A COMPARATIVE ANALYSIS OF SELECTED CLUSTERING ALGORITHMS FOR CRIMINAL PROFILING

J. A. Adeyiga[1],*, S. O. Olabiyisi[2] and E. O. Omidiora[3]

[1,] DEPARTMENT OF COMPUTER SCIENCE, BELLS UNIVERSITY OF TECHNOLOGY OTA, OGUN STATE, NIGERIA.
[2, 3,] DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, LADOKE AKINTOLA UNIVERSITY OF TECHNOLOGY, OGBOMOSHO, OYO STATE, NIGERIA

E-mail addresses: [1] *jayadeyiga@gmail.com,* [2] *soolabiyisi@lautech.edu.ng,*
[3] *eoomidiora@lautech.edu.ng*

ABSTRACT

Several criminal profiling systems have been developed to assist the Law Enforcement Agencies in solving crimes but the techniques employed in most of the systems lack the ability to cluster criminal based on their behavioral characteristics. This paper reviewed different clustering techniques used in criminal profiling and then selects one fuzzy clustering algorithm (Expectation Maximization) and two hard clustering algorithm (K-means and Hierarchical). The selected algorithms were then developed and tested on real life data to produce "profiles" of criminal activity and behavior of criminals. The algorithms were implemented using WEKA software package. The performance of the algorithms was evaluated using cluster accuracy and time complexity. The results show that Expectation Maximization algorithm gave a 90.5% clusters accuracy in 8.5s, while K-Means had 62.6% in 0.09s and Hierarchical with 51.9% in 0.11s. In conclusion, soft clustering algorithm performs better than hard clustering algorithm in analyzing criminal data.

Keywords: Clustering Algorithm, Profiling, Crime, Membership value

1. INTRODUCTION

Security of citizens is the major concern of the police. Rather than focusing on enforcement, police can deter crime through the knowledge benefits they derive from information and its associated technologies [1]. As the rate of crime increases each year there is need to develop a computerized crime analysis tool that would assist the Nigeria Police Force and other Law Enforcement Agencies to better combat crime. However, challenges to Law Enforcement and Intelligence Agencies is the difficulty in analyzing large volume of data involved in criminal activities, as the crime rate is increasing [2].

Therefore, detecting patterns in crimes is the next step in predicting and subsequently responding to crime. As such, it is very important to attempt to detect patterns in crime [3]. In detecting patterns in crime, it is very important to gather data from different data sources, store and maintain the data,

generate information and generate knowledge. Due to the vast use of computers and electronic devices and tremendous growth in computing power and storage capacity, there is an explosive growth in data size. The storing of the data in data warehouse enables entire enterprise to access a reliable current database. To analyse this vast amount of data and drawing fruitful conclusions and inferences, it needs special tool called data mining. And clustering algorithm is one aspect of data mining to be considered in this paper work.

Consequently, the inability to analyze criminal accurately based on their behavioral characteristics in situation where hard clustering technique (Simple K-Means) is been used. Therefore, this paper investigated the performance of fuzzy clustering algorithm and hard clustering algorithms by selecting two hard clustering algorithms (Simple K-Means and Hierarchical) and one fuzzy clustering algorithm (Expectation Maximization). Fuzzy

* Corresponding author, tel: +234-703-345-3778

clustering on the other hand accepts the fact that the clusters or classes in the data are usually not completely well separated and thus assigns a membership degree between 0 and 1 for each cluster to every datum which is obviously a good technique that can be used for clustering crime data. Although the extension from deterministic (hard) to fuzzy clustering seems to be an obvious concept, it turns out that to actually obtain membership degrees between zero and one, it is necessary to introduce a so called fuzzifier in fuzzy clustering. The main purpose of the fuzzifier is to control how much clusters are allowed to overlap.

Based on the issues identified above, this paper work is to establish the strength of fuzzy clustering algorithm over hard clustering algorithm by performing comparative analysis of the three selected clustering algorithm to justify the fact that a criminal can belong to as many groups as possible which is the major strength of fuzzy clustering.

2. RELATED WORK

There have been many research works in data mining especially in the area of security that has to do with criminal data analysis and profiling. For example, [4] proposed a method to employ computer log files as history data to search some relationships by using the frequency occurrence of incidents. They analyzed the result to produce profiles, which can be used to perceive the behavior of criminal. Because criminal profiling helps in identifying crime characteristics, which is the first step in developing crime analysis. [5] Also introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly.

While [5] and [6] both used data mining approaches in criminal career analysis. Four important factors play a role in the analysis of criminal career (crime nature, frequency, duration and severity). The tool that was developed in both works extracts these important characteristics of a criminal from the database and creates a digital profile for all offenders, compared all these individuals by a new distance measure and clusters them accordingly. But in the work of [6] there was significant improvement over the work of [5]by making four major key enhancements mainly to improve the semantics and the efficiency, thereby improving the existing methods of automated

criminal career analysis. A new distance measure was introduced that more closely resembles the reality of policing. Instead of the previous, more rigid, comparison of career changes over time in the work of [5] they proposed the collection of crimes in a single year as a multiset, which then describes severity, nature and frequency inherently. A new distance measure called jaccard distance was employed for calculating the difference between two crime-multisets. Instead of the former method of a strict number wise comparison between years (comparing the first year of criminal A with the first year of criminal B, the second year of A with the second year of B, etc.), with the possibility of stretching or shrinking careers, they proposed a novel alignment of the mentioned multisets.

In conclusion, the enormous cloud of one-time offenders gave an unclear sketch of their distance space and the runtime of the chosen approach was not optimal yet as the clustering method (push and pull clustering) used in the former one was too intensive in a computational way causing performance delays, as this is highly inefficient in the real world where the volumes of data increase daily. This was addressed with the used of torus clustering. This was able to solve the problem of time complexity and some major problem encountered in the former method. Finally, it was suggested by [6] to equip the tool with a sub-cluster detection algorithm to provide better insights into the comparability of criminal careers. And also to set fuzzy borders between the different years' crimes within months ending or beginning such a time unit can be (partly) assigned to the next or previous year respectively as well, thus eliminating the problems arising with strict coherence to the change of calendar year. It was also suggested that other clustering techniques be adopted, if they adhere to the demand of incremental addition of single items. According to [7], which profile and analyse a criminal with the presence of evidence from the crime scene as oppose to the works of [5] and [6] who did their own analysis without evidence from the crime scene by constructing a Bayesian networks for criminal profiling from limited data with the presence of evidence from the crime scene. The method adopted in their work was that they developed a Bayesian network (BN) model of offender behaviour from a database of cleared homicides.

The BN was able to infer the characteristics of an unknown offender from the crime scene

evidence, and help narrow the list of suspects in an unsolved homicide. Their research shows that 80% of offender characteristics are predicted correctly on average in new single-victim homicides, and when confidence levels are taken into account this accuracy increases to 95.6%. Their model is shown in Figure 1.

The knowledge that is gained from data mining approaches is a very useful tool which can help and support police forces [8].These made [9] studied the application of fuzzy association rule mining for community crime pattern discovery. They were able to develop an approach that relieves the need of law enforcement personnel to go through uninteresting, obvious rules in order to find interesting and meaningful crime patterns of importance to their community. Rules discovered in their study offers utility for use from the national level down to the state and community level.

Also, [10] was able to use decision tree and Simple K-Means algorithm to cluster crimes according to their attributes using an online police data in Iraq. The model of their work is shown in Figure 2. While [11] used a clustering/classify based model to anticipate crime trends which could be used to lessen and even prevent crime for the coming years. The first task they did was the prediction of the size of the population of a city and the basic approach used to do this was to cluster population sizes, create classes from the clusters, and then classify records with unknown population sizes, the reason why they used clustering to create classes was because classes from clusters are more likely to

represent the actual population size of the cities. WEKA mining software package was used and adopted for the clustering. The next task was the prediction of future crime trends by tracking crime rate changes from one year to the next and using data mining to project those changes into the future. The basic method they used was to cluster the cities having the same crime trend, and then using "next year" cluster information to classify records.

They concluded that from the encouraging results, it is believed that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis, and that there was still lots of options to explore. Visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern. While [12] used various data mining techniques to combat crime and terrorism in Nigeria, they analyzed how data mining techniques can be adopted by law enforcement agencies in tracking the activities of terrorist and their criminal activities, and also examined the limitation of data mining in fighting crime in Nigeria.

Malathi and Santhosh [11] looked at the use of missing value and clustering algorithm for crime data using data mining. They also looked at MV algorithm and Apriori algorithm with some enhancements to aid in the process of filling the missing value and identification of crime patterns. They also used semi-supervised learning technique for knowledge discovery from the crime records and to help increase the predictive accuracy.
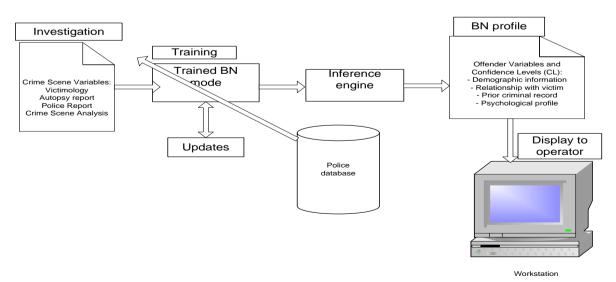

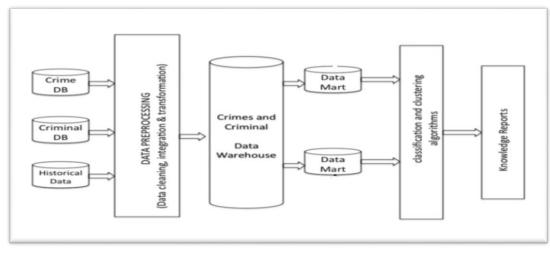*Figure 1   A BN model as decision-support tool for police investigation.*
Source [7]

Figure 2: Framework Architecture for Crime Analysis, Source: [10]

Two clustering techniques, K-Means and DBScan (Density-Based Spatial Clustering Application with Noise) were used to test the data. Thus the two clustering techniques were analyzed in their efficiency in forming accurate clusters, speed of creating clusters, efficiency in identifying crime trend, identifying crime zones, and crime density of a state and efficiency of a state in controlling crime rate. Experimental results showed that HYB algorithm show improved results when compared with k-means algorithm and therefore was used in further investigations.

In conclusion the research focused on developing a crime analysis tool for Indian scenario using different data mining techniques that can help law enforcement department to efficiently handle crime investigation. The tool was proven to be effective in terms of analysis speed. While [13] did an intelligent analysis of crime data using data mining methods (clustering and classification) and auto correlation models which aim to identify a criminal based on the witness/clue at the crime spot. The clustering of data was based on criminal/ crime and thereby minimizing the search space, based on the clusters the classification algorithm is then applied to classify the criminal then the auto correlation model authenticates the criminal.

3. THE SELECTED CLUSTERING ALGORITHMS

Clustering is the technique that is used to group objects (crimes and criminals) without having predefined specifications for their attributes [10]. It ensures that the data objects are similar to one another within the same cluster and dissimilar to other objects in a different cluster. There is no golden standard or benchmark for the type of clustering technique to use, it depends on the goal of the researcher [14]. The clustering methods is divided into

Hard and Fuzzy clustering. The clustering algorithm selected for comparison in these research work includes Simple K-Means, Hierarchical and Expectation Maximization Algorithm. These are combination of hard and soft clustering.

3.1 Simple K-Means

These methods of clustering are based on the classical set theory, and require that an object either does or does not belong to a cluster [9]. It means partitioning the data into specified number of mutually exclusive subsets. That is, each document belongs to exactly one cluster, it only assign a value of 1 and 0. Construct various partitions and then evaluate them by some criterion. Partition techniques create a flat partitioning of the data points. If K is the desired number of clusters, then partitioning approaches typically find all K clusters at once. Given k, the k-means algorithm is implemented in four steps:

1. Partition objects into k nonempty subsets
2. Compute seed points as the centroids of the clusters of the current partition (the centroid is the center, i.e., *mean point*, of the cluster)
3. Assign each object to the cluster with the nearest seed point
4. Go back to Step 2, stop when no more new assignment. [15]

The weakness of k-means is that it is applicable only when mean is defined, then what about categorical data?, the need to specify k, the number of clusters, in advance, Not suitable to discover clusters with non-convex shapes, the k-means algorithm is sensitive to outliers. A major example is K-Means clustering which aims to partition a dataset into k clusters in which each observation belongs to the cluster with the nearest mean.

3.2 Hierarchical Algorithm

Hierarchical clustering builds a cluster hierarchy (a tree of clusters). Create a hierarchical decomposition of the set of data (or objects) based on a criteria. Use distance matrix as clustering criteria. This method does not require the number of clusters *k* as an input, but needs a termination condition.Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. As such, these algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will be form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances [15]. Types include Agglomerative (bottom-up) which simply means starting with one point cluster and recursively merging two or more most similar cluster to one parent cluster until the termination and Divisive (Top-down) which starts with one cluster of all objects and recursively splitting each cluster until the termination criterion is reached.

3.3 Expectation Maximization

This belongs to the family of Fuzzy clustering also called soft clustering. This allows objects to belong to several clusters simultaneously, with different degrees of membership. It is more natural than hard clustering because objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership [16]. For example, the clusters or groups that are identified will be overlapping, meaning that one instance may fall into several clusters.

While Expectation Maximization algorithm extends the basic approach to clustering in two important ways. Instead of assigning cases or observations to clusters to maximize the differences in means for continuous variables, the EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the final clusters [17]. It's a framework to approach maximum

likelihood of parameters in statistical models by using an iterative refinement technique

The EM algorithm
1. Select an initial set of model parameters.
2. Repeat
3. E-Step: for each object, calculate the probability that each object belongs to each distribution, i.e., calculate $P(\Theta_j \backslash o_i, \Theta)$.
4. M-Step: given the probabilities from E-step, find the new estimates of the parameters that maximize the expected likelihood.
5. Until the parameters do not change.
6. Or, stop if the change in the parameters is below a specified threshold. [18].

4. THE DATA SET

In this paper work, a comparative analysis of the selected clustering algorithm was performed using WEKA software package. WEKA Version 3-6-12 software was used for clustering the three selected clustering Algorithms, two hard clustering algorithms and one fuzzy clustering algorithm (Simple K-Means, Hierarchical and Expectation Maximization Algorithms) these were also chosen for testing our data due to their popularity, wide acceptability [19]. The performance was determined by calculating the percentage of the correctly clustered instances and also comparing the actual instances against the clustered instances.

The data set was gathered from the law enforcement agencies in Nigeria, the data gotten from the Nigeria police force headquarters in Abuja. Two hundred and sixty two (262) crime records was used in this work with the following attributes, Nature of offence, offence committed, severity level, and mode of operation. This involves data cleaning/entity extraction. Due to the increased number of crimes in recent times large amount of crime and criminal data have to be stored and analyzed before any Mining technique can be effectively utilized because the resulting knowledge depends greatly on the quality of the training data than the mining technique used. Therefore, preprocessing will be applied to filter the data so as to fill in the missing values, identify outliers and smooth out noisy data, correct inconsistent data and lastly resolve redundancy caused by data integration. This is a store of the data gathered from different sources and transformed after the preprocessing. The data that have been pre-processed i.e. cleaned, transformed and integrated is moved to the data warehouse. Its focus is on the modeling and

analysis of the data for decision making. Different factors such as the crime nature, severity, and weapon used and the frequency of the crime will be extracted from the data warehouse so as to compute the criminal profile per offender.

5. RESULT AND DISCUSSION
5.1 Experimental Setup
In order to improve on the existing crime analysis techniques, this study carried out a comparative analysis of different clustering algorithms on the data set gathered from law enforcement agencies in Nigeria. This was implemented using WEKA Version 3-6-12 on Windows 7 64-bit operating system with Intel Pentium processor. Two hundred and sixty two (262) crime records were used in this work with the following attributes, Nature of offence, offence committed, weapon used, severity level, and mode of operation. And classes to cluster evaluation was chosen as the cluster mode with attribute offence committed chosen as the class to cluster.

5.2 Comparison of Selected Clustering Algorithms
Three clustering algorithms (Simple K-Means, Hierarchical and Expectation Maximization Algorithms) were selected for testing our data due to their popularity and wide acceptability (Sharmila and Mishra, 2013). WEKA is also chosen for this research work because it provides easy-to-use interface with lots of algorithms already implemented.

5.2.1 Expectation Maximization Algorithm Result
The following set parameters were used, maximum iterations was set to 1000, seed no was set within the range of 5 to 100 and number of clusters to -1, this select no of clusters automatically by cross validation. The seed no of 10 was the best with 90.5% of clustered accuracy in a time of 8.5s as compared with others. This was able to generate 3 clusters which were also expected, compared with others with 4 and 5 clusters. The clustered instances were much closed to the actual instances with simple offences having exactly the same number clusters. And it does not classify any instances as no class.

5.2.2 Simple k-Means Clustering Algorithm Result
The following set parameters were used; the distance function was set to Euclidean distance, maximum iteration to 1000, Number of clusters to 4 while the seed value was also set within the range of 5 to 100. The seed no of 50 performed optimally with 62.6% of clustered accuracy in a time of 0.09s as compared with others and with just 3 no of iterations to converge that is, it simply measure how similar the points in the cluster are to one another and reflects how accurate the final centers are. It can also be deduced from the result that the same result was generated when the distance function was changed.
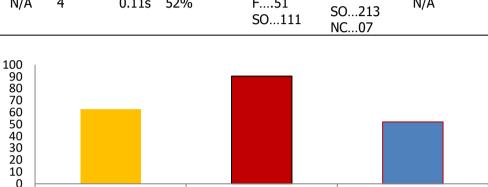
5.2.3 Hierarchical Clustering Algorithm Result
The following set parameters were used, the distance function was set to Euclidean distance and Manhattan distance, link type to Single, Complete, Average, Mean and Centroid with number of clusters also set to 4. Because it expected that the number of clusters must be specified from the beginning. It was also discovered that there was no change in the result regardless of the distance function used. The best result gotten from the hierarchical clustering Algorithm was when the link type was set to Average with 52% of clustered accuracy in a time of 0.11s. It can also be seen that the clustered instances against the actual instances generated a lot of disparity even with the experiment that performed optimal.

5.3 Comparison Analysis of the Three Selected Algorithms Result
The best result was selected from the experiment carried out on each of the Clustering Algorithms (Simple k-means, hierarchical and Expectation Maximization Algorithms) that was chosen for testing our data. Table 5.1 is shown the comparison analysis result. Expectation Maximization Algorithm performed best with 90.5% of clustered accuracy in a time of 8.5s, as compared to the 62.6% in a time of 0.09s and 52% in a time of 0.11s of the K-Means and Hierarchical Clustering Algorithm respectively. This can be proven by plotting the graph of the selected algorithms against the percentage of clustered accuracy and this is shown in Figure 5.1. Also, the result of the actual instances against the clustered instances is very close in Expectation Maximization Algorithm as compared to the Simple k-Means and Hierarchical Clustering Algorithm respectively. This can also be proven by plotting the graph of the selected algorithm against the No of clustered instances and it is shown in Figure 5.2.

Table 1 Comparison Performance of the Selected Clustering Algorithms

Algorithm	Seed No	No of Cluster	Time	% of Clustered Accuracy	Actual Instances	Clustered Instances	Log Likelihood	Sum of Square Error
Expectation Maximization	10	3	8.5s	90.5%	M....100 F....51 SO...111	M....103 F....111 SO...48	-19.33	N/A
Simple K-Means	50	4	0.09s	62.6%	M....100 F....51 SO...111	M....75 F....38 SO...121 NC...28	N/A	1180.0
Hierarchical	N/A	4	0.11s	52%	M....100 F....51 SO...111	M....01 F....41 SO...213 NC...07	N/A	N/A

Figure 3: Bar Chart of the selected algorithm against the percentage of Clustered Accuracy.
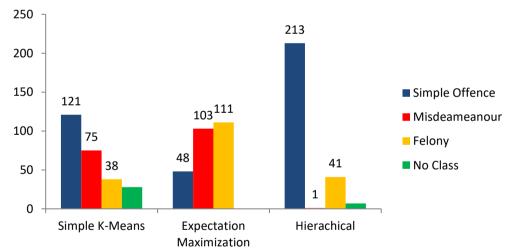

Figure 4: Chart of the selected algorithm against the no of Clustered Instances generated.

6.0 CONCLUSION

This paper established the fact that the most commonly used Simple K-Means algorithm does not give good results for clustering crime data due to the human nature of overlapping. Since it was established that fuzzy clustering algorithm is an improvement over other hard clustering algorithms based on the fact that it gives rooms for overlapping depending on the degree on membership. That is, a criminal can belong to as many clusters has possible due to the human nature which has the tendency of changing behavior. The paper therefore recommend that fuzzy clustering algorithms be further explore for clustering criminal data.

7. REFERENCES

[1] Brown Mary Maureen and Brudney Jeffery. L (2003), "Learning Organization in the Public Sector, A study of Police Agencies employing Information and Technology to Advance

Knowledge" *Public Administration Review* Vol 63, Issue 1, pp. 30- 43

[2] Chen, H. Atabakhsh. H. Petersen, T. Schroeder J. Buetow, T..Chaboya, L. & Huang, Z (2003, may).COPLINK: Visualization for Crime Analysis In proceedings of the 2003 annual national conference on Digital government research, pp, 1-6, Digital Government Society of North America

[3] Askeriya I, H. Jahankhani, S. Wee Lee &Ameer Al-Nemrat, (2010): "Education, Training and Awareness (ETA) Four Dimensional Cybercrime prevention model"

[4]Abraham, T. and de Vel, O. (2006) Investigative profiling with computer forensic log data and association rules, in Proceedings of the IEEE International Conference on Data Mining (ICDM' 02),Pp. 11 – 18.

[5] Jeroen S, De Bruin, Tim K. Cocx, Walter A, Kosters, Jeroen F, J. Laros and Joost N, Kook (2006) "Data Mining Approaches to Criminal Career Analysis," in *Proceedings of the Sixth International Conference on Data Mining* (ICDM' 06), pp, 171 – 177

[6] Tim k. Cocx, Jeroen F. J Laros, Walter A. Kosters (2008) "Enhancing the Automated analysis of criminal careers" [15]Abbas Osama Abu (2008),"Comparisons between data clustering algorithm" *International Arab journal of information technology,* vol 5 No 3 July 2008.

[7] Baumgartner, K, Ferrari S and Palermo G (2008): "Constructing Bayesian Networks for Criminal profiling from limited data" Elsevier 21 March 2008.

[8] Keyvanpour, M.R, Javideh, M. and Ebrahimi, M.R. (2010) Detecting and investigating crime by means of data mining : a general crime matching framework, *Procedia Computer Science,World Conference on Information Technology,* Vol. 3, pp. 872 - 830

[9] Buczak, L. Anna, and Christopher M, Gifford (2010)"Fuzzy Association Rule Mining for Community Crime Pattern Discovery" Published in Proceeding ISI-KDD' 10 ACM SIGKDD Workshop on Intelligence and Security Informatics, Article No, 2.

[10] Swadi B, Kadhim. Al- Janabi (2011) "A Proposed Framework for Analyzing Crime Data Set Using Decision Tree and Simple K – Means Mining Algorithms" *Journal of Kufa for Mathematics and Computer* Vol. 1. No 3, may 2011, pp 8 – 24.

[11] Malathi and Santhosh, (2011) ''an enhanced algorithm to predict a future crime using data mining" *International Journal of Computer Applications* (0975 - 8887) volume 21 – no. 1, may2011.

[12] Raphael O. O & Francis O.E (2011): "Combating Crime And Terrorism Using Data Mining Techniques", 10th International Conference On *Information Technology for People Centred Development Nigeria Computer Science Conference Proceedings*, Vol 22, pp80 – 89 2011.

[13] MandeUttam, Y. Srinivas, J.V.R.Murthy (2012) An Intelligent analysis of Crime Data using data mining and auto correlation models. *International Journal of Engineering Research and Applications* (IJERA) Vol. 2, Issue 4, July – August 2012. Pp. 149 – 153

[15] Khare Pallavi, Anagha Gaikwad, and Pooja Kumari, (2015). Fuzzy C- Means Clustering withKernelMetric and Local Information for Image Segmentation; *International Journal of Computer Applications* (0975 – 8887) *National Conference on Emerging Trends in Advanced Communication Technologies* (NCETACT-2015).

[16] Esh Narayan, Yogesh Birla and GauravTax(2012) "Enhancement of fuzzy C-means clustering using Expectation Maximization algorithm". *International Journal of Computer Applications* Vol 43-no 13 April 2012.

[14] Krishnamurthy Revatthy (2012) "Survey of Data Mining Techniques on Crime Data", *International Journal of Data Mining Techniques and Applications* vol. 01, pp. 48 – 55

[17] Naser Sara, RawnAlkhaldi and Gregory Vert (2005) "A modified fuzzy k-means clustering using Expectation maximization algorithm.

[18] Han, Kamber and Pei,(2012) "Knowledge discovery and data mining"

[19] Zadeh L.A,(1975)"Fuzzy logic and approximate reasoning, synthese ,vol, 30 pg 407-428.