can be built from the different bits of information from each of the crime incidents. For instance, one crime site reveals that suspect has black hair, the next incident/witness reveals that suspect is middle aged and third one reveals there is tattoo on left arm, all together it will give a much more complete picture than any one of those alone. Without a suspected crime pattern, the detective is less likely to build the complete picture from bits of information from different crime incidents.

Today most of it is manually done with the help of multiple spreadsheet reports that the detectives usually get from the computer data analysts and their own crime logs. We choose to use clustering technique over any supervised technique such as classification, since crimes vary in nature widely and crime database often contains several unsolved crimes. Therefore,

classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Also nature of crimes change over time, such as Internet based cyber crimes or crimes using cell-phones were uncommon not too long ago. Thus, in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

V. CRIME DETECTION

Intelligence agencies are actively collecting and analyzing information to investigate criminal's activities. Local law enforcement agencies have also become more alert to criminal activities in their own jurisdictions. When the local criminals are identified properly and restricted from their crimes, then it is possible to considerably reduce the crime rate. Criminals often develop networks in which they form groups or teams to carry out various illegal activities. Data mining task consisted of identifying subgroups and key members in such networks and then studying interaction patterns to develop effective strategies for disrupting the networks. Data is used with a concept to extract criminal relations from the incident summaries and create a likely network of suspects [12]. Co-occurrence weight measured the relational strength between two criminals by computing how frequently they were identified in the same incident.

VI. COMBAITING CRIMES

Using data mining, various techniques and algorithms are available to analyze and scrutinize data. However, depending on the situation, the technique to be used solely depends upon the circumstance. Also one or more data mining techniques could be used if one is inadequate. Data mining applications also uses a variety of parameters to examine the data start investigation as to the likely causes of the attack and the individuals who might have responsible attack. We have stated that crime investigation remains the prerogative of the law enforcement agencies concern, but computer and computer analysis can be useful in solving detecting.

VII. 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 technique 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 findings of the study.

VIII. PROBLEM STATEMENT

In my country Libya the current system of the Supreme Security Committee (SSC) is a manual system. We aim to explore in this work the applicability of data mining technique in the efforts of crime prevention with particular emphasis to the dataset we collected from Benghazi**,** Tripoli, and Al-Jafara SSC's. We propose to implement a model that could help us to extract crime patterns. These patterns will be applied to some data mining algorithms such as association rules mining and clustering to classify crime records on the basis of the values of attributes crime. Applying such algorithms will illustrate the overall results of using both algorithms to perform better results rather in association rule mining or in clustering. The rules generated by association rule mining could be easily presented in human language which might be used by SSC officers to help them decided a crime prevention strategy.

IX. DATA MINING TASK

We will use some task as the follow:

A. 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 IDs.
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.

More than 350 crime records were collected to test the proposed model. The distribution of the collected Data is shown in table (a) 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

Table(a):attributes for crime and criminal

B) Data Preprocessing Phase

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 Purpose. The preprocessing includes the following tasks [1],[2],[5],[6].

1) Data cleaning: fill in missing values, smooth noisy Data, identify or remove outliers, and resolve Inconsistencies
2) Data integration: using multiple database
3) Data transformation.
4) Data reduction.
5) Data discretization.

figure (1) shows the distribution of offenses versus different crime and criminal attributes.

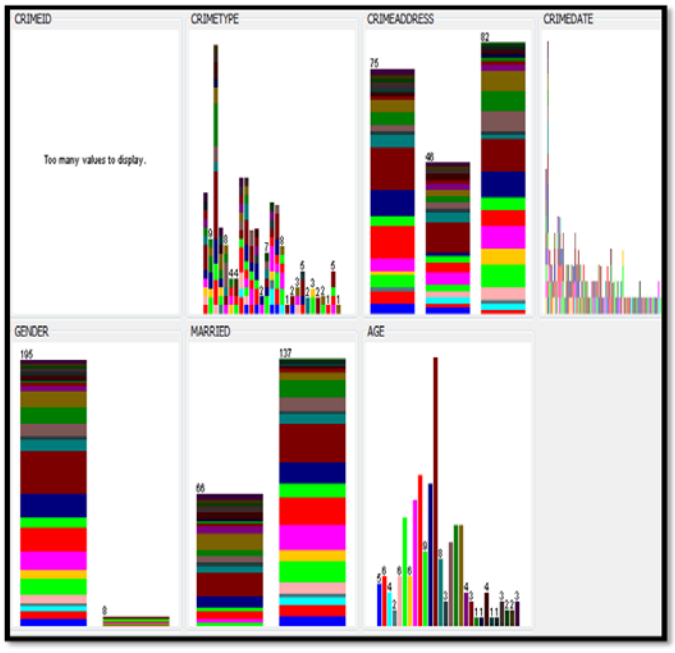

Fig.1 attributes for crime and criminal

X. DATA SET

In this paper, we will consider crime database as a training dataset used in our proposed 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.

XI. PROBLEM SOLUTION

The proposed model will be named as Mining Criminal (MLCR).Law enforcement agencies today are faced a large volume of data that must be preprocessed and transformed into useful information. Data mining can improve crime analysis and aid in reducing and preventing crime.

The purpose of this study to explore the applicability of data mining techniques in the efforts of crime analysis and prevention. The data was collected manually from Benghazi, Tripoli, and AL-Jafara SSCs. 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 the values of crime attributes.

The MLCR proposed system will be implemented to conduct to interact with two types of mining algorithm 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:

A. Mining Libyan Criminal Records using Assocaition Rules (MLCR-AR)

In this prototype we will use the Libyan notional criminal record dataset gathered from many legal criminal resources such as Benghazi, Tripoli, and Al-Jafara SSC.

In this prototype we will use the Libyan notional criminal record dataset gathered from many legal criminal resources such as Benghazi, Tripoli, and Al-Jafara SSC.

Association rule mining is a method used to generate rules from crime dataset based on frequents occurrences of the patterns to help the decision makers of our security society to take 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)

B. Mining Libyan Criminal Records using Assocaition Rules (MLCR-C)

This prototype will use the same dataset indicated in MLCR-AR prototype.

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

A cluster is a 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-Means algorithm clusters 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.

Step 1: Randomly Choose cluster centers.
Step 2: Assign instance to cluster 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 X0,X1,X2,X3.

Output

		Actual	
		Positive	Negative
Predicted	Positive	a	B
	Negative	c	D

Table(2): 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 (2)

FPR = b/b+d (3)

Accuracy= a+d/a+b+c+d 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 result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. 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 at minimizing an objective function know as squared error function given by:

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

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 centers.

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:

Attribute	Full Data (13)	Cluster# 0 (9)	1 (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%)

The K-Mean algorithm is fast, robust and easier to understand. and gives best results when data set are distinct or well separated from each other.

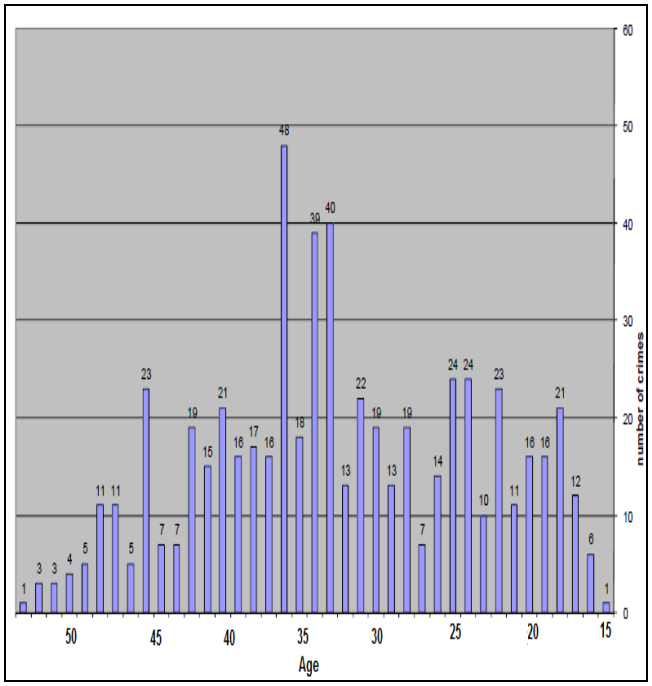

Fig.3: criminal age vs. number of crimes

XII. IMPLEMENTATION

A software tools and frameworks that we use 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 paper.

- 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 and criminal data.

- National dataset contents more than 340 records with 7 attributes.

- WEKA is another prototype Data mining tool available over the internet. This is being developed by The University of Waikato. New Zealand, though it is implemented primarily in Java, recently many more computer languages 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 Attribute Related file Format (ARFF) files or create a file in Common delimited Format (CSV) files can then be input to the WEKA program.

XIII. 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.

First of all, the collected data were preprocessed to transform dataset from numeric to nominal by using Numeric-To-Nominal which is in Unsupervised>attribute in WEKA, and split the dataset into teo,30% testing and 70% training. Into format suitable for analyze purpose.

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.

This model aims to help the Libyan Security Committee to identify the criminal behavior and specifying offense types related to criminal groups in Libya.

XIV. ACKNOWLEDGMENT

we would like to thank all the members of Supreme Security Committee for Tripoli, Benghazi, and AL-Jafara.

XV. 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.

[11] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society April 2004.

[12] Brown, D.E. The regional crime analysis program : A frame work for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics.