# Design and Implement Proposed Crime Analysis using Modified Association Rule

**Emad K. Jabar[1] Soukaena H. Hashem[1] Enas M. Hessian [2]**

[1]Computer Sciences/ University of Technology/ Baghdad-Iraq

[2] Computer Sciences/ Al-Mustansaria University / Baghdad-Iraq

## Abstract

This research presents a proposal to advance crime analysis that through employee data mining association rules on crime's data with a proposed strategy consists of three levels, each level present suggestion to suite and consistence crime analysis and predictions. First level will deal with the challenges in mining crime data, where the last often comes from the free text field. While free text fields can give the newspaper columnist, a great story line, converting them into data mining attributes is not always an easy job. The proposal will look at how to arrive at the significant attributes for the data mining systems. That through suggested view organized the crime to three dimensions these are crime attributes, criminal attributes and geo-crime attributes. Second level will use AR (apriori) as a miner technique of crimes, but apriori in case of large dataset is not efficient, also has no security to protect the mined data from unauthorized users. The proposal modify apriori (MAR) to avoid the degradation of performance with crime analysis by reduce unimportant and redundant transactions. Advance MAR with modest suggestion to be secure. Third level, applying the MAR on each dimension separately then according need and on demand of correlate among these dimensions, the correlation done using proposed mixing.

The proposal applied on real crime data from a dependable sheriff's office depended in our previous work (reference 6), then a comparison done between the previous and current work. The results of comparisons show the current work advance previous work by optimizing time and space consumed in mining through apply suggested MAR in current work, where the previous work apply traditional apriori AR. Also the proposed MAR give precision in prediction since it omitting the redundant and ineffective data.

**Keywords:** Crime Analysis, Criminal, Data Mining, AR, apriori.

**المستخلص**

يقدم هذا البحث اقتراحا لتعزيز تحليل الجريمة من خلال قواعد ارتباط تعدين البيانات، تتكون الاستراتيجية المقترحة من ثلاثة مستويات، كل مستوى يقدم اقتراح مناسبة لتحليل الجريمة والتنبؤ بها. يتناول المستوى الأول بيانات الجريمة في مجال التعدين، وهذا المستوى يبحث في كيفية التوصل إلى سمات هامة للجريمة التي تستخدم في أنظمة تعدين البيانات. المقترح ينظيم بيانات الجريمة في ثلاثة أبعاد، سمات الجريمة، سمات المجرم والسمات الجغرافية للجريمة. والمستوى الثاني استخدام قواعد الارتباط (نحو استدلالي) كأسلوب منجم للجرائم، ولكن في حالة البيانات الكبيرة هذه الطريقة ليست فعالة، وأيضا لا توجد حماية للبيانات المعدنه من المستخدمين الغير مخوليين. البحث يقوم باقتراح (MAR) لتجنب تدهور الأداء مع تحليل الجريمة من خلال الحد من المعاملات غير المهمه والزائدة عن الحاجة.

المستوى الثالث تطبق MAR على كل بعدا بشكل منفصل وفقا للحاجة والطلب على وجود علاقة بين هذه الأبعاد. الطريقة المقترحه تم تطبيقها على بيانات الجريمة الحقيقية من مكتب مأمور شرطة وتم الاعتماد على هذه البيانات في العمل السابق (مرجع 6)، ثم مقارنة ذلك بين العمل السابق والحالي. نتائج المقارنات إظهرت ان العمل الحالي يتسم بفعالية أكثر من العمل السابق عن طريق الاستفادة المثلى من الوقت والمساحة المستهلكة في مجال التعدين, كذلك الدقة في التنبؤ لأنه بحذف البيانات المكررة وغير فعالة.

## 1. Introduction

Crime is an intentional act in violation of criminal law committed without defense or excuse, and is penalized by the state as a felony or misdemeanor [1]. Crime management is defined as controlling, directing, and coordinating police resources (money, equipment, and personnel) to prevent the violation of law and where it has been violated, to apprehend the criminals and take them to court and recover the stolen property [2]. Crime is a major issue where the top priority has given by our government. Law enforcement agencies like that of police today are faced with large volume of data that must be processed and transformed into useful information. The idea here is to try to capture years of human experience into computer models via data mining [3, 4]. The nature of each crime is also categorized by a well-defined taxonomy used by various anticrime and anti-terrorism agencies across the globe. But such a rich data set loses its usefulness if we do not know what to look for. Ideally, with so much of information about each place and associated crimes, citizens could infer the degree of inhabitability of a particular area and so on. People involved in law enforcement could use identified hotspot areas to fight crime using a more targeted approach [5].

## 2. Related work

***In [6] Jabar E. K. et.al.,*** propose three correlated dimensional model; crime, criminal and geo-crime. By apply the secure AR data on each of the three correlated dimensions separately then using Genetic Algorithm GA as mixer of the resulted ARs to exploit the relational patterns among crime, criminal and geo-crime to help to detect universal crimes patterns and speed up the process of solving crime with more accurate. ***In [7] Malathi A. et. al.,*** they discuss that, a major challenge facing all law-enforcement and intelligence gathering organizations is accurately and efficiently analyzing the growing volumes of crime data. There has been an enormous increase in the crime in the recent past. They look at MV algorithm, DB Scan and PAM outlier detection algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. ***In [8] Mande U. et. al.,*** introduce binary clustering and classification techniques have been used to analyze the criminal data. The crime data considered in this paper is from Andhra Pradesh police department this paper aims to potentially identify a criminal based on the witness/clue at the crime spot an auto correlation model is further used to ratify the criminal. ***In [9] Malathi. A et. al.,*** they use a clustering/classify based model to anticipate crime trends. The data mining techniques are used to analyze the city crime data from Police Department. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years. ***In [10] Sathyaraj S. R. et. al.,*** they studies to integrate a large volume of data sets into useful information by adopting a various information techniques in the hottest technology world. The adopted approaches of Single variate Association Rule for Crime to Crime based on the knowledge discovery techniques such as, clustering and association-rule mining. The present study of this paper was focuses through the real crime dataset by using various algorithms. ***In [11] Mande U. et. al.,*** they aims towards the construction of new methodologies based on Data mining concepts and serves as a decision support system. Given a set of available clues, from the forensic labs and the clues collected at the crime spot, a methodology is presented to map the evidence and identify a criminal. ***In [12] Chen N. et. al.,*** their model, firstly, we predict the residence of the offender based on the locations of the last crime scenes with three methods (distance analysis: the location that has the shortest distance to each crime site, circle fitting, probability theory); secondly, they predict the time of the next crime based on previous data with the method of fitting a straight line; next, predict the location of the next crime based on the locations of the last crime scenes and the time predicted in the second step with the method of weighted average; finally, generate a predicted location based on the three predicted locations with

the method of weighted average. ***In [13] Yu C. H. et. al.,*** they discuss the preliminary results of a crime forecasting model developed in collaboration with the police department of a United States city in the Northeast. They first discuss approach to architecting datasets from original crime records. Additional spatial and temporal features are harvested from the raw data set. Second, an ensemble of data mining classification techniques is employed to perform the crime forecasting.

## 3. The Proposal of Crime Analysis

To explain the proposal of crime analysis and prediction using MAR-Mixer data mining on suggested three dimensions, will introduce figure (1), then explain it is details:

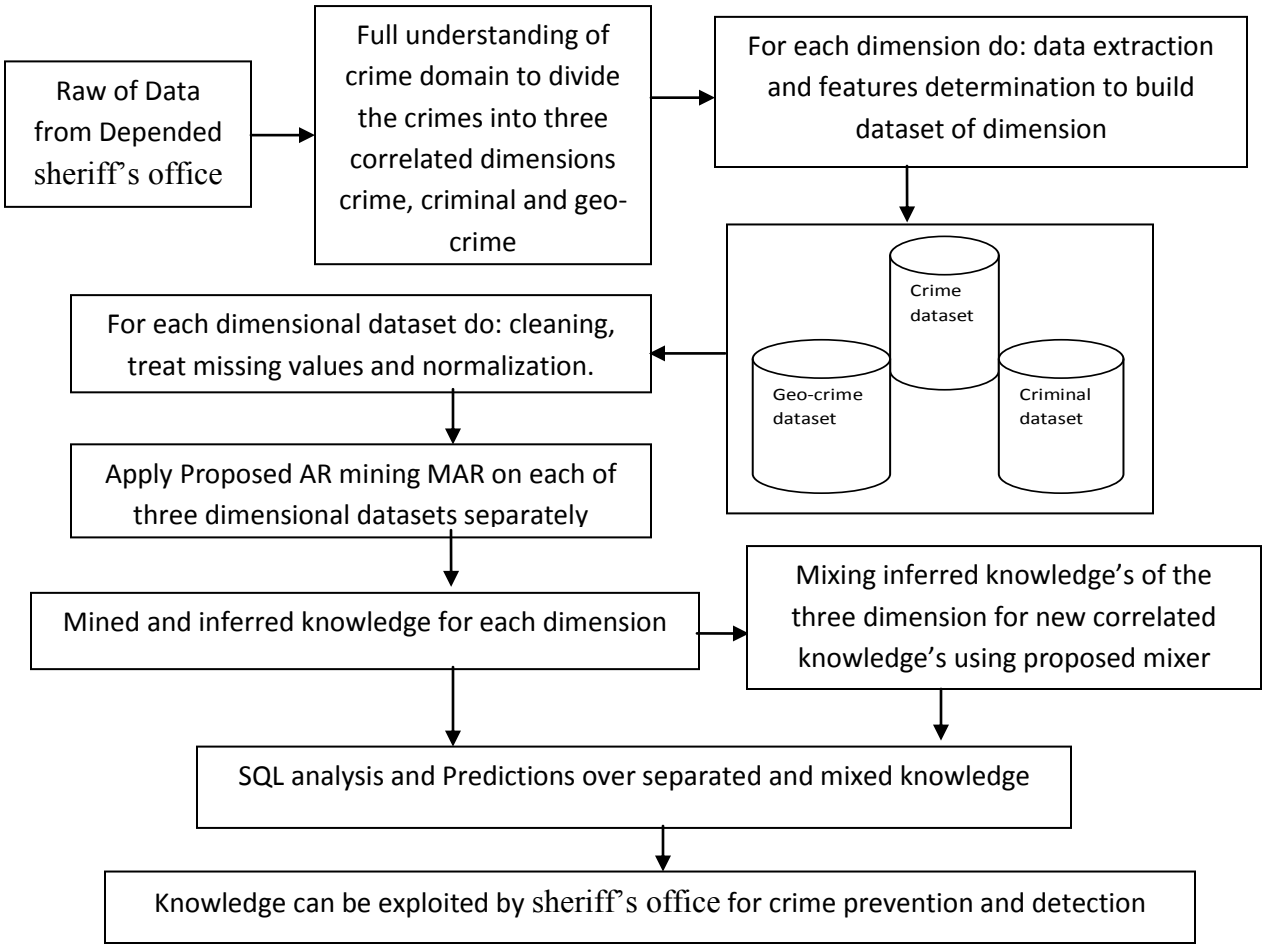


Figure (1): a proposed three correlated dimensional MAR-Mixer system.

### 3.1 Understanding the Crime Domain

Crime is neither systematic nor entirely random, so crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection. The major challenges are encountered. Crime information volume has been increased. Different methods and structures used for recording crime data. The data available is inconsistent and are incomplete thus making the task of formal analysis a far more difficult. The main focus is to develop a crime analysis tool that assists the police in: To perform crime analysis to detect crime patterns. Provide information to formulate strategies for crime prevention and reduction. Identify and analyze common crime patterns to reduce further occurrences of similar incidence. The proposal has main objectives of crime analysis can be classified into: Extraction of crime patterns by analysis of available correlated crime, criminal and geo-crime data. Prediction of a crime based on the correlation of existing data and anticipation of crime rate using data mining techniques.

**3. 2 Extracting the Target Dataset**

Here will see any crime investigation highlights primarily on three dimensions; these are crime dimension, criminal dimension and crime-geo dimension. For each dimension will select the most critical attributes (variables) are very interested and repeated in crime registration. Each dimension is a dataset has it is own attributes but all of the three datasets are correlated such that each transaction in them are related to one crime considering it is dimensions. Now will display the considered attribute in each dataset of dimensions: ***Crime dataset*** take the following attributes (crime_id, crime_type, crime_location, crime_date, crime_weapon, crime_victum, crime_witness, crime_clues). ***Criminal dataset*** take the following attributes (criminal_id, criminal_gender, criminal_age, criminal_address, criminal_income, criminal_job, criminal_maritalstatuse, criminal_signs_diffrence, criminal_ religon, criminal_natioal) ***Crime-geo dataset*** take the following attributes (geo-id, geo-population (high, small), geo-size (large, small), geo-type (open, closed), geo-earth (agriculture, industrial), geo_longitude, geo_latitude).

**3.3 Data Preprocessing**

The preprocessing systems include the following tasks:

***Data cleaning:*** fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies (there are many modest proposals for filling missing values).
Different preprocessing techniques were used to get clean data, these include: Using KNN-based imputation method, in this method, the missing values of an instance are imputed by considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a Euclidian distance function.

***Data integration:*** using multiple databases, data cubes, or files (since data are collected from many of sheriff's office so, the data are integrated to build uniform three datasets. That by uniform the selected attributes and uniforms the types of value's attributes.).

***Data transformation:*** normalization and aggregation (in system where the AR is the used technique for mining the attributes are all converted to letters appear if agree with attribute condition and disappear if not agree).

***Data reduction:*** reducing the volume but producing the same or similar analytical results (in system reducing done in two faces these are: face of omit some unusefull attributes such as criminal color eyes which the criminal could change it. The second face is omitting entire records because they have more than three missing values so the filling will cause noisy.).

***Data discretization:*** part of data reduction, replacing numerical attributes with nominal ones (in system for example the attribute of age instead of filling it as number will be replaced by letter such as A, where A will appear if criminal age more than 35 year and disappear if age smaller).

### 3.4 Modified Association Rules MAR

The efficient discovery of such crime, criminal and geo-crime rules has been a major focus in proposed data mining. Many algorithms and approaches have been proposed to deal with the discovery of different types of association rules discovered from a variety of databases. However, the proposed datasets relied upon are alphanumerical and often transaction-based. The problem of discovering association rules is to find relationships between the existence of a crime-attribute (or characteristic) and the existence of other crime-attributes (or characteristics) in a large repetitive collection. Then after finding all

association rules from the three datasets separately will mixing these rules to extract new rules to correlate these three datasets using mixing algorithm.

Because of large number of records in database results in much more space and time consuming. Here will propose suggestion to optimize apriori algorithm which reduces the size of database. The suggestion adds related table (relational model) has one attribute called Transaction Length (TL), containing number of items in specified transaction in database. The process of omitting a transaction in database will made according to the value of Z. For any value of Z, algorithm will search the same value for TL in database. If value of Z matched with value of TL then omit only those transactions from database. Algorithm (1), will explain all details of MAR.

**Algorithm (1): The Proposed MAR**

**Input**: Database of transactions (D) and minimum support threshold (*min_sup*).

**Output**: Frequent itemsets in D (L).

**Process:**

```
L1=find_frequent_1-itemsets(D);
For(k=2;Lk-1≠∅; k++)
{ Ck=apriori_gen(Lk-1, min_sup); // see algorithm (1-1)
for each transaction t∈D
{ Ct=subset(Ck,t);
for each candidate c∈Ct
c.count++;  }
Lk={ c∈Ck |c.count≥min_sup };
if(k>=2)
{  Omitting Itemset (D, Lk, Lk-1);  // see algorithm (1-2)
 Omitting Complete Transaction (D, Lk);       // see algorithm (1-3) }   }
return L=UkLk ;
```

**End of Process.**

**Algorithm (1-1): apriori_gen**

**Input**: Frequent(k-1)-itemsets (Lk-1) and *min-sup*

**Output**: Ck

**Process:**

```
for each itemset l1∈ Lk-1
{ for each itemset l2∈ Lk-1
```

{ If (l1 [1]= l2 [1])∧ (l1 [2]= l2 [2]) ∧…∧(l1 [k-2]= l2 [k-2]) ∧(l1 [k-1]< l2 [k-1]) then
{ c=l1 ∞l2;
for each itemset l1∈Lk-1
{for each candidate c ∈Ck
{ if l1 is the subset of c then
c.num++; } } } }
C'k={ c∈Ck |c.num=k};
return C'k;

**End of Process.**

**Algorithm (1-2): Omitting Itemset**

**Input**: Database of transactions (D), frequent(k)-itemsets (Lk) and frequent(k-1) – itemsets (Lk-1)

**Output**: D with deleted values of attributes

**Process:**

for each itemset i ∈Lk-1 and i ∉ Lk
{ for each transaction t∈D
{ for each value of attribute ∈t
{ if (value of attribute =i)
update value of attribute =null; } }}

**End of Process.**

**Algorithm (1-3): Omitting Complete Transaction**

**Input**: Database of transactions (D) and frequent(k)-itemsets (Lk)

**Output**: D with deleted records

**Process:**

for each transaction t∈D
{ for each value of attribute ∈t
{ If (value of attribute!=null and value of attribute!=0 )
{ Record-of-data.count++; } }
If (Record-of-data.count<k)
{ delete Record-of-data; } }

**End of Process.**

In this proposal will using Mixing algorithm will be applied on selected association rules resulted from the three datasets as in the following proposed steps:

1. A mixing representation or encoding schema for potential solutions to the problem. Each association rule will be presented as a series of numbers (all

alphabets representing the attributes will be encoded by numbers each number has two digit since will have many attributes distributed over all the three datasets. Such as A=01, B=02, and finally the symbol ---> will take the number 00). For example the association rule from the third dataset is UVW---> XZ has the following encoding (212223002426).

2. One way to create an initial population of potential solutions, the initial population already created with association rules algorithms which established on the three datasets. So this means the initial population of Mixer will be the selected association rules extracted by the three datasets separately and encoded as series of numbers.
3. An evaluation function that pays the role of the problem environment (novel association rules), rating solutions in term of their "fitness", Here the proposed evaluation function for each rule is consist of three parts, these are:
   - Each number in each series is a two digit.
   - Each number will appear only on the left or right of zero.
   - Confidence of the rule must pass the minimum confidence.
4. Mixing operators that alter composition of offspring. One-point crossover is the most basic crossover operator, where a crossover point on the mixing code is selected at the zero number which occur in the series of numbers, and two parents rules are interchanged at this point.
5. Crossover exploits existing rule potentials, but if the population does not contain all the encoded information needed to find the novel rules, no amount of rules mixing can produce satisfactory solution. For this reason a mutation operator capable of spontaneously generating new frame is included. The most common way of implementing mutation is to flip a bit with a probability equal to a very low, given Mutation Rate (MR). A mutation operator can prevent any single bit from converging to a value through the entire population and, more important, it can prevent the population from converging and stagnating at any local optima.
6. Values for the various parameters that used by the mixing algorithm (population size, rate of applied operators, etc.). For particular problem will use the following parameters of the mixing algorithm: population size (pop-size) = 3000 (the

parameter was already used) for each dataset will take the 1000 higher confidence rules. Probability of crossover (PC) =1, probability of mutation (PM) = 0.001 (the parameter will be used in mutation operation).

7. Continue with mixing processing until the optimized rules will be optimized to be the novel rules.

## 4. Discussion and Results Experimental works

As in our previous work prove idea of building proposed three correlated dimensional system for advancing crime analysis and prediction will do the following experimental works. Collecting data crimes from three sheriff's office, these data was collected from the year 2003 to year 2013. These data are divided in to two parts these parts are: First part was from 2003 to 2010, this part conducted to construct proposed system. Second part was from 2010 to 2013, this part conducted to verify the analysis and predictions discovered from applying proposed system on the first part data. Also as in previous work verify the constructed system analysis and predictions on second part of data will see that, system advances the following crime analysis: Tactical crime analysis since 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. Strategic crime analysis since it 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. Administrative crime analysis since it is focuses on providing summary data, statistics, and general trend information to police managers. Investigative crime analysis since involves profiling suspects and victims for investigators based on analysis of available information. Intelligence analysis since it is focuses on organized crime, terrorism, and supporting specific investigations with information analysis and presentation. Operations analysis since examines how a law enforcement agency is using its rescues. It focuses on such topics as deployment, use of grant funds, redistricting assignments, and budget issues. Also system proves and meets the following: By crime analysis will 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. Asses' crime

through analysis helps in crime prevention efforts. Preventing crime costs less than trying to apprehend criminals after crimes occur. Will arrive at the significant attributes for the data mining systems, since there are some attributes never appear in prediction and analysis. By analyze crimes will could inform law enforcers about general and specific crime trends, patterns, and series in an ongoing, timely manner. The comparisons done between our previous work [6] and our current proposal in this research according the following measures: Time measure: this measure calculate the time consumed in mining the same size of three dimensions (three datasets those are crime, criminal and geo-crime) in both our previous work [6] and our current proposal, see figure (2), which explain the superior of the current work in minimize time consuming in mining using proposed MAR.

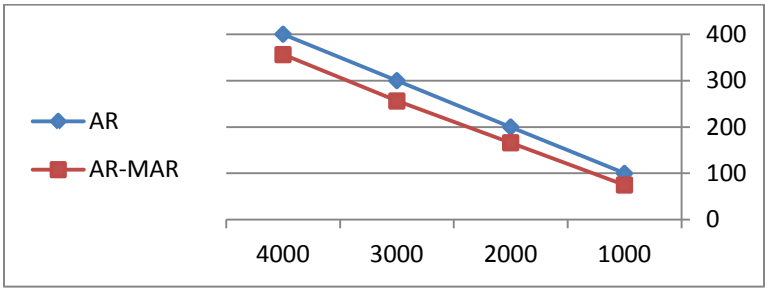


Figure (2): comparison between AR and MAR in time consuming with four cases of data (X-axis) and time is measured in seconds (Y-axis)

Space measure: this measure calculate the space consumed in mining the same size of three dimensions (three datasets those are crime, criminal and geo-crime) in both our previous work [6] and our current proposal, see figure (3), which explain the superior of the current work in minimize space consuming in mining using proposed MAR.

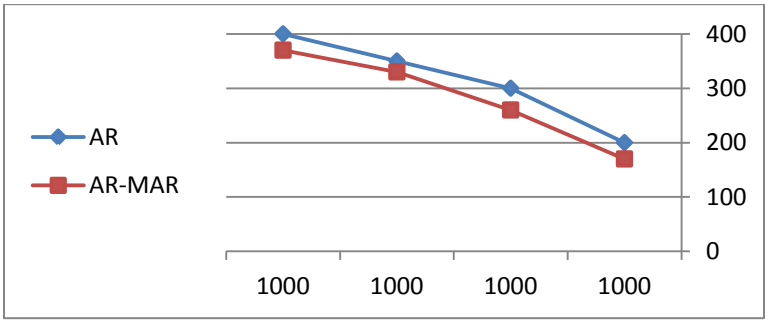


Figure (3): comparison between AR and MAR in space consuming with four cases of data (X-axis) and space is measured in KB (Y-axis)

Precision measure: this measure calculate the precision of prediction in mining the same size of three dimensions (three datasets those are crime, criminal and geo-crime) in both our previous work [6] and our current proposal, see figure (4), which explain the superior of the current work in maximize the precision of prediction in mining using proposed MAR.

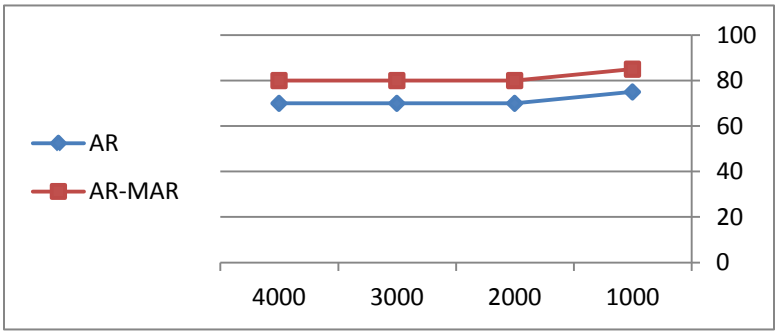


Figure (3): comparison between AR and MAR in precisions of prediction with four cases of data (X-axis) and precisions is measured in % (Y-axis)

## 5. Conclusion and Recommendation

This research reaches the conclusions:

Considering the three dimensions of the crime model make our previous and current proposal can help the police to find the criminal quickly by predicting location of the next crime and the residence of the criminal. The proposed three correlated dimensional model MAR-Mixer of identifying a criminal, in the absence of witness or any clue by the forensic experts. In these situations, here we have tried to identify the criminal by correlate the criminal with crime and location of the crime using mining and genetic. The advantages of using KNN imputation in preprocessing are; KNN can predict both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean among the k nearest neighbors). It can easily treat instances with multiple missing values. It takes in consideration the correlation structure of the data. After extracting the association rules from each dataset separately we tend to mix these associations rules using proposed mixer. Customizing the mixer to suite our proposal model, that by propose a schema for encoding the rules, and then making the initial pool is all the extracted rules from the three datasets. Making the crossover point is ----> provide much justify in generating new child rules from parent rules. Fitness function customizes to satisfy the basic conditions in building the association rules according data mining techniques.Proposed MAR-mixer avoided the limitations of our

previous proposed model AR-GA include that crime pattern analysis can only help the detective, not replace them. And data mining is sensitive to quality of input data that may be inaccurate, have missing information, be data entry error prone etc. Also mapping real data to data mining attributes is not always an easy task and often requires skilled data miner and crime data analyst with good domain knowledge. They need to work closely with a detective in the initial phases. Finally the proposed MAR was the base stone of superior our current work over our previous work in optimize the time, space and precisions of prediction.

## References


1. Chen H. , Chung W. , Qin Y., Chau l. , Xu J.ennifer, Wang G., Zheng R., Atabakhsh H., "Crime Data Mining: An Overview and Case Studies", AI Lab, University of Arizona, proceedings National Conference on Digital Government Research, 2003, available at: http://ai.bpa.arizona.edu/
2. Fayyad U.M. and Uthurusamy R. ," Evolving data mining into solutions for insights". Communications of the ACM, 45(8), 28-31, 2002.
3. Chau M., Xu J., and Chen H., "Extracting meaningful entities from police narrative reports". In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA.
4. Nath S. V., "Crime Pattern Detection Using Data Mining", Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology , 41- 44, 2006.
5. Yan J. , Liu N., Yang Q., Zhang B., Cheng Q., Chen Z., "Mining Adaptive Ratio Rules from Distributed Data Sources", Data Mining and Knowledge Discovery, 12, 249–273, 2006, 2005 Springer Science+Business Media, Inc. Manufactured in the United States. [IVSL]
6. Jabar E. K., Hashem S. H. and Hussein E. M., " Propose Data Mining AR-GA Model to Advance Crime analysis", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 5 (Sep. - Oct. 2013), PP 38-45 www.iosrjournals.org.
7. Malathi. A and. Baboo S. S., "Enhanced Algorithms to Identify Change in Crime Patterns", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, No.3, Sep-Dec, 2011, pp. 32-38, ISSN: 2007-1558.
8. Mande U., Srinivas Y. and Murthy J.V.R., "An Intelligent Analysis Of Crime Data Using Data Mining & Auto Correlation Models", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 4, July-August 2012, pp.149-153 149 | P a g e
9. Malathi. A , Baboo S. S. and Anbarasi A., " An intelligent Analysis of a City Crime Data Using Data Mining", 2011 International Conference on Information and Electronics Engineering, IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore.
10. Sathyaraj S. R., Thangavelu A., Balasubramanian S., Sridhar R., Chandran M. and Prashanthi M. D., "Clustered Spatial Association Rule To Explore Large Volumes Of Georeferenced Crime To Crime Data", 3rd International Conference On Cartography And Gis 15-20 June, 2010, Nessebar, Bulgaria.
11. Mande U., Srinivas Y. and Murthy J.V.R., "Criminal Mapping Based On Forensic Evidences Using Generalized Gaussian Mixture Model", The International Journal of Computer Science & Applications (TIJCSA), Volume 1, No. 4, June 2012 ISSN – 2278-1080, Available Online at http://www.journalofcomputerscience.com/
12. Chen N. and Wang Y., " Prediction of Series Criminals: An Approach Based on Modeling", 2010 International Conference on Computational and Information Sciences.
13. Yu C.H., Ward M. W., Morabito M., and Ding W., " Crime Forecasting Using Data Mining Techniques", 2011 11th IEEE International Conference on Data Mining Workshops.