International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 24 (2015) pp 44597-44602
© Research India Publications. http://www.ripublication.com

An Efficient Density Based K-Medoids Clustering With Expectation-Maximization Algorithm to Analyze Crime Data

MuhammedShafi. P[1]
Research Scholar [1]Assistant Professor, [1]Department of Computer Science, [1]N.A. M college Kallikkandy, Kannur, Kerala, India, [1]pullaratshafi@gmail.com

Selvakumar.S[2]
[2]Department of Computer Science, [2]Peryar University, Salem, [2]infoselva@yahoo.co.in

Periyasamy.S[3]
[3]Department of Computer Science, [3]PRUCAS, Harur,Dharmapuri, [3]periyasamysps@gmail.com

Abstract

Data mining is the process of discovering unknown patterns from the huge volume of data,it is used for analyzing countless set of data and then extracting process. In data mining, crime management is a vital application where it plays a significant role in handling of crime data prediction and analysis. In recent year enormous crimes are increased and it had been social nuisance. In any country crime, investigation is a very important role, but huge crime data management is one of the challenging process of police department. The rapid popularity of the mobile phones and the internet, the crime data can be sent to the police message centers through SMS or e-mail. But, manual processing of huge volumes of crime data is a difficult process. Thus, in this paper proposed a clustering technique for the crime data analysis. An efficient Density Based K-Medoids (DBK) clustering algorithm is used for cluster the crime data and Expectation-Maximization Algorithm improves DBK clustering process. The Clustering process clusters the SMS and e-mail into subgroups, which are used toredirect the SMS and e-mail to the section where action are taken based on the crime. This proposed system helps the police crime data analysis system for better prediction and classification of crimes.

Keyword: Crime management, Data mining, Density Based K-Medoids (DBK), Expectation-Maximization Algorithm.

1. Introduction

Data mining process is one the emerging field that can be utilized in a wide range of applications likebanking,marketing,health insurance,city planning and so on. Recently, the crime data analysis is one of the significant application in data mining field.Data mining approach comprises numerous techniques and tasks containing Association, Clustering, Link Analysis Prediction and Classification. Each of them has itsown applications and importance.

Crime analysis permit to analysis of the huge volume of data. Crime data analysis is very important for law enforcement department and police departments. Normally, the crime analysis involves in mapping crime data for command staff and generating crime statistics. Additionally, the crime analysis approach mainly focusing on analyzing different suspect information and police reports to help investigators in main crime units to detect the serial robbers and so on.Crime analysis is acting as one of the efficient method of analyzing crime. More particularly, crime analysis is to find knowledgeable information in a huge volume data and distribute this information to investigators. This process leads to avoid suppressingcriminal activity and apprehend criminals and it also helps to prevent the crime activity.

Additionally, the rapid popularity of the mobile phones and the internet, the crime data can be sent to the police message centers through SMS or e-mail. But, manual processing of huge volumes of crime data is a difficult process. Thus, in this paper proposed a clustering technique for the crime data analysis. An efficient Density Based K-Medoids (DBK) clustering algorithm is used for cluster the crime data and Expectation-Maximization Algorithm improves DBK clustering process. The Clustering process clusters the SMS and e-mail into subgroups, which are used to redirect the SMS and e-mail to the section where action are taken based on the crime. This proposed system helps the police crime data analysis system for better prediction and classification of crimes.

2. Related work

In [4] author presented a model for criminal and crime data analyzes utilizing simple K- Means algorithm using data Aprior and clustering algorithm with Assoication rules. This process helps to find the trends and patterns, finding possible explanation and relationships, making forecasts, identifying possible suspects and mapping criminal networks. The clustering process depends on the different criminal and crime attributes and unknown common characteristics. To take any decision about the crime by make use of association rules mining on crime dataset which help to create a secure society and prevention action on crime. Here the police department in a Libyapolice department in the Libya data set is used for processing.

In [5] binary clustering and classification approach has been utilized to analyze the criminal data. Here the crime data are collected from the Andhra Pradesh police department. Mainly in this paper concern about the security issues and potentially identify a criminal and their witness and clues. An auto correlation model is used for crime spot is further utilized to ratify the criminal. Thisproposedwork is to effectively avoid the hindered crime analysis.

In [6] crime prevention and analysis are done by using systematic approach which creates the analyzing trends and patterns of crime. This system can help crime data analysts and Law enforcement officers to accelerate the process of resolving crimes. This system mainly focuses on visualizing crime prone areas and high probability crime occurrence area. Utilizing the data mining concept between criminal justice and computer science to solve the crimes faster.Causes of crime occurrence like political enmity,criminal background of offenders, etc. Here the crime factors of each day, take an account for processing.

In [7] data mining process employed in the context of intelligence analysis and law enforcement analysis. The paper concern about the national security attacks in Mumbai since the 26/11. However, enhancing technology and information are the drawback of effective analysis of the terrorist and criminal activities. Thus, a classify/clustering based approach is to apply on the anticipate crime trends. Here the crime data are obtained from the Tamil Nady Police Department. The proposed system shows the potentially lessen time to take prevent the crimes.

In [8] author mainly focuses on the influence of neighboring state crime rate with the reference state utilizing spatial data mining techniques. In last few years spatial data analytical process is tremendously increased in the field of decision making process. Maintain law and order, regional governance, disaster prediction and so on.In this study different types of factors are taken for the analyzing process such as literacy-rate, GDP, police-rate, various crimes like dacoit, murder, and riots and the state as location data and Employment-rate. The correlation between different crimes find by using spatial autocorrelation and compare the different clustersattribute with their relation. Here, analyzing process using the states of India crime data in the year of 2012. The clustering process is utilized to find the patter with the various Police force distribution, Employment and crime densities. The outcome of this proposed approach reveals that the crimes of the state of India have positive spatial correlation. The crime clusters can be utilized for creating different security measures in the states.

Nowadays computer crimes are drastically increased and it provides harm to computer as well to a human being. But it is very rigid to inspect such kind of case in a small period of time.Thus, in order to analyze digital forensic along with multi-relation classification is utilized in whichfacilitates pattern mining from multiple tables. The author in [9] examined digital forensic associated with log files to investigate the crime scene and the author also précised the file with fuzzy rule.

3. Research Problem

The Kerala Government announced a cash award to those who provides clues or evidences through photos taken with mobile phone camera which will help police to solve crimes. This technique is very helpful to the police in the present condition of our state, where the crimes have come with a high degree. It also cherishes the cooperation of people. The crime information can be sent to the police message center through SMS or e-mail. Now they can handle manually by reading all the content and send to the corresponding section (traffic, crime, etc) for further enquiry. But, manual processing is difficult with the large volume of SMS processing. So the computers can process thousands of SMS in seconds. Furthermore, running and installing software often less costs that training and hiring personnel. The computers, prone to less errors than humans, especially those who work long hours. But here different kind of issues arising such as proprietary file formats, lack of linguistic tools for investigative purposes, lack of SMS authorship, SMS has unknown written language, short form issues in SMS, noun or verb detection, visualization reporting of SMS analysis.

4. Database Representation

The crime data analysis presented in this proposed work is based on traffic data given bytire-1 cellular operator in the Kerala Government. The sample crime data comprises Call Detail Records (CDR) of 10000 crime account and almost 20000 genuine accounts. Additionally, includes about 8000 machines to machine devices and 15000 post-paid family plans, from the two year period between Jan 2012 and Jan 2014. In CDRs different types of records are included such as log each phone call, the network exchange data, a text message data, etc. If same providers communicating two ends which means a duple record are stored.The Mobile Originated (MO) is used for records the log data for transmitting party and receiver log informationisstored in Mobile Terminated (MT). If MT and MO records have same transaction means it contains the duplicate data such as terminating number and originating number. The Internet Protocol (IP) data traffic creates only MO logs.

In the collected dataset was utilized for testing and training data. These data contain both criminals and crimes with the their different attributes as shown in table 1.

Attributes	Description
CrimeID	Individual Crimes are labelled by unique Crime IDs
CrimeName	Disguised crime’s name
CrimeType	Indicates crime type.
CriminalID	Individual Criminals are labelled 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	Marital status of the criminal.
Mobile No	Individual phone No
Mobile Network	Mobile network type
Mobile connection	Indicate postpaid, or prepaid
Text message	Detail of text message
Time	Text message received and send Time
Call type	SMS/Voice calls
IMEI No	Mobile IMEI number
Duration	Mobile calls time duration

5. Data Preprocessing

Normally, the real world data have some of the drawbacks such as Inconsistence,Noisy and Incompleteness [10]. So these data have to be preprocessed and the preprocessing tasks include following subtasks such as Data cleaning,Data transformation,Data integration,Data reduction,Data discretization. Different kind of preprocessing techniques are applied in this preprocessing such as Removing outliers,Filling missing data, Data reduction utilizing aggregation and normalization. The outliers and missing values in both criminal and crime data set have absolutely various meanings than in many other datasets.For instance, finding an unknown criminal address tasks from the database is difficult,so unknown data removed from the database.

6. Density Based K-Medoids (DBK) For Criminal Investigation

Data mining is an efficient technique and it has excessive potential to simply the criminaldata investigation on the most significant crime data [14].The discover the knowledge from existing crime data is one of the value-added advantages. The successful data mining approach depends on the specific choice of methods utilized by analysts. In this proposed work has constructed a framework for mining data with the intention ofcatching professional criminals. In this section,Density Based K-Medoids (DBK)which is aimed to discover the clustersby using crime data (SMS).

K-medoid is one of the standard partitioning technique of clustering, which clusters the crime dataset of n objects into K number of clusters [11]. The algorithm operates on the minimizing principle with a sum of dissimilarities between each and every object and their corresponding reference point. The DBK algorithm randomly selects the n objects in crime dataset D and the initial representative object is named as medoids. The mediod can be represented as cluster object, which object in the cluster has minimal dissimilarity i.e. it is a most centrally placed in the crime data set. After that, all the dataset objects are assignedto their corresponding object to the nearest cluster relied upon the object’s distance to the medoid of the cluster. A new mediod is decided after assigning each data object to corresponding cluster.

The Density Based K-Medoids creation does not create the same result with iteration, due do theirresulting cluster is relied upon the initial random assignments. The DBK algorithm creates more robust clusters in the crime data presence of outliners and noise data. But, the optimal number of clusters $k$is hard to project. Thus, in this proposed work use Expectation-Maximization Algorithm to improve the DBK cluster results.

In fig. 1 shows the two contiguous uniformly distributed cluster areas such as $R1 \text{ and } R2$, such that $R2$ is α time denser than $R1$, with $\alpha > 1$. The minimum difference in cluster density is indicated by using α. The two cluster regions will be combined into a single cluster. Here assume the present object to be treated, p is located between boundary of two regions. The density of any object between q and p will be greater than$m$ but less than $(1+\alpha)\frac{m}{2}$. Similarly, the density of each objectbetween$p$ and r will be greater than$(1+\alpha)\frac{m}{2}$but less thanαm. When a transition region is met cluster growth in that particular direction, the cluster formation may get stopped. A cluster transition region may be explored while going from a minimum density region to a greater density region or from greater density region to minimum density region. Thus, two various density factors α_1 and α_2 are required to eliminate the order of dependency.

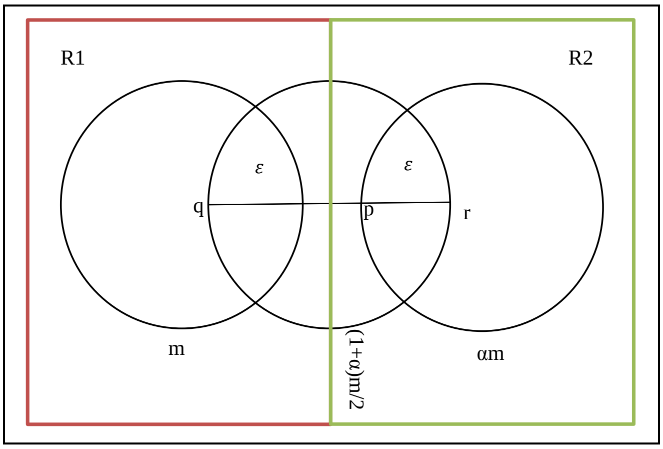

Fig.1 Density variation

6.1 Expectation-Maximization

In this proposed crime data analysis method making an initial cluster partition to be utilized by Expectation-Maximization clustering algorithm. The presented EM algorithm work based on the cross entropy approach. Initially, the clusters are created by using density estimation perception. Here the data is treated as independent data obtained from a combined population, while different types of (cluster identifiers and Attributes) labels are hidden. Particularly, consider the crime data $\{x_1, \dots x_m\}$ to be a set of cluster vectors in a cluster subset X is in dimensional Euclidean space(cluster region) R^n having clusters $\{C_j\}, j = 1, \dots k$ then the primary distributionμ of X is defined as follows

$$\mu = \sum_{j=1}^{k} P_j, \mu_j \qquad (1)$$

Where set $P = \{p_j, j = 1, \dots k\}$ is defined as cluster probabilities and $\mu_j\, j = 1, \dots, k$ is defined as the inner cluster distributions.

The EM clustering algorithm suggests the Cross-Entropy (CE) method of data fitting [12].In EM algorithm one of the generic approach is the Cross-Entropy (CE) method [13]. The general working procedure of EM method is as follows

Cross-Entropy method

Step 1: Initially, create a random sample data the allocate the parameters for each sample which are underlying distributed.

Step 2: Update the parameters with the purpose of generating a "better optimal" sample in the succeeding iteration

Step 3: Iterate the process until the cluster is "stabilized".

The EM clustering algorithm has been shown to be robust regarding of initial mediods creation. This process may be considered as an optimization issue. The initial mediod creation is as follows

$$L = \min_{c_1 \dots c_k} R(c_1 \dots c_k) = \min_{c_1 \dots c_k} \sum_{j=1}^{k} \sum_{x \in C_i} \|x - C_j\|^2 \qquad (2)$$

Where$c_1 \dots c_k$are defined as the decision variables, in this situation The mediods of the clusters are defined as $\{C_j\}, j = 1, \dots k.$

Density Based K-Medoids (DBK) with Expectation-Maximization algorithm

Step 1: Let$D$ is the Crime Dataset with k points and k defined as the obtained number of clusters.

Step2: Letε defined as Euclidean neighborhood radius and η defined as Minimum number of cluster neighbors needed in ε and p characterized as any point in Crime Dataset$D$ and N is defined as a set of points in ε neighborhood object of p.

Step 3: For each unvisited point p in dataset D

```
{
$c = 0$
$N = getNeighbors\ (p\ and\ \varepsilon)$
$if\ (sizeof\ (N)) < \eta$
$make\ p\ as\ Noise$
$++c$
}
{
$mark\ p\ as\ visited$
$add\ p\ to\ cluster\ c$
$recurse\ (N)$
}
```

Step 4: To find obtained M clusters centers c_m by taking equation 2.

Step 5: Find the total number of optimal points in each cluster.

Step 6: If$m > k$, choose two clusters based on a number of points satisfying the pre-defined crime rules criteria and densityvalue.Then, joint these two cluster centers.

Step 7:repeat until attaining k clusters.

Step 8:else $l = k - m$, choose two clusters based on a number of points satisfying the pre- defined crime rules and density value. Then, split the cluster by utilizing $kmedoids$ clustering algorithm.

Step 9:repeat until attaining k clusters.

Step 10: Finally, C_k centers are obtained. Then employ one iteration of $kmedoids$clustering with new C_k centers and $k$are defined as the initial parameters and named all the clusters by k labels.

7. Evaluation And Results

In this section evaluate the clustering performance of the proposed Density Based K-Medoids (DBK) with Expectation-Maximization algorithm and compare the results with DBSCAN clustering andK-Medoids clustering.

A. Performance metrics

In order to measure the performance of clustering by using different types of metricsuch as Rand Index, F-measure,efficiency and Run Time, error rate respectively.Clustering metrics as shown in table 1.

Table 1Clustering metrics

Clustering metrics	
The Number Of cluster classes	50
Crime data Dimensions	2
The Number Of object Per cluster Class	2500,5000,7500,10000, 12500

The Rand index is a measure clustering structure of C_i. The R index is defined as follows

$$Rand\ index = \frac{w + x}{w + x + y + z} \qquad (3)$$

Where w signifies the number of optimal particle in theclusters, z signifies the number of optimal particle in dissimilar cluster, y indicates the number of optimal particlelabelled in the cluster C_j but not in C_i (region $r1$ not in $r2$) and lastly x indicates the number of optimal particlelabelled in the cluster C_i but not in C_j(region $r2$ not in $r1$)

Sum of Squared Error (SSE) is a simplerstandard measure for clustering. Cluster density is measured by utilizing SSE. It is considered as

$$SSE = \sum_{K=1}^{K} \sum_{\forall x_i \in C_k} \|x_i - \mu_k\|^2 \qquad (4)$$

Where $\mu_k$is cluster $k$vector mean value, C_k as a set of cluster $k$occurrences. The μ_k is defined as follows

$$\mu_{k,j} = \frac{1}{N_k} \sum_{\forall x_i \in C_k} x_{i,j} \qquad (5)$$

Where, $N_k = |C_k|$ is the number of occurrencesequivalent to the cluster k.

The F-measure, merging recall and precision values of finally obtained clusters.

$$precision\ (i,j) = \frac{c_{ij}}{c_j}$$

$$recall(i,j) = \frac{c_{ij}}{c_i}$$

The f-measure, $F(i,j)$ of a class i with cluster j then is computed as

$$F(i,j) = \frac{2 * precision\ (i,j) * recall(i,j)}{precision\ (i,j) + recall(i,j)} \qquad (6)$$

The Clustered Support (CS) and Clustered Confidence (CC) is used for make decision about crime.

$$CS = \frac{(clusters_{areas}(antecedent) \Rightarrow clusters_{areas}(consequent))}{area(S)} \qquad (7)$$

$$CC = clusters_{areas}(X \Rightarrow Y) clusters_{areas}(X) \qquad (8)$$

The predefined rules denoted as $X \Rightarrow Y$. Here the clustered crime data are related with the predefined rules to find the crime.

B. Results

Fig. 2 shows the comparison result of threeclustering algorithms such as proposed Density Based K-Medoids (DBK) with Expectation-Maximization algorithm, DBSCAN clustering andK-Medoids clustering in term of the time taken for clustering creation. The proposed DBK+EM algorithm shows the promising result to compare with other two algorithms. The proposed algorithm takes minimum time to create all the clusters.

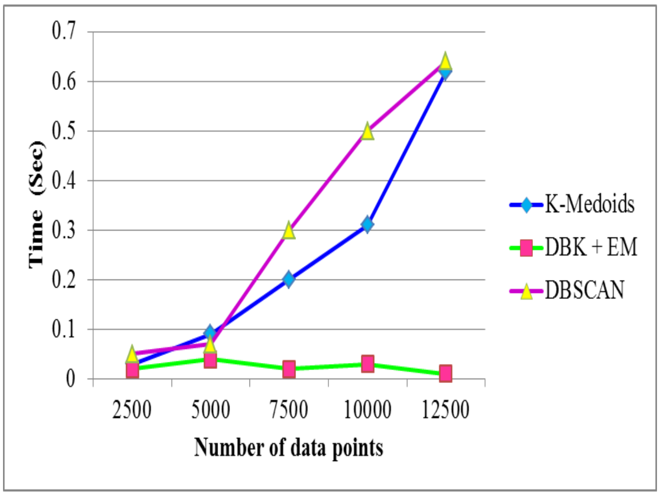

Fig. 2 Time Taken for Clustering

Fig. 3 shows Figure 2 shows the comparison result of three clustering algorithms such as proposed Density Based K-Medoids (DBK) with Expectation-Maximization algorithm, DBSCAN clustering and K-Medoids clustering in term of Rand Index. The proposed system shows the promising results to find the accurate clusters and cluster centroids. Here the proposed approach classification accuracy is greater than the other cluster algorithm.

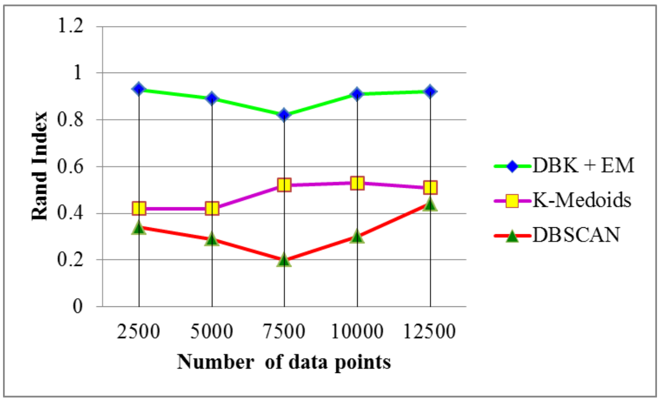

Fig. 3 Rand Index

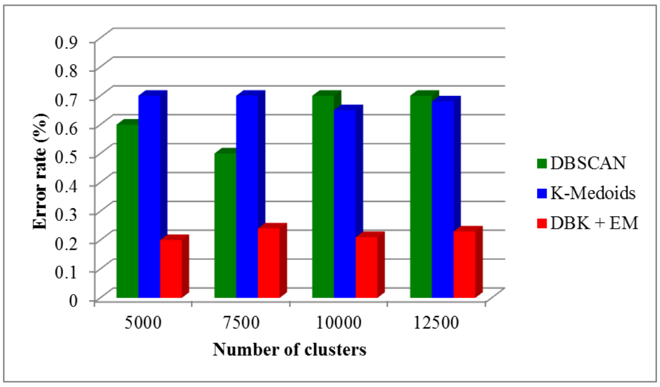

Fig. 4 Error Rate

Fig. 4 shows that cluster generation and classification results in term of Error rate. The proposed system presets the promising results to compare with other two algorithms, which create clusters with minimum error, the error rate has not exceeded 0.2 %. In the proposed system “misclassified” rate is very low.

8. Conclusion

In this proposed Density Based K-Medoids (DBK) with Expectation-Maximization algorithm clustering algorithm efficiently reaching information from raw data (crime Mobile Originated (MO) and Mobile Terminated (MT) data) of the Kerala Government crime dataset. Although, the proposed work easy to implement and this work efficiently avoids the drawback of the huge data dimensionality problem and robustly eliminate the outliers. The experimental results show the DBK + EM algorithm performed very well than DBSCAN clustering and K-Medoids clustering in terms of Rand Index, F-measure,efficiency and Run Time, error rate respectively. One of the major challenges in the crime data analysis is an obtain the understandable knowledge from crime data. It will be avoided by using Cross-Entropy (CE) measures. In addition the EM algorithm help to reveal Density Based K-Medoids (DBK) clustering. In this method helps to crime and police department very well.

References

[1] Malathi. A, S. Santhosh Baboo, Anbarasi. A, "An intelligent Analysis of a City Crime Data Using Data Mining",International Conference on Information and Electronics Engineering, Vol.6, PP.130-134, 2011.

[2] R. G. Uthra "Data Mining Techniques To Analyze Crime Data", International Journal For Technological Research In Engineering,Vol.1, Issue 9, 2014.

[3] Revatthy Krishnamurthy,J. Satheesh Kumar, "Survey Of Data Mining Techniques On Crime Data Analysis",International Journal of Data Mining Techniques and Applications, Vol 01, Issue 02, 2012.

[4] Zakaria Suliman Zubi,Ayman Altaher Mahmmud, "Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data",WSEAS International Conference on Remote Sensing (REMOTE), Research Gate Publication, PP.79-85, 2013.

[5] Uttam Mande,Y.Srinivas, J.V.R.Murthy, "An Intelligent Analysis Of Crime Data Using Data Mining & Auto Correlation Models",International Journal of Engineering Research andApplications (IJERA), Vol. 2, Issue 4, pp.149-153, 2012.

[6] Sathyadevan.S, Amritapuri, Devan, M.S., Surya Gangadharan, S. "Crime analysis and prediction using data mining",First International Conference onNetworks & Soft Computing (ICNSC), IEEE, PP. 406 – 412, 2014.

[7] Malathi. A, S. Santhosh Baboo, "An Intelligent Analysis of Crime Data for Law Enforcement Using Data Mining",International Journal of Data Engineering (IJDE), Vol.1, Issue 5, PP.77-83, 2010.

[8] Ahamed Shafeeq B M, Binu. V. S, "Spatial Patterns of Crimes in India using Data Mining Techniques",International Journal of Engineering and Innovative Technology (IJEIT),Vol.3, Issue 11, 2014.

[9] Deepak Meena, Hitesh Gupta, "Digital Crime Investigation using Various Logs and Fuzzy Rules: A Review",International Journal of Advanced Research in Computer and Communication Engineering,Vol. 2, Issue 4, 2013.

[10] Kadhim B. Swadi Aljanabi, "An Improved Algorithm for Data Preprocessing in Mining Crime Data Set ",Journal of Kufa for Mathematics and Computer,Vol.1, No.4, pp.81- 87, 2011.

[11] Raghuvira Pratap A,K Suvarna Vani,J Rama Devi, K Nageswara Rao, "An Efficient Density based Improved K- Medoids Clustering algorithm", International Journal of Advanced Computer Science and Applications(IJACSA),Vol. 2, No. 6, 2011

[12] Z. Volkovich, R. Avros, M. Golani, "On Initialization of The ExpectationmaximizationClustering Algorithm", Global Journal of Technology and Optimization, Vol.2, PP.117-120, 2011.

[13] Gue Jun Jung, Hoon Young Cho, Yung-Hwan Oh, "Data-Driven Subvector Clustering Using The Cross-Entropy Method", ICASSP, IEEE, PP. 977-980, 2007.

[14] S.Yamuna, N.SudhaBhuvaneswari, "Datamining Techniques to Analyze and Predict Crimes", The International Journal of Engineering And Science (IJES),Vol.1, Issue 2, PP.243-247, 2012.