A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K-Means Mining Algorithms

Kadhim B. Swadi Al-Janabi

Department of Computer Science

Faculty of Mathematics and Computer Science

University of Kufa /Iraq - kadhimbs@yahoo.com

Abstract

This paper presents a proposed framework for the crime and criminal data analysis and detection using Decision tree Algorithms for data classification and Simple K Means algorithm for data clustering. The paper tends to help specialists in discovering patterns and trends, making forecasts, finding relationships and possible explanations, mapping criminal networks and identifying possible suspects. The classification is based mainly on grouping the crimes according to the type, location, time and other attributes; Clustering is based on finding relationships between different Crime and Criminal attributes having some previously unknown common characteristics. The results of both classifications and Clustering are used for prediction of trends and behavior of the given objects (Crimes and Criminals).

Data for both crimes and criminals were collected from free police departments' dataset available on the Internet to create and test the proposed framework, 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

Data Mining, Classification, Decision Tree, Clustering.

1. INTRODUCTION

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, 2, 3]. 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. A simple taxonomy of knowledge discovery techniques looks like the following [1, 5, 9]:

- Manual search
- OLAP (On Line Analytical Processing)
- Knowledge Engineering
- Visualization
- Automated search
- Auto-clustering
- Link analysis
- Regression

Rule Induction

Data mining represents one of the emerging fields 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 own importance and applications [1,2,3,4, 6, 7].

Advances in technology, which allow analyses of large quantities of data, are the foundation for the relatively new field 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. Some analysts do all this and other types of analysis [5]. The role of the crime analyst varies from agency to agency.

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 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, 11]. Preventing crime costs less than trying to apprehend criminals after crimes occur.

Crime analysis is defined as a set of systematic, analytical processes directed at providing timely and pertinent information relative to crime patterns and trend correlations to assist operational and administrative personnel in planning the deployment of resources for the prevention and suppression of criminal activities,

known as *crime analysis.* Crime Analysis

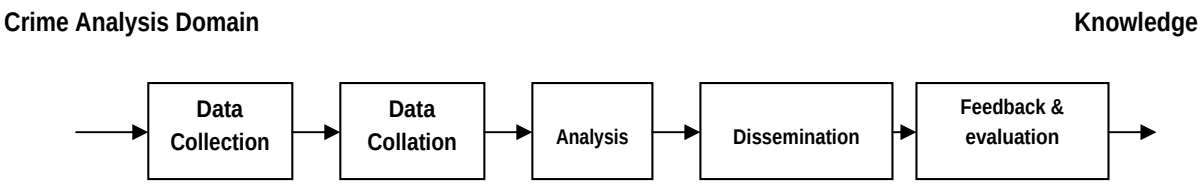
aiding the investigative process, and increasing apprehensions and the clearance of cases. Within this context, crime analysis supports a number of department functions including patrol deployment, special operations and tactical units, investigations, planning and research, crime prevention, and administrative services [5, 11, 12].

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 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 of the abundance of information existing in law enforcement agencies, the criminal justice system, and the public domain.

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



Fig(1). Steps for Crime Analysis Process.

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. The main objectives of crime analysis can be classified into:

- **Extraction of crime patterns by analysis of available crime and criminal data. The study in this paper focuses on this objective.**
- **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.**

3. TYPES OF CRIME ANALYSIS

3.1 Tactical Crime Analysis

Tactical crime analysis involves analyzing data to develop information on the where, when, and how of crimes in order to assist officers and investigators in identifying and understanding specific and immediate crime problems [5]. Tactical crime analysis units focus on and will work closely with patrol officers and investigators. The goal

of tactical analysis is to promote a rapid response to a crime problem happening right now. One of your roles as a tactical crime analyst is to detect current patterns of criminal activity to predict possible future crime events.

3.2 Strategic Crime Analysis

Strategic crime analysis is concerned with long-range problems and planning for long-term projects. Strategic analysts 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 [5]. For example, 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.

3.3. 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.

3.4. Investigative Crime Analysis

Investigative crime analysis involves profiling suspects 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.

3.5. 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.

3.6. 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.

4. DATA COLLECTING AND PREPROCESSING

The dataset used as training and testing data for the proposed frame work were extracted from the Internet [10]. These data contain data about both crimes and criminals with the following main attributes:

- Crime ID :Individual Crimes are designated by unique Crime IDs
- Crime Name :Disguised crime's name
- Crime Type: Indicates crime type.
- Day, date, time: indicate when a crime happened.
- Criminal ID: 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.
- Criminal Income.

4.1 Data preprocessing

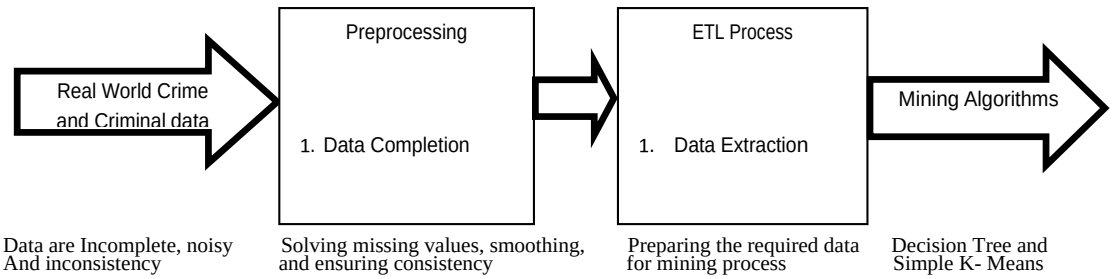
Real world data usually have the following drawbacks: Incompleteness, Noisy and Inconsistence. So, these data need to be preprocessed to get the data suitable for analysis purposes. The preprocessing includes the following tasks [1, 2 ,8, 9]:

- 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 cannot 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.

The process is show in figure (2).



Figure(2). Crimes and Criminals Data Preprocessing.

Samples of crime and criminal data before and after the preprocessing steps are given in table(I) and table(II).

Crime-ID	Criminal Gender	Criminal Job	Income	Marital status	Age
1	Male	Non employment	20k	S	22
2	Male	building worker	40k	M	33
3	Male	Teacher	80k	M	36
4	Male	cleaning worker	30k	M	38
5	Male	building worker	40k	M	24
6	Male	Teacher	80k	D	41
7	Male	Non employment	20k	S	18
8	Male	Student	0	S	22
9	Male	Dealer	110k	M	33
10	Male	Teacher	80k	D	35
11	Male	cleaning worker	30k	D	34
12	Male	building worker	40k	M	36
13	male	Non employment	20k	S	22
14	male	Non employment	20k	S	20
15	male	cleaning worker	30k	S	19
16	male	Teacher	80k	W	44
17	male	Dealer	110k	W	45
18	male	Technical	70K	W	51
19	male	Non employment	20k	W	50
20	male	Non employment	20k	W	49
21	male	cleaning worker	30k	M	37
22	male	building worker	40k	M	34
23	male	Student	0	S	22
24	male	Non employment	20k	S	26
25	male	Teacher	80k	M	40
26	male	Engineer	110 K	M	41
27	male	Doctor	100 K	W	52
28	male	building worker	40k	S	33
29	male	building worker	40k	S	31
30	male	Non employment	20k	S	29
31	male	cleaning worker	30k	M	37
32	male	building worker	40k	M	37
33	male	Non employment	20k	M	37
34	male	building worker	40k	M	35
35	male	building worker	40k	M	33
36	male	Non employment	20k	M	31
37	male	Technical	70K	S	36
38	male	Technical	70K	S	38
39	male	Dealer	110k	D	40
40	male	Doctor	100 K	D	30
41	male	cleaning worker	30k	D	46
42	male	Mechanical	90k	D	28
43	female	cleaning worker	30k	S	25
44	female	Mechanical	90k	S	25
45	male	Non employment	20k	S	25
46	male	Mechanical	90k	S	24
47	male	Doctor	100 K	S	44
48	male	Doctor	100 K	M	45
49	male	Non employment	20k	S	17
50	male	Non employment	20k	S	19

Table(I). Real World Crime and Criminal Data

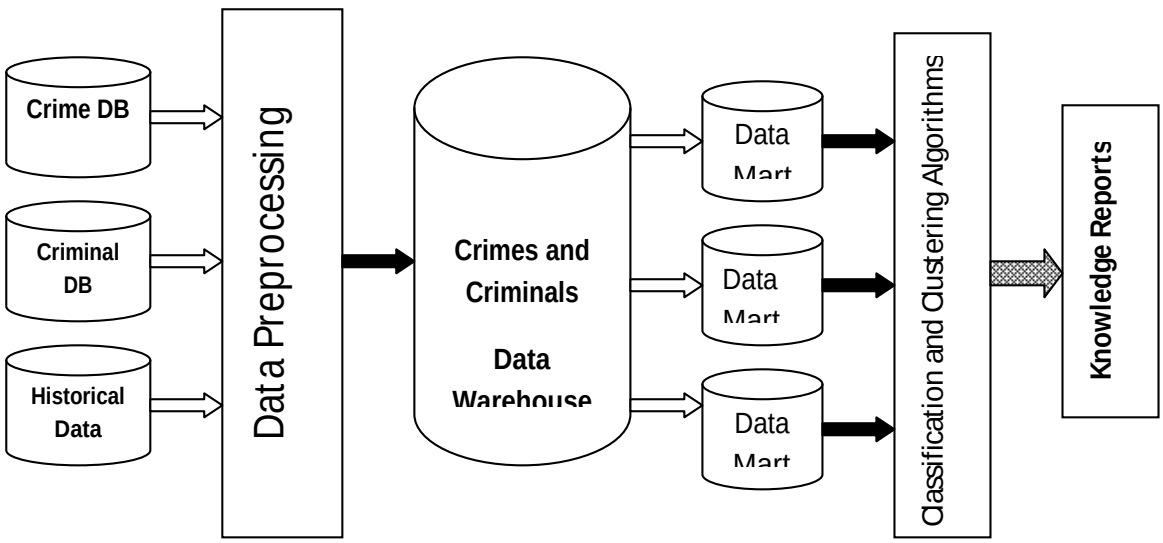
Table(II). Samples of the Preprocessed Data.

Crime ID	Offense Type	Day of the Week 1=Sun"	Criminal ID	Criminal Sex 1=Male	Criminal Income In K	Criminal Marital Status 1=Single	Criminal Age Category 1=Age<20
24	1	6	271	1	120	4	4
36	1	5	50	1	20	2	1
123	1	5	408	1	50	2	1
218	1	6	586	1	35	2	2
231	1	7	10	1	80	1	3
242	1	2	286	1	20	2	3
316	1	5	554	1	65	2	2
364	1	7	165	1	20	1	4
404	1	6	575	1	40	1	3
444	1	6	446	1	40	2	3
592	1	6	356	1	40	1	3
602	1	7	74	1	40	1	4
678	1	7	411	1	55	2	2
686	1	7	315	1	40	2	3
945	1	6	31	1	30	3	4
955	1	6	194	2	50	1	2
102	2	1	165	1	20	1	4
106	2	1	510	1	50	2	1
116	2	2	235	1	35	3	3
211	2	6	468	1	40	2	3
212	2	6	420	1	110	1	4
415	2	1	194	2	50	1	2
596	2	6	45	1	20	2	2
762	2	6	481	1	40	1	4
773	2	6	310	2	20	2	2
826	2	1	53	1	70	1	4
835	2	1	301	1	30	3	3
857	2	2	221	1	60	3	4
860	2	2	545	1	55	2	2
865	2	2	489	1	100	3	4
871	2	2	515	1	65	2	2
958	2	6	192	2	90	2	2
995	2	7	274	1	90	3	3
1001	2	7	143	1	50	2	1
85	3	1	512	1	55	2	3
95	3	6	165	1	20	1	4
96	3	6	468	1	40	2	3
312	3	4	347	1	60	2	2
670	3	6	43	2	30	2	2
718	3	2	571	1	100	4	4
855	3	1	481	1	40	1	4
705	4	2	295	1	110	4	5
710	4	2	36	1	20	3	3
973	4	7	405	1	35	2	3
978	4	7	126	1	40	4	4
988	4	7	594	1	110	1	4
993	4	7	435	1	120	4	4
9	5	4	382	1	50	2	2
27	5	7	63	1	30	3	3
55	5	1	586	1	35	2	2

5. Proposed Framework Architecture

Most of the data mining techniques are applied on a data repository called Data Warehouse that integrates data from different sources on different integrals, these data may come in different formats e. g. Databases, spreadsheets, filing systems, XML files and other formats. These data are cleaned up and converted into the required formats then moved to the data warehouse that may contain detailed, lightly summarized and highly summarized data depending on the analysis required.

Data Marts can be drawn from the data warehouse depending on the system performance required and the data ownership, and then data mining techniques are applied on either data warehouse or the data marts. Figure 1 shows the proposed system architecture.



Figure(3). Proposed Framework Architecture for Crime Analysis

6. Classifiers and Clustering Models

Classification models are highly used in mining crimes and criminals data to get some patterns that have a wide variety of applications in different fields, one of which is to understand the different reasons behind crimes of specific types and the grouping of different crimes and criminals according to some related attributes including crime type, criminal age, marital status, employability, and other attributes. Decision Trees are very well fitted for classifying crimes and criminals objects. Information gain[1]is a critical criteria to choose the splitting attribute that is used to construct the decision tree, and in order to define information gain precisely, we need to define a measure commonly used in information theory, called Entropy and can be calculated using equation (1) and (2).

$$Entropy(S) = \sum_{i=1}^{n} - P_i \log_2 P_i \qquad ...(1)$$

and the Information Gain is given in equation (2)

$$Gain(S, A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \qquad ...(2)$$

The importance of defining and finding the information gain based on the entropy is to put the most important attribute on the top level of the decision tree and it can be used to cancel those attributes that do not affect the decision tree structure.

Decision trees are simple knowledge representation and they classify examples to a finite number of classes, the nodes are labeled with attribute names, the edges are labeled with possible values for this attribute and the leaves labeled with different classes. Objects are classified by following a path down the tree, by taking the edges, corresponding to the values of the attributes in an object.

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

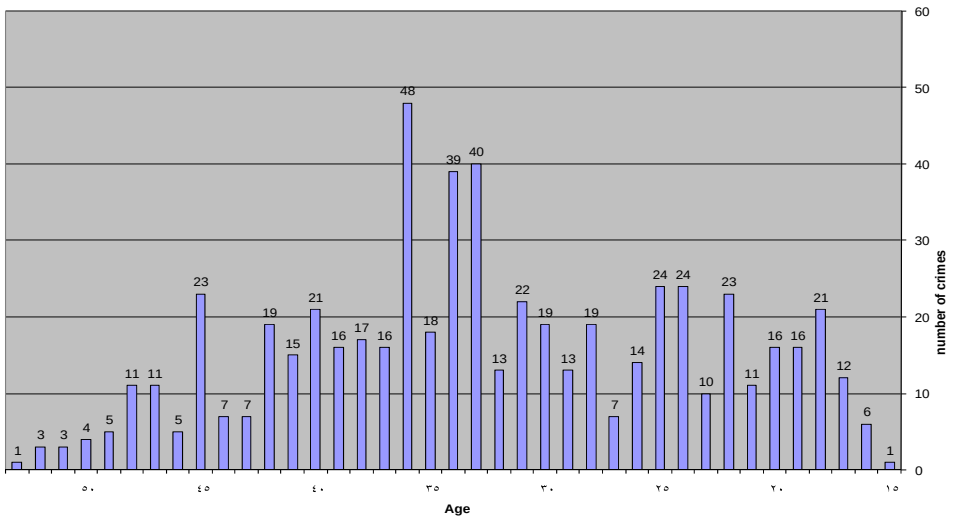
A cluster is a collection of data objects having the following characteristics:

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

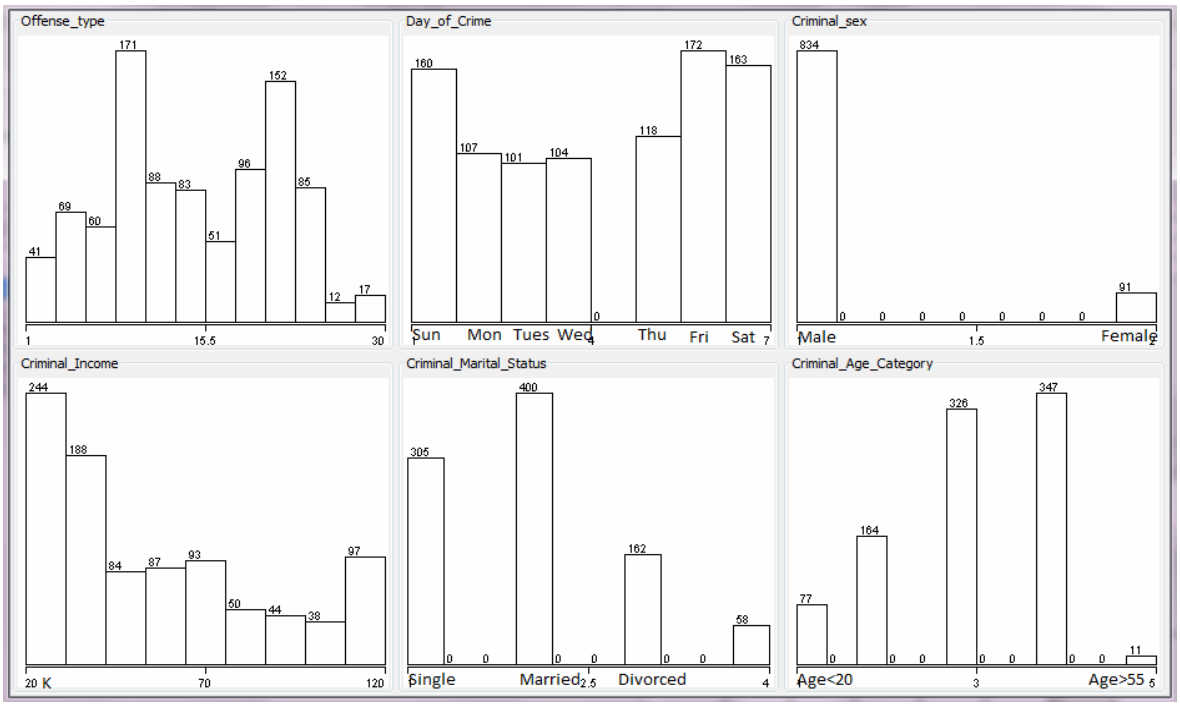
Cluster analysis: Grouping a set of data objects into clusters

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

Figures (4) shows the relationship between the criminal ages and the number of crimes, figure (5) shows the distribution of offenses versus different crime and criminal attributes.



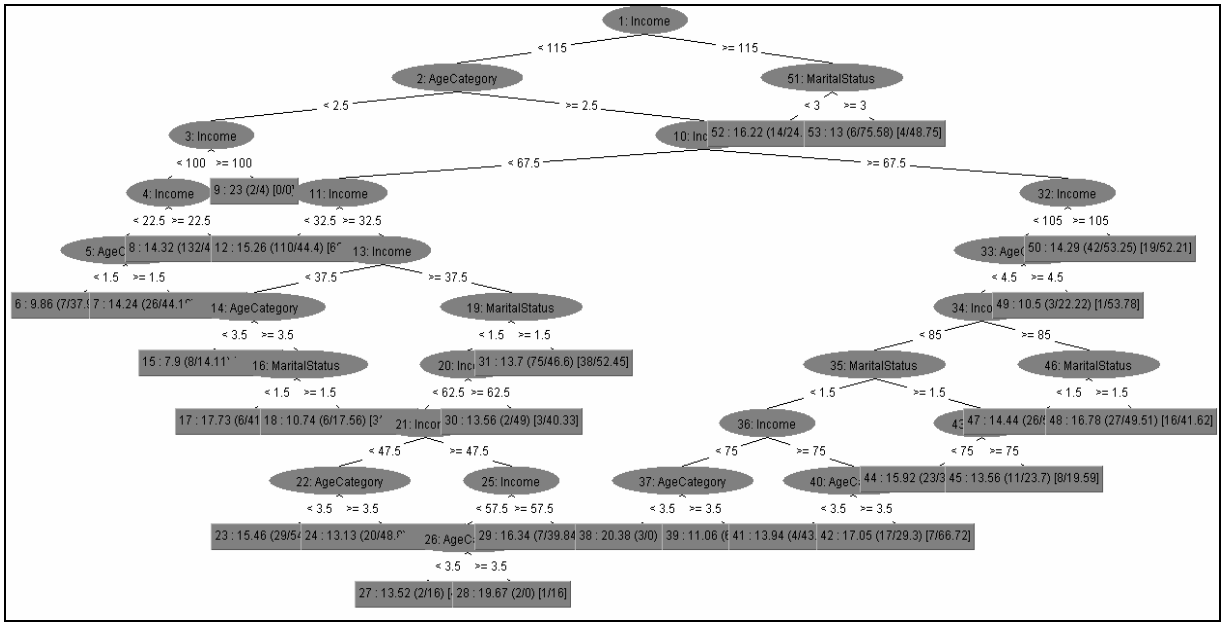
Figure(4). Crimes vs. Criminal Ages.



Figure(5). Offense Distributions vs. Different Crime and Criminal Attributes.

7. Decision Trees

Applying decision tree algorithms in WEKA led to the diagrams shown in figure 6, 7, 8, and 9.



Figure(6). Decision Tree for Offense Type VS. Income, Marital Status, and Age Category

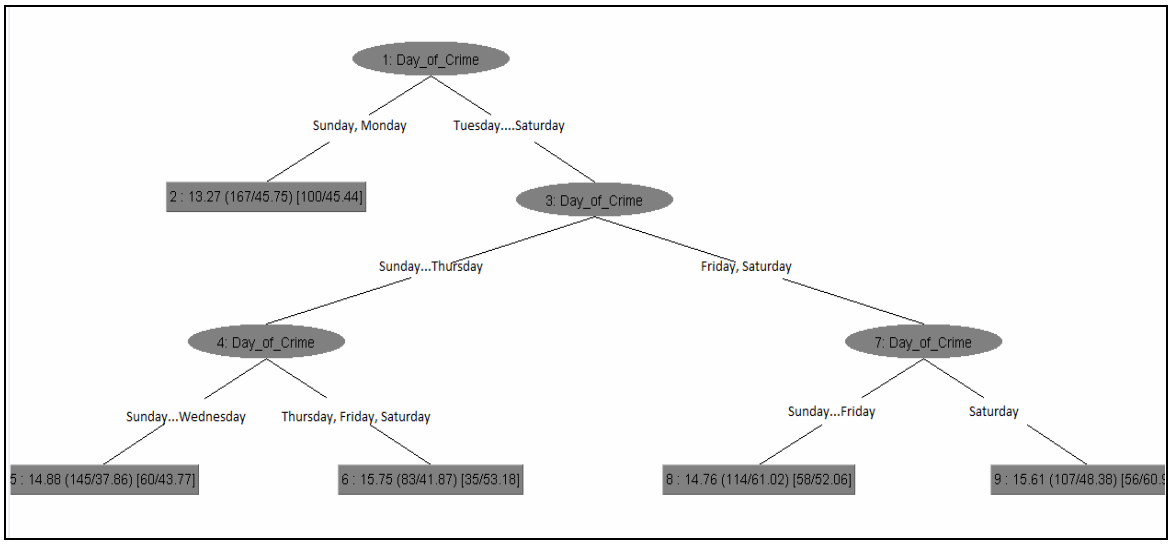
Algorithm information:

Instances: 925

Attributes: 4 {Offense type, Income, Marital Status, Age Category}

Size of the tree: 53

Correlation coefficient	-0.0049
Mean absolute error	6.0996
Root mean squared error	7.0614
Relative absolute error	100.7451 %
Root relative squared error	101.2513 %



Figure(7). Decision Tree for Offence Type VS. Day of the Crime.

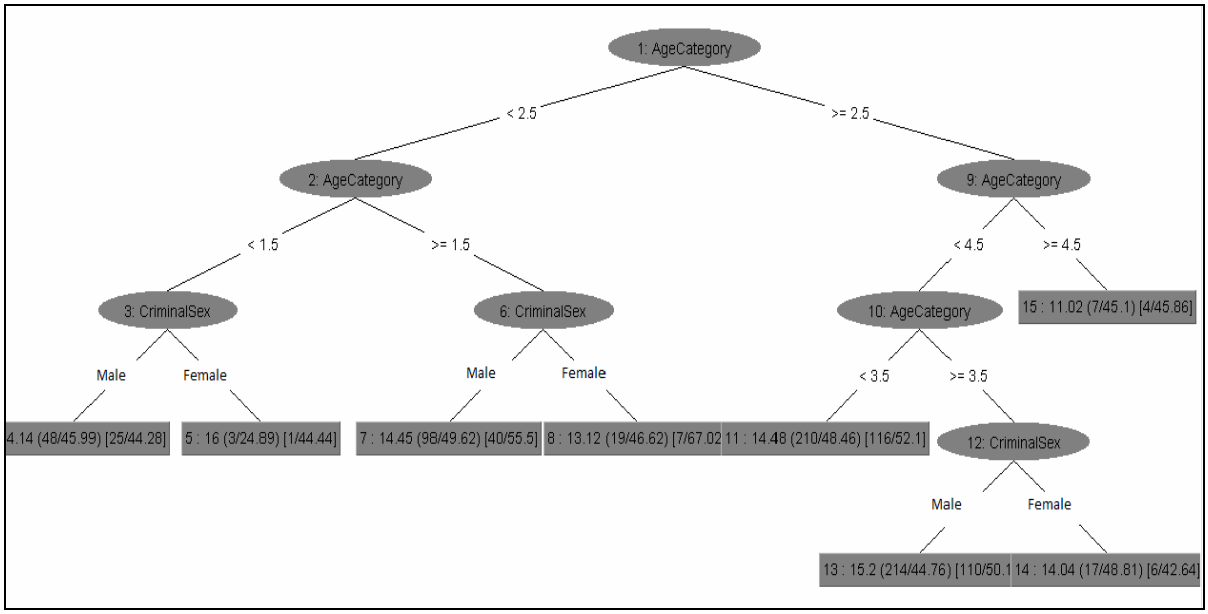
Algorithm Information

Instances: 925

Attributes: 2 {Offense type, Day of Crime}

Size of the tree : 9

Correlation coefficient	0.0709
Mean absolute error	6.0167
Root mean squared error	6.9582
Relative absolute error	99.3767%
Root relative squared error	99.7715%



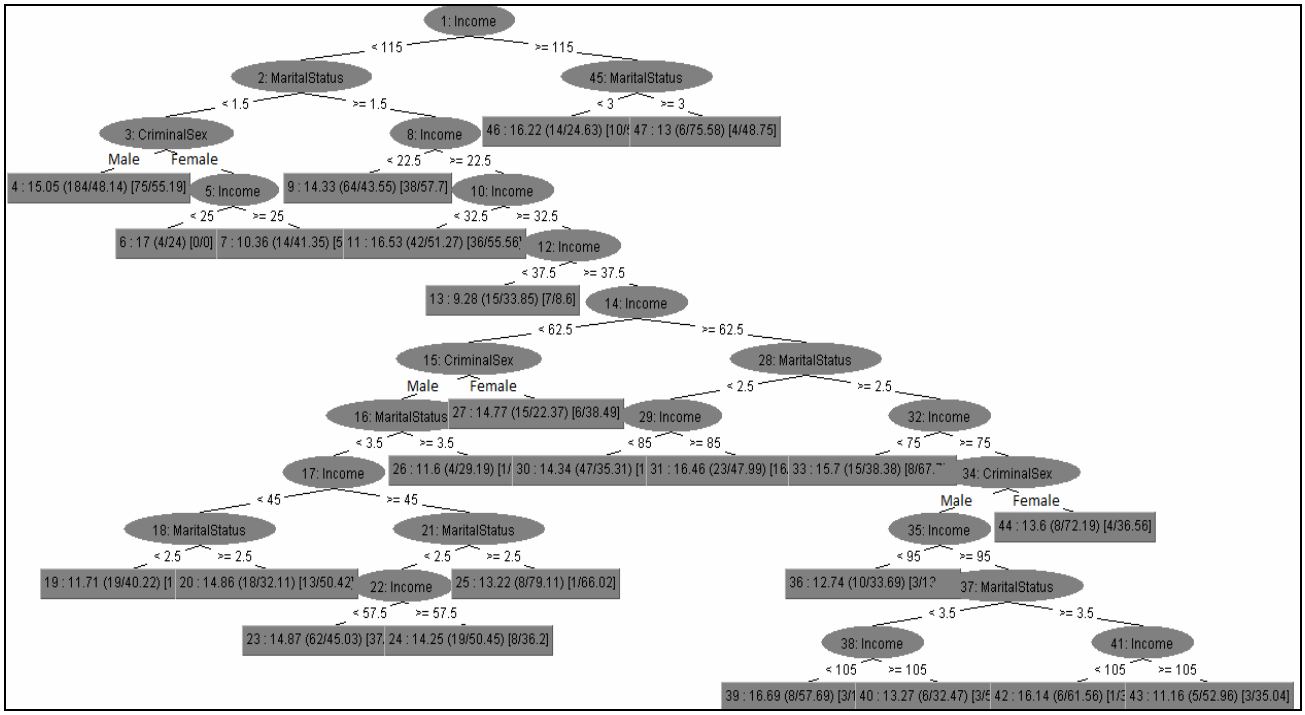
Figure(8). Decision Tree for Offence Type VS. Sex and Age Category.

Instances: 925

Attributes: 3 {Offense type, Criminal Sex, Age Category}

Size of the tree : 15

Correlation coefficient	-0.0679
Mean absolute error	6.0747
Root mean squared error	7.0026
Relative absolute error	100.3338 %
Root relative squared error	100.4081 %



Figure(9). Decision Tree for Offense Type VS. Sex, Income, and Marital Status.

Instances: 925

Attributes: 4 {Offense type, Criminal Sex, Income, Marital Status}

Size of the tree : 47

Correlation coefficient	-0.0028
Mean absolute error	6.1026
Root mean squared error	7.0858
Relative absolute error	100.7946 %
Root relative squared error	101.6 %

7. Clustering Simple K-Means Algorithms

Clustering represents one of the most efficient way for grouping criminals into groups without identifying previous attribute specifications, according to some types of similarity and dissimilarity between these objects. In the proposed framework, this can help in identifying some common criminal behaviors according to their groups. Simple K-Means clustering algorithm was used in this paper since it is very well fitted for such type of data.

7. 1. Clustering Algorithm 1

Instances: 925

Attributes: 4 {Criminal Sex, Income, Marital Status, Age Category}

K Means Algorithm

Number of iterations: 2

Within cluster sum of squared errors: 211.1029

Attribute	Full Data (925)	0 (91	1 (834)
Criminal Sex	1.0984	2	1
Income	55.7838	54.011	55.9772
Marital Status	1.9708	2.0659	1.9604
Age Category	3.0551	2.8791	3.0743

Clustered Instances:

0 91 (10%)

1 834 (90%)

7.2. Clustering algorithm 2

Instances: 925

Attributes: 2 {Offense type, Day of Crime}

Test mode: evaluate on training data

Number of iterations: 8

Within cluster sum of squared errors: 78.049

Attribute	Full Data (925)	0 (472)	1 (453)
Offense type	14.6411	14	15.3091
Day of Crime	4.1686	2.3157	6.0993

Clustered Instances

0 472 (51%)

1 453 (49%)

8. Conclusions

To get a satisfactory model for data mining and to get excellent results of analyzing crime data set, it requires huge historical data that can be used for both creating and testing the model. One thousand crime records and more than six hundred criminal records that were used in this work can give good estimation and lead to an acceptable model. WEKA (Weikatto Environment for Knowledge Analysis) and Excel software were used to preprocess and analyze the collected crime and criminal data.

First of all, the collected data were preprocessed to fill in the missing attributes, remove outliers, and to reduce data using attribute extraction technique, and then data were normalized and transformed into formats suitable for analysis purposes. Tables (I) and (I) show sample of the data before and after preprocessing. It is clear that data in table(II) can be very well fitted for analysis using decision tree and clustering algorithms.

Figure(4) and figure(5) give the overall statistical knowledge about the data prepared as input to the mining algorithms. Decision tree in figures(6) through (9) give the paths for different types of offenses depending on different crime and criminal attributes, this will help in identifying what attributes highly affect a specific type of offense. Entropy and information gain locate the attributes highly affecting the results at the top of the tree.

Clustering results given in algorithm 1 and 2 tend to group criminals according to some attributes (criminal gender, income, age category and marital status), this will help in identifying the criminal behavior and specifying offense types related to criminal groups.

9. 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] Haider k. and Kadhim Aljanabi, "Crime Data Analysis Using Data Mining Techniques To Improve Crimes Prevention Procedures", ICIT2010, October 2010, University of Kufa, Iraq.

[8] 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.

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

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

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

[12] Hsinchun Chen, Homa Atabakhsh, Tim Petersen, "Visualization for Crime Analysis", Proceedings of The National Conference on Digital Government Research

CiteSeerx, COPLINK, 2006.

الملخص:

تقدم هذه الورقةِ البحثية إطارا ونموذجا لتحليل بياناتِ الجريمة باستخدام تقنيات وخوارزميات مفاهيم التنقيب عن البيانات (التصنيف والتجميع Classification and Clustering)بهدف تقديم افضل المعلومات الى المختصين في علم الجريمة للمساعدة في الكشف عن الجريمة. يهدف البحث إلى مساعدة الإختصاصيين في إكتِشاف الأنماطِ والإتّجاهاتِ للجرائم والمجرمين و إيجاد عِلاقاتِ وتفسيراتِ محتملةِ للجرائم ومتابعة الشبكاتَ إلاجرامية وتمييز مشتبه بهمَ محتملينَ. إنّ التصنيفَ بشكل رئيسي يستند اساسا على تصنيف الجرائم طبقاً للنوع، العنوان، وقت حصول الجريمة، صفات المشتبه بهم وغيرها. اضافة الى إيجاد العلاقات بين الجرائم المختلفة والخواص الإجراميةِ. ولتحقيق ذلك تم استخدام خوارزميات مختلفة لما يسمى بشجرة القرارات Decision Tree Algorithms لاجراء عملية التصنيف وتقنيات المتوسط البسيط Simple K-Mean للتجميع.

تم تجميع البيانات عن الجرائم والمجرمين من البيانات الحرة على الانترنت، حيث استخدمت هذه البيانات لانشاء واختبار النموذج المقترح، وقد تم استخدام خوارزميات مختلفة لاعداد هذه البيانات لكي تتلائم مع خوارزميات التنقيب المختلفة وبعد ذلك تم تطبيق خوارزميات التصنيف والتجميع للحصول على المعلومات التي تساعد في اعطاء رؤية واضحة عن الجرائم والمجرمين. وقد استخدمت برامجيات WEKAو Excel لمعالجة وتحليل تلك البيانات.