International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 12– No.34, November 2020 – www.ijais.org

Crime Analysis Tool using Kernelized Fuzzy C-Means (KFCM) Algorithm

Adeyiga J.A.
Bells University of Technology
Ota, Ogun State, Nigeria

Achas M.J.
Bells University of Technology,
Ota, Ogun State, Nigeria

Adewumi O.A.
Bells University of Technology,
Ota, Ogun State, Nigeria

ABSTRACT
Several criminal analysis tools have been developed to assist the Law enforcement agency LEA in solving crimes but the techniques employed in most of the systems lack the ability to analysis criminal based on their behavioral characteristics. Hence, this research therefore developed a criminal analysis tool using the KFCM algorithm and compared the result with the FCM algorithm.

The data used was downloaded online and it is available at https://portal.chicagopolice.org/portal/page/portal/ClearPath/News/Crime%20 statistics from the city of Chicago Police Department with over one million records.

The paper reviewed the Fuzzy C-Means (FCM) clustering algorithm and the Kernelized Fuzzy C-Means algorithm and then implemented and compared the results of both algorithms using confusion matrix as the metric of evaluation.

The result analysis shows that the KFCM and the FCM algorithms both performed at par to each other but the KFCM had a better accuracy over the FCM algorithm with a higher execution time.

The FCM algorithm is therefore recommended to be modified along with the KFCM to give a more robust cluster with higher performance.

Keywords
Kernelized fuzzy c-means, law enforcement agency, clustering, clustering algorithm, Analysis, Crime data, Euclidean distance

1. INTRODUCTION
To detect patterns in crimes are next to predicting and then responding to crime in order to assist the Law Enforcement Agencies. As such, it is very important to attempt to detect patterns in crime [1]. In detecting patterns in crime, it is very pertinent to gather data from different data sources, store and maintain the data, generate information and generate knowledge.

According to [10], about 10% of the criminals commit about 50% of the crimes. Therefore, it is imperative to profile criminals in other to detect patterns from criminal records so as to discover clues that will help Law Enforcement Agencies in the investigation of crime, because an ideal Crime Analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detections and action [10].

The purpose of data grouping or clustering is simple in its nature and is close to the human way of thinking. Therefore, data clustering is a process of putting similar data into groups. Clustering techniques and algorithm are based on real life model that data with certain qualities must cluster together. The goal of clustering is to group similar objects into one cluster and dissimilar objects into another cluster based on characteristics of data [8]. Based on the characteristics of data, certain occurrences of data can be further placed under detailed surveillance.

The step from the well-known Simple K-Means clustering algorithm to Fuzzy C-Means algorithm and its vast number of sophisticated extensions and generalizations involves an additional clustering parameter, the so called fuzzifier [2].

Fuzzy clustering 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 [5]. 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 [5]. The main purpose of the Fuzzifier is to control how much clusters are allowed to overlap.

According to [4], detecting crime from data analysis can be difficult because the daily activities of criminal generate large amounts of data and stem from various formats. Therefore, Crime Pattern Analysis is paramount to effectively support action planning and decision making to manage crime.

Consequently, Several crime analysis system have been developed to assist the Law Enforcement Agencies in solving crimes but the techniques employed in most of the analysis system lacks essential components which would have made the quality of their output more in tune with reality. Such as, the inability to cluster criminal accurately based on their behavioral characteristics in situation where hard clustering technique (Simple K-Means) is been used. Also, the Euclidean distance function used in the FCM Algorithm to measure the distance between the data point and cluster center. The limitation in using the Euclidean distance is that it measures only noise free data and Euclidean Shaped data set [9]. Therefore the Kernelized Fuzzy C-Means (KFCM) is being proposed to overcome the shortcomings of both the Simple K-means and the FCM algorithm. This paper therefore developed a criminal analysis tool using the KFCM algorithm and compared the result with the FCM algorithm.

2. LITERATURE REVIEW
The earliest and simplest known clustering technique, which is the simple K-Means, has the limitation of not always guarantee convergence and the methods of clustering are based on the classical set theory, that require an object to either belong to or not belong to a cluster [3]. This is too rigid to cluster crime data, in order to show the behavioral characteristics of human. Based on the limitations identified

with the simple K-Means, this paper, further reviewed the Fuzzy C-Means (FCM) clustering algorithm that has the privilege of overlapping, in which a data item can belong to two or more clusters based on the degree of membership, These methods of clustering could better describe our crime data set based on the behavioral tendency of human. But the challenge with FCM algorithm according to [9] is with the Euclidean distance function used to measure the distance between the data point and cluster center. The limitation in using the Euclidean distance is that it measures only noise free data and Euclidean Shaped data set. Based on the issue identified with the FCM, an improvement was made on the FCM with the introduction of a kernel function to properly measure data set of various forms. In view of the shortcomings identified above, this paper implemented and compared the results of both FCM and KFCM using confusion matrix as the metric of evaluation by calculating the accuracy, precision sensitivity, specificity and the time.

3. THE PROPOSED SYSTEM

3.1 Data Acquisition

The data used was downloaded online and it is available at https://portal.chicagopolice.org/portal/page/portal/ClearPath/News/Crime%20 statistics from the city of Chicago Police Department with over one million records with the following attributes; Crime nature, crime description/mode of operation, location of crime, scene description, date and time, number of crimes committed and so on. And it was used to test the performance of the criminal analysis system

The criminal data downloaded online were preprocessed through the use of concept hierarchies where the raw data were replaced by higher level concepts. The increase in the number of crimes in recent times which has led to large number of criminal data being stored and analyzed before the clustering technique was effectively applied because the resulting knowledge depends greatly on the quality of the training data than the clustering technique used. Based on these, some inconsistencies in the data were addressed by filing some missing values before selecting the necessary attribute needed for profiling. This was further transformed into a research questionnaire so as to seek the service of professionals/experts in the area of crime investigation in Nigeria to give pertinent information as regards the classification and description of the data.

The data gotten after preprocessing was normalized using the feature scaling concept, where all the attributes have being converted into a numeric values and the range of the feature of the data are reduced to a scale between 0 and 1. This is necessary because the variables of the different attributes in the data set gathered have disparate ranges and it is expected that all data lie in the same range. This was achieved using equation 3.1 that computes z, the normalized value of a member of the set of observed values of x.

$$Z = (x - \min[(x)]/(\max(x) - \min \quad (x)) \qquad (1)$$

Where min and max are the minimum and maximum values in x given its range.

The normalization attributes such as crime nature, severity, weapon used, frequency of the crime and the computed criminal profile per offender was then clustered together using FCM and the KFCM algorithm for knowledge creation from the results generated.

3.2 Description of the System Architecture for the Criminal Analysis System

Crime data was gathered online from the city of Chicago Police Department and both crime nature and crime description attributes were both transformed into a research questionnaire for factor analysis with the involvements of police experts to give pertinent information as regards the classification of the crime nature attribute and weight on the crime description attribute. A data warehouse was then generated after analysis of the questionnaire, from the data warehouse, attributes needed for profiling the data such as the crime nature, crime location (longitude and latitude), mode of operation, number of crime committed, criminal identification number and the severity of the crime computed were selected. Based on the attributes, a profile was created for individual criminal, and clustering algorithm was then applied on the profile created in order to generate different clusters and the knowledge acquired from the clusters formed was useful to the Law Enforcement Agencies. The clusters formed were able to categorize the criminals into groups based on their behavioral tendencies, such as severity of the offences of the different criminals. The proposed system architecture is shown in Figure 1.

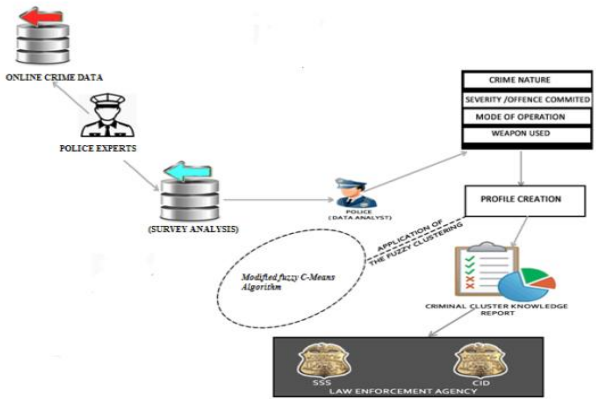

Fig 1: System Architecture for Criminal Profiling

4. THE ALGORITHMS

4.1 The Fuzzy C-Means Algorithm

Step 1. Compute the membership matrix (U) and initialize randomly in Equation (2).

$$\sum_{i=1}^{c} U_{ij} = \quad 1 \forall \quad j = 1.....n \qquad (2)$$

This equation represents the membership matrix; the summation of the membership values must be equal to 1.

Step 2. Calculate the centroids (Ci) in Eq. (3).

$$C_i \quad = \quad \frac{\sum_{j=1}^{n} u_{ij}^{m} x_{ij}}{\sum_{j=1}^{n} u_{ij}^{m}} \qquad (3)$$

Centroid is main point of the cluster analysis system, in clustering this value of C_i depends on the membership matrix function and related parameter of Xi.

Step 3.Calculate the dissimilarities between centroid and data points in Equation (4). This check the threshold value using membership matrix and Euclidian distance between ith centroid (C_i) and jth data point

$$J(U, c_1, c_2, \ldots, c_c) = \sum_{i=1}^{c} j_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} \quad (4)$$

$$\text{Where } d_{ij}^{2} = \|X_j - C_i\|^2 = (X_j - C_i)^T \quad (5)$$

is a squared inner-product distance norm, and

$$m \in [1, \infty) \quad 6)$$

is a parameter that determines the fuzziness of the resulting clusters and is set to 2.

$$\text{If } \| U(k+1) - U(k) \| < 0.01 \quad (7)$$

Eq. (7) checks the difference between the value of the present classes and the next classes of the membership function and compares it with the threshold value. If values are satisfied, then we forwarded the next steps.

Else go back to step two until the values are satisfied.

The weakness and strength of Fuzzy C-Means Algorithm are highlighted below.

Strength

1. FCM gives best result for overlapped data set and is comparatively better then k-means algorithm.

2. Data point is assigned membership to each cluster center as a result of which data point may belong to more than one cluster center

Weaknesses

1. Apriori measurement of the number of clusters.

2. With subordinate value of β we get the better result but at the overhead of extra number of iteration.

3. Euclidean distance measures can inequitably weight underlying factors. [6]

4.2 The Kernelized Fuzzy C-Means Algorithm

Step 1. Compute the membership matrix (U) and initialise randomly in Eq. (8)

$$\sum_{i=1}^{c} U_{ij} = 1 \quad \forall = 1 \ldots\ldots\ldots n \quad (8)$$

This equation represents the membership matrix; the summation of the membership values must be equal to 1.

Step2. Calculate the centroids (Ci) in Eq. (9)

$$C_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_{ij}}{\sum_{j=1}^{n} u_{ij}^{m}} \quad (9)$$

Centroid is main point of the cluster analysis system, in clustering this value of C_i depends on the membership matrix function and related parameter of Xi.

$$J(U, c_1, c_2, \ldots, c_c) = \sum_{i=1}^{c} j_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2} \quad (10)$$

Step 3.Calculate the dissimilarities between centroid and data points in Equation (10). This check the threshold value using membership matrix and Euclidian distance between ith centroid (C_i) and jth data point.

The Euclidean distance used in optimizing the objective function in Eq. (5) is being replaced by kernel induced distance function due to the inconvenience in using the Euclidean distance that measures only noise free data and Euclidean shaped dataset [9].

$$\text{where } d_{ij}^{2} = \| \psi X_j - \psi C_i \|^2 \quad (11)$$

$$\| \psi X_j - \psi C_i \|^2 = \langle \psi(Xj) - \psi(Ci), \psi(Xj), (Ci) \rangle \quad (12)$$

And

$$\langle \psi(Xj), \psi(Ci), \rangle = \mathrm{K}(Xj, Ci) \quad (13)$$

Hence, the kernel induced distance function is

$$\| \psi X_j - \psi C_i \|^2 = K(Xj, Xj) + K(Ci, Ci) = 2K(Xj, Ci) \quad (14)$$

$$m \in [1, \infty) \quad (15)$$

is a parameter that determines the fuzziness of the resulting clusters and is set to 2.

If

$$\| U^{k+1} - U^{(k)} \| < 0.01 \quad (16)$$

Eq. (16) checks the difference between the value of the present classes and the next classes of the membership function and compares it with the threshold value. If values are satisfied, then we forwarded the next steps.

Else go back to step two until the values are satisfied.

A precarious issue related to KFCM clustering is the selection of an "optimal" kernel for the problem at hand. The kernel function in use must conform to the learning objectives in order to obtain meaningful results for un-labeled data [6]. And not flexible enough to support data from different heterogeneous sources [7].

5. RESULT ANALYSIS FROM THE CLUSTERING TECHNIQUES

The results gotten from the implementation of the existing Fuzzy C-Means Clustering Algorithm and the Kernelized Fuzzy C-Means Algorithm are shown in Figure 2 and Figure 3 respectively. The three clusters formed categorized the criminals into groups based on their behavioral characteristics in terms of their degree of membership to the three clusters created, such as light, intermediate and heavy weight criminals. This in essence is useful for Law Enforcement Agency in reducing the search space during investigation especially in the case where there is no evidence at the scene of the crime. They only need to focus their search on the cluster of interest instead of searching through the whole record. Each data item in each of the cluster represent a criminal and as such, a particular data item of interest can further be investigated by looking through its crime profile history such as, its usual mode of operation, nature of previous crime committed and weapon used etc for intelligence and knowledge in order to ascertain whether it fits into the description of the crime under investigation.

From the different clusters generated in Figure 2, its membership value is shown along the x axis that shows the extent to which each data item belongs to a cluster. Based on the cluster formed, criminals that exhibit the same or having similar crime attributes are clustered together along with their level of membership that shows the extent to which such criminal belong to the particular cluster. This could be described as a light, intermediate or heavy criminal severity cluster depending on its placement on the graph. The membership value is just giving the degree of overlapping of each data item to the clusters. That is, to what extent a criminal belongs to the different clusters based on the attribute of that criminal. Based on this logic, the idea is that the entire criminal in a particular cluster tends to exhibits the same behavioral patterns based on the attributes extracted from the data set to form the cluster.

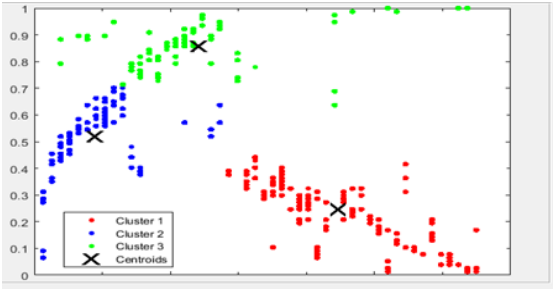

Fig 2: Fuzzy C-Means Cluster

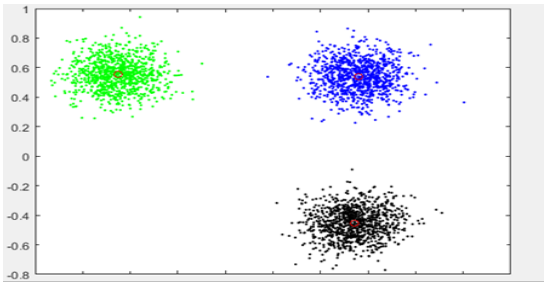

Fig 3: Kernelized Fuzzy C-Means Cluster

6. EVALUATIONS OF RESULTS

To evaluate the performance of the results generated from the clustering algorithms, the performance of the (FCM and KFCM) on the crime data set were compared. Two experiments were carried out, each made of 2150 number of instances, and six numbers of attributes, 1000 instances were used for training while 1150 instances were used for testing. The first experiment was carried out using the existing FCM and the confusion matrix is shown in Table 1. The confusion matrix is pertinent, since the evaluation metric chosen for this research are all based on the confusion matrix, and its performance evaluation results is shown in Table 2 while the second experiments was carried out using the existing KFCM with the confusion matrix shown in Table 3 and its performance evaluation results shown in Table 4. The performances were evaluated in terms of Sensitivity, Accuracy, Specificity and Execution time.

Table 1: Experiment One: Confusion Matrix Using the FCM Algorithm

Cluster No	TP	FN	FP	TN
1	213	35	26	69
2	169	33	29	153
3	148	28	9	100
Average	177	32	21	107

Table 2: Experiment One: Performance Analysis Result Using the FCM Algorithm

Cluster No	SPEC(%)	SENS(%)	PREC(%)	ACC(%)	TIME(S)
1	85.86	86.17	94.22	86.09	0.32
2	84.05	83.33	90.30	83.59	0.32
3	91.74	84.09	94.26	87.01	0.32
Average	87.22	84.53	92.93	85.57	0.32

Table 3: Experiment Two: Confusion Matrix Using the KFCM Algorithm

Cluster No	TP	FN	FP	TN
1	256	16	51	56
2	259	23	8	92
3	168	15	9	93
Average	228	18	23	80

Table 4: Experiment Two Performance Analysis Result Using the KFCM Algorithm

Cluster No	SPEC	SENS(%)	PREC(%)	TIME(s)
1	64.22	94.35	84.17	2.31
2	92.21	92.16	93.89	2.31
3	91.17	91.80	94.91	2.31
Average	82.54	92.77	90.99	2.31

6.1 Comparative Analysis of the Algorithms

The comparisons analysis table shown in Table 5, further revealed the analysis of the results obtained from the clustering algorithms used in the experiments. A total of five metrics were used for the evaluation of the performance of the algorithms. The mean performance of all the metrics used for evaluating the algorithm were computed, compared and analyzed in Table 5. This revealed that the KFCM algorithm performed better than the existing FCM. The KFCM algorithm had 82.5% specificity as against 87.2% for FCM and KFCM; 92.8% sensitivity as against 84.5% for FCM and KFCM; 90.9%; precision as against 92.9% for FCM and KFCM; 89.4% accuracy as against 85.6% for FCM. The

KFCM had a computational time of 2.32s as against 0.32s for FCM. The analysis is also depicted graphically in fig 4; this gives a clearer view of the comparism analysis carried out.

Table 5: Comparative Analysis Table of both algorithms

Algorithm	SPEC(%)	SENS(%)	PREC(%)	ACC(%)	TIME(S)
FCM	87.2	84.5	92.9	85.5	0.32
KFCM	82.5	92.8	90.9	89.4	2.32

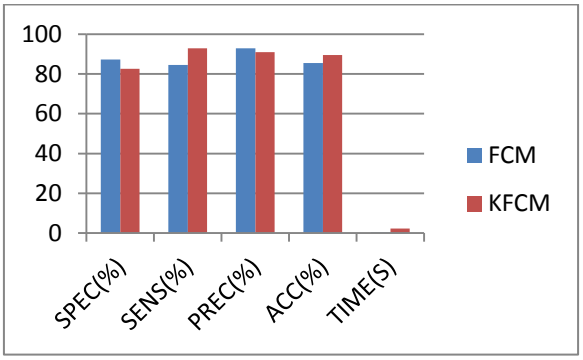

Fig 4: Comparative Analysis Table of both algorithms

7. CONCLUSION

In conclusion, the KFCM algorithm and the FCM algorithms both performed at par to each other but the KFCM had a better accuracy over the FCM algorithm with a higher execution time. The KFCM is better recommended because the execution time can be trade off, since the system is a security system and the accuracy of a security system is paramount. Hence, the high computational time of the KFCM algorithm could be traded off for a better and accurate security system.

8. REFERENCES

[1] Askeriya I, H. Jahankhani, S. Wee Lee and Ameer Al-Nemrat. (2010). Education, Training and Awareness (ETA) Four Dimensional Cybercrime prevention model; 24th European Conference on Information Systems, (12-15) Istanbul, Turkey.

[2] Dellaert Frank (2002). The Expectation Maximization Algorithm; Technical report Georgia Institute of Technology, 2002.

[3] Esh Narayan, Yogesh Birla and Gaurav Tax (2012). Enhancement of Fuzzy C-means Clustering using Expectation Maximization Algorithm; International Journal of Computer Applications, Vol 43 no 13 April 2012.

[4] Krishnamurthy Revatthy (2012). Survey of Data Mining Techniques on Crime Data; International Journal of Data Mining Techniques and Applications, Vol. 01: 48 –55.

[5] Klawonn Hoppner F, F Fruse R and Runkler (2003). Fuzzy Cluster Analysis; J. Wiley and Sons, Chichester, England 2003.

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

[7] Prabhakar Sandhya H, Sandeep Kumar, (2017). A Survey on Fuzzy C-Means Clustering Techniques; International Journal of Engineering Development and Research 2017 IJEDR | Vol 5, Issue 4 | ISSN: 2321-9939.

[8] Raphael O. O and 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: 80 – 89, 2011.

[9] Sheeba M. S. and A. Sathya (2015) "Hybrid approach of Kernelized Fuzzy C-Means and Support Vector Machine for Breast Medical Image Segmentation" Journal of Chemical and Pharmaceutical Research, 2015, 7(2):281-291.

[10] Shyam Varan Nath, (2006). Crime pattern detection using Data Mining; Proceeding of the 2006, International Conference on Web Intelligence and Intelligence Agent, 41-44.

[11] City of Chicago police Department (2001) via city of Chicago.org/public-safety/crimes-2001-to present. available at https://portal.chicagopolice.org/portal/page/portal/ClearPath/News/Crime%20Statistics