Fuzzy C-Mean Clustering Algorithm Modification and Adaptation for Applications

Bassam M. El-Zaghmouri

Department of Computer Information Systems
Jerash University
Amman - Jordan

Marwan A. Abu-Zanona

Department of Computer Science
Imam Muhammad Ibn Saud Islamic
University- Al-Ehsa Branch
Al-Ehsa – Saudi Arabia

Abstract - Many clustering algorithms with different methodologies are subjected to be common techniques and main step in many applications in the computer science world. The need of adapting efficient clustering algorithm increases in critical applications (i.e. wireless sensors networks). Utilizing the Fuzzy Logic power; Fuzzy C-mean (FCM) clustering has a major role in most clustering applications. But in many cases, the result of FCM is considered to be non-complete clustering strategy. This paper adapted the FCM algorithm to enable of generating clusters with equal sizes. Also, scattered points that are located far away from all clusters are grouped out of clusters. Another modification is to localize specific points that have ability to locate in more than one cluster; hence this has a non-negligible importance in some fields such as cellular communications.

Keywords – Fuzzy C-Mean; Clustering; Euclidean Distance

I. INTRODUCTION

Clustering of numeric data is a basic procedure of many classifications and modeling algorithms. The goal of clustering is to create new groups of data from large data set. Where that is very useful and needed in many applications (i.e. data management in space, cellular communications, wireless sensors network ... etc.). Fuzzy C-Means (FCM) clustering algorithm minimizes the cost function. It has been studied and applied (dunn, 1974; Bezdek, 1974; Bezdek et 1987) [1] [2]. FCM is the one of the most recent contributions to the field of AI and data clustering. It has several benefits compared to existing development approaches for data clustering, in particular the ability to divide data for different size clusters with all fuzzy logic benefits. This research is an enhancement of the basic FCM algorithm. This paper developed a new contribution that enable to get a set of equal-size clusters in robust and fast processing abilities [2].

From the fact that the world reaches becomes grow too fast; this paper aims to develop new approach which leads us to solve clustering methodology and FCM algorithm in order to achieve goals by improving that algorithm to divide data to achieve accurate data clustering.

Fuzzy C-Mean (FCM) algorithm is a away to show how data can be classified and clustered in organization or in any application such as cellular, but it's clear to observe that data has some attributes such as distance between points of data, distance, weight and potential value for data points that makes it difficult to understand how to cluster data points such away to achieve better classification and use for data points [1]. FCM algorithm divide data for different size cluster by using fuzzy system depend on many criteria like distances between one data point and another's, choosing center point and membership function that mean we don't have accurate data cluster size. This paper is adding a new development to the FCM algorithm to get equal size clustering method. The modification is an addition to FCM and not internal modification of the algorithm. FCM choosing cluster size and central point depend on fuzzy model but this paper solve this problem by using fuzzy model to define central point of cluster then use Euclidean function and distance value for cluster size.

FCM algorithm work such away all points Out of clusters and put it to cluster that near to this point by measuring distance between point and clusters but in my

new enhancement will have new weight concept to decide if we will include this point to any cluster or cancel it [3].

If the data point included in two clusters or more the problem is how to decide in which clusters it can be used? This problem is so important in some industrial application. The traditional clustering using FCM algorithm locates that point to the first arranged cluster and omits it from the other cluster(s), whereas that cluster may be the wrong one or even that point may subjected to interest for all clusters that shares this point. This paper relates the points to its cluster depending on its potential regardless of its arrange.

II. LATERAL SURVEY

Some different researches have been concerned in Fuzzy Model identification based on cluster estimation present efficient method for estimating cluster center of numeric data and can be use to determine number of clusters and their initial value, and also it related to change the algorithm internally.

In [2], Feature-weight assignment that if all values of feature weights are either 1 or 0, a number in [0 1] can be assigned to a feature for indicating the importance of the feature that an appropriate assignment of feature-weight can improve the performance of fuzzy c-means clustering. "The weight assignment is given by learning according to the gradient descent technique; Experiments on some UCI databases demonstrate the improvement of performance of fuzzy c-means clustering",

Alternative c-means clustering algorithms was presented in [1] explains a new alternative hard c-means (AHCM) and alternative fuzzy c-means (AFCM) clustering algorithms. These alternative types of c-means clustering have more robustness than c-means clustering. Numerical results show that AHCM has better performance than HCM and AFCM is better than FCM.

From that point, this paper introduces a new development of FCM-based algorithm which solves some of the weakness of above methodology. Proposed methodology must have all development life cycle from requirement to testing, easy to use and to learn

III. METHODOLOGY

This algorithm uses the FCM traditional algorithm to locate the centers of clusters for a bulk of data points. The potential of all data points is being calculated with respect to

specified centers. The availability of dividing the data set into large number of clusters will slow the processing time and needs more memory size for the program. Figures-1 shows the data set, where figure-2 displays the center of clusters that gotten from the traditional FCM algorithm. Figure-3 shows the traditional result of FCM for cluster the data set into four clusters.
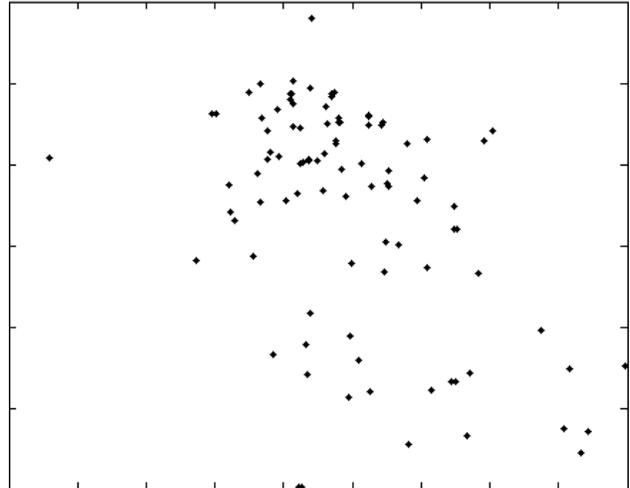

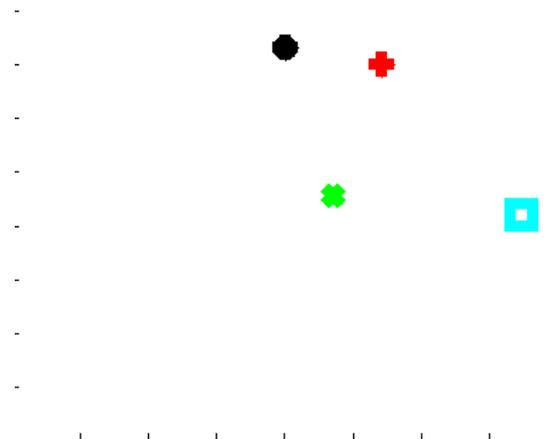
Figure 1. Original testing data set.

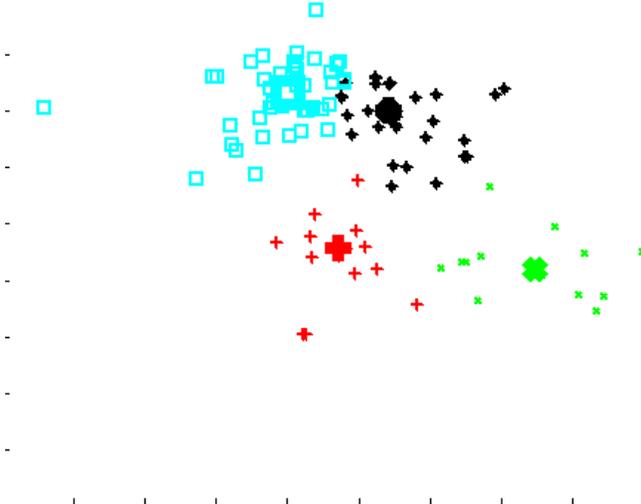Figure 2. Cluster Centers using Traditional FCM.

Figure 3. Clustering using Traditional FCM.

The tested data is real scope data consists of 100 sample point. Hence traditional clustering should device the data to four clusters and each data point should be located in one specified cluster. From figure-3 the results will be as; the first cluster contains 20 elements (Green), the second cluster contains 41 elements (Red), the third cluster contains 18 elements (Black), and the fourth one contains 21 elements (Blue).

When looking to analyze the data clustering, that is obvious; some clusters contains double size of data that included in another clusters, the red cluster contains a point (the most left) that is too far from the center of that cluster whereas the most points of the green and blue clusters is more closest to the center of red than that far point.

Thus, there are three problems in this algorithm:

1. The size of clusters is not equal, whereas in many applications it is needed to be equals.
2. The far point is being merged to a specified cluster, and this will be a big deal in most applications (i.e. communication applications).
3. There is a points that located out of cluster and it is closest than another points in that cluster to its center (points located between two clusters).

This paper is proposing a solution to generate equi-sizes clusters and solve the problem of points that located between two clusters.

Point Potential with respect to Cluster Center

First step in clustering is locating centers for assumed clusters. This is done by fuzzy method depending on FCM traditional algorithm. The centers will be located as shown in figure-2. The next step is to grouping each data point to a specified cluster. The potential of the point with respect to each center is calculated and compared, and the center that has maximum potential with respect the data point will be its cluster center.

The Euclidean distance between two points is defined as the length of the line segment connecting them. In Cartesian coordinate system, the two dimensional representation of Euclidean distance between the points p and q will be as shown in equation-1.

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \tag{1}$$

The potential is the Euclidean distance. As this paper aims to get equi-sizes clusters; each center of clusters will be related to a number of closest points with respect to it by calculating the potential using Euclidean distance. The total number of data set is being divided to the number of clusters.

IV.   RESULTS

By grouping data of high potential to specified clusters, each center will group equal number of data points that will be the closest points to it. Thus a point that has high potential with respect to two clusters will be a common point between them. This result is important in industrial applications, and also, in communication systems to determine the common data points that can be used by two clusters or more, thus, managing the cluster transfer and control easily.

Figure-4 shows the result of the new modification, in the same arrangement of figure-3. Many points are omitted from this cluster map, because of its farness from the real cluster center in the same case that there are closer points to the center from those omitted.
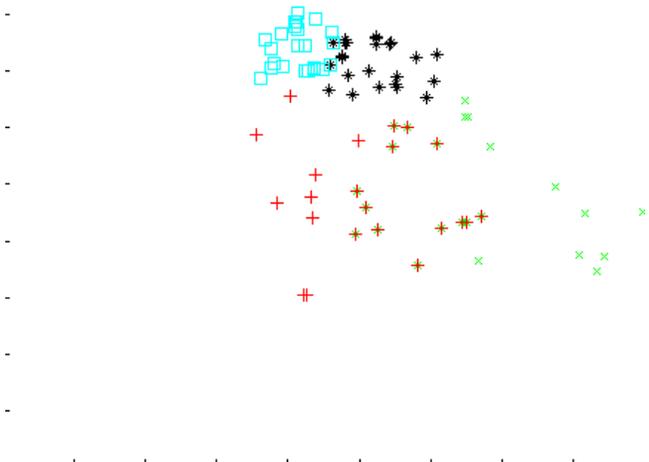
Figure 3. Clustering using Modified Algorithm.

The red cluster and the green one have many common points, those points are very important to be clustered as common in many applications. Also, the same issue between the blue and black colored clusters.

V. CONCLUSION

Many algorithms can be implemented to develop clustering of data sets. Fuzzy C-mean clustering (FCM) is efficient and common algorithm. This research develop the FCM to add computational grouping of data in order to get equal sizes clusters and also to determine the belonging of points that located between clusters. The benefit of this research is to develop a high performance algorithm that sort and group data set in variable number of clusters to use this data in control and managing of those clusters.

In future; the next research will study and develop this algorithm to get a solution for the rest of data point which omitted because of its farness from all clusters.

REFERENCES

[1] Kuo-Lung Wu and Miin-Shen Yang, **"**Alternative c-means clustering algorithms", Department of Mathematics, Chung Yuan Christian University, Chung-Li 32023, Taiwan, 29 November 2001.

[2] Stephen L.Chiu, "Fuzzy model identification based on cluster estimation", Rockwell science Center Thousand Oaks, California 91360, june 1994.

[3] Yue Yafan, Zeng Dayou, Hong Lei, "Improving Fuzzy C-Means Clustering by a Novel Feature-Weight Learning", Dept. of Fundamental Sci., North China Inst. Of Aerosp. Eng., Langfang ; Computational Intelligence and Industrial Application, 2008. PACIIA '08. Pacific-Asia Workshop; Dec. 2008.

[4] K.C. Gowda, E. Diday, "Symbolic clustering using a new similarity measure", IEEE Trans. System Man Cybernet. 22 (1992) 368–378

[5] A. Isazadeh, and M. Ghorbani, "Fuzzy c-means and its generalization by Lp-norm space", WSEAS Transaction on Mathematics, 2 (2003), 168-170