MASS ACTION BASED DATA CLUSTERING METHOD AND ITS WEIGHTED FUZZIFICATION

Umut ORHAN Mahmut HEKİM

e-mail: umutorhan@gop.edu.tr e-mail: mhekim@gop.edu.tr

Gaziosmanpasa University, Electronics and Computer Department, 60250, Tokat, Turkiye

Key words: Mass action law, position of cluster center, weighted fuzzification, MAB clustering

ABSTRACT

Data clustering is an interesting approach for finding similarities in data and putting similar data into groups. A clustering method partitions a data set into several groups such that the similarity within a group is larger than among groups. It has been playing an important role in solving many problems in image processing and pattern recognition. In this study, by inspired from mass action law, a new method called Mass Action Based (MAB) clustering is developed in order to detect the positions of cluster centers without dependence on data density. It is performed by using several data sets of different fields.

I. INTRODUCTION

Data clustering is grouping of objects into homogenous groups based on same object features; and it is considered an interesting approach for finding similarities in data and putting similar data into groups. This approach is an important for image processing; remote sensing, data mining, and pattern recognition. Generally speaking, clustering is one method to find most similar groups from given data, which means that data belonging to one cluster are the most similar; and data belonging to different cluster are the most dissimilar. In the literature, researchers have proposed many solutions for this issue based on different theories, and many surveys focused on special types of clustering algorithm have been presented [1 - 6].

Clustering algorithms are used not only to categorize data, but are also useful for data compression and model construction. By finding similarities, similar data can be represented with fewer symbols. Also, if we can find groups of data, we can build a solution based on those groupings.

According to mass action law, an object is gravitated by other object proportional to its mass and the inverse of square of distance between them. If it is adapted to data clustering, it is supposed that each center of cluster is a mass gravity center and each data point's mass is constant and the equation of mass action law m / d^2 is transformed to $1/d^2$. This idea is the subject of paper. Also, in order to determine all data belonging to each cluster, we propose a new fuzzification method called weighted fuzzification, which is the independent on data density. End of the clustering process, we carry out the optimum positions of cluster centers and membership values.

II. DATA CLUSTERING OVERVIEW

In this section, a detailed discussion of each technique is investigated. Hard clustering, which is also called k-means clustering, is an algorithm based on finding data clusters in a data set such that the cost function of dissimilarity measure minimized. In most cases this dissimilarity measure is chosen as the Euclidean distance [7, 8].

A set of N vectors x_j are to be partitioned into C groups G_i (where i=1,…,C and j=1,…,N). The cost function, based on the Euclidean distance between a vector x_k in group j and the corresponding cluster center c_i, can be defined by:

$$J = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \left(\sum_{k, x_k \in G_i} \left\| x_k - c_i \right\|^2 \right) \tag{1}$$

where J_i is the cost function within group i. The partitioned groups are defined by a $C \times N$ binary membership matrix U, where the element u_{ij} is 1 if the jth data point x_j belongs to group i, and 0 otherwise. Once the cluster centers c_i are fixed, the minimizing u_{ij} for Equation 1 can be derived as follows:

$$u_{ij} = \begin{cases} 1 \text{ if } \left\| x_j - c_i \right\|^2 \leq \left\| x_j - c_k \right\|^2, \ \forall k \neq i \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

This means that x_j belongs to group i if c_i is the closest center among all centers. On the other hand, if the membership matrix is fixed, i.e. if u_{ij} is fixed, then the optimal center c_i that minimizes Equation 1 is the mean of all vectors in group i:

$$c_i = \frac{1}{|G_i|} \sum_{k, x_k \in G_i} x_k \tag{3}$$

where $|G_i|$ is the size of G_i , or $|G_i| = \sum_{j=1}^{N} u_{ij}$.

The method determines the cluster centers c_i and the membership matrix U iteratively. The performance of the K-means algorithm depends on the initial positions of the cluster centers, thus it is advisable to run the algorithm several times, each with a different set of initial cluster centers [3, 4].

Fuzzy c-means clustering relies on the basic idea of k-means clustering. But, each data point belongs to a cluster with a degree of membership while in k-means every data point either belongs to a certain cluster or not. So fuzzy clustering employs fuzzy partitioning such that a given data point can belong to several groups with the degrees of membership between 0 and 1. However, fuzzy clustering also uses a cost function. In this method, the membership matrix U is allowed to have elements with values between 0 and 1. Thus the summation of degrees of membership of a data point to all clusters is always equal to 1 [3, 6, 9]:

$$\sum_{i=1}^{C} u_{ij} = 1, \quad \forall j = 1, \ldots, n \tag{4}$$

The cost function for fuzzy clustering is a generalization of Equation 1:

$$J(u, c_1, \ldots, c_C) = \sum_{i=1}^{C} J_i = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m d_{ij}^2 \tag{5}$$

where u_{ij} is between 0 and 1; c_i is the cluster center of fuzzy group i; $d_{ij} = \|c_i - x_j\|$ is the Euclidean distance between the ith cluster center and the jth data point; and $m \in [1, \infty]$ is a weighting exponent. The necessary conditions for Equation 5 to reach its minimum are

$$c_i = \frac{\sum_{j=1}^{N} u_{ij}^m x_j}{\sum_{j=1}^{N} u_{ij}^m} \tag{6}$$

and

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{7}$$

The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. As in k-means, the performance of fuzzy clustering depends on the initial membership matrix values. Thereby it is advisable to run the algorithm for several times with different values of membership degrees of data points [3,4].

Subtractive clustering method is based on a measure of the density of data points in the feature space. The idea is to find regions in the feature space with high densities of data points. The point with the highest number of neighbors is selected as center for a cluster. Since each data point x_i is a candidate for cluster center, a density measure for first cluster center $c1$ is defined as

$$D_i = \sum_{j=1}^{N} \exp\left(-\frac{\|x_i - x_j\|^2}{(r_a/2)^2} \right) \tag{8}$$

where r_a is a positive constant representing a neighborhood radius. Hence, a data point will have a high density value if it has many neighboring data points. The first cluster center x_{C1} is selected as the point having the largest density value D_{C_1} . Next, the density measure of each data point x_i is revised as follows:

$$D_i = D_i - D_{c_i} \exp\left(-\frac{\|x_i - x_{c_i}\|^2}{(r_b/2)^2} \right) \tag{9}$$

where r_b is a positive constant which defines a neighborhood having measurable reductions in density measure. Therefore, the data points near the first cluster center x_{C_1} will have significantly reduced density measure. After revising the density function given by Equation 9, the next cluster center is chosen as the point having the greatest density value. This process continues until a sufficient number of clusters are attained [4, 7].

III. MASS ACTION BASED DATA CLUSTERING METHOD

In general clustering techniques, according to the dimension of data and the number of clusters, desired result may not be reached for many applications. In common clustering techniques, if the number of clusters is sufficient, then this algorithm can obtain optimum result, otherwise can not. Cluster centers detected by algorithms are fault when the number of clusters is given incomplete or excess. Although the number of clusters is optimum, the position of cluster centers which is calculated by general clustering techniques may not be signed desired point.

In this study, we focus on improving of the cluster center positions. Mass Action Based (MAB) clustering method is a new approach to finding the best cluster centers positions and allowing being reduced computation times, significantly. It uses both of k-means and subtractive clustering methods which are affected by mass action. Here, before subtractive clustering to detect the initial cluster center points for k-means, neighborhood density is calculated for each data points. Then, k-means method detects optimum positions of the cluster centers. While the performance of common clustering methods depends on the initial membership matrix values, this new approach removes the problem of initial condition.

In subtractive clustering, first it is used Equation 8 and continued by Equation 9. But in new approach, Equation 10 is substituted for Equation 8. By using Equation 10 obtained from the action law $m \, / \, d^2$, we may detect the positions of the cluster centers independent of data density.

$$D_i = \sum_{j=1}^{n} \frac{1}{d_{ij}^2} \quad , \quad d_{ij} = \left\| x_i - x_j \right\| \tag{10}$$

Each data point belongs to all cluster centers with the ratio of distances from each cluster centers. In order to remove the parameter *m* used in Equation 7, a new membership matrix *U* is calculated by using Equation 11.

$$u_{ij} = \frac{s_{\min}}{s_j} \sum_{j=1}^{N} \sum_{i=1}^{C} \left(\frac{\sum_{i=1}^{C} \frac{1}{d_{ij}^2}}{d_{ij}^2} \right), \; d_{ij} = \left\| c_i - x_j \right\| \tag{11}$$

where $s_j = \sum_{i=1}^{C} \left\| c_i - x_j \right\|$ and $s_{\min} = \min\{s_j\}$.

This approach determines the cluster centers c_i and the membership matrix *U* using the following steps:

*Step*1. Normalize the data points.

*Step*2. Calculate the density measure of each data point according to Equation 9 and 10.

*Step*3. Select the point having the largest density value as cluster center c_i, and continue until the number of clusters equals to *C*.

*Step*4. Compute the hard membership matrix *U* using by Equation 2 with computed cluster centers.

*Step*5. Calculate the cost function according to Equation 1. Go to Step 7 if it is smaller than selected threshold value.

*Step*6. Update the cluster centers according to Equation 3. Go to Step 4.

*Step*7. Determine the fuzzy membership matrix *U* using by Equation 11.

IV. NUMERIC EXAMPLES

In this section, we show several examples of mass action based clustering to illustrate the ideas presented in the previous section. We first present three simple examples to provide insights into new approach. We then present more realistic examples, and compare performance of new method with those of the corresponding fuzzy and subtractive algorithms.

Example 1: The butterfly data set [7] consists of 15 two dimensional vectors that have two clusters. This data set have used as test data in many experiment and it has symmetric data distribution. We focus on symmetric point belonging to both clusters with membership value of 0.5. Figure 1 shows Mass Action Based Clustering method for butterfly data set and our results for this data set are summerized in Table 1. It shows membership value of cluster centers.

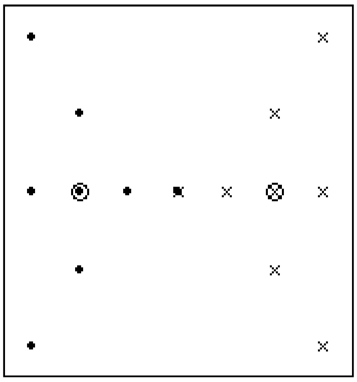
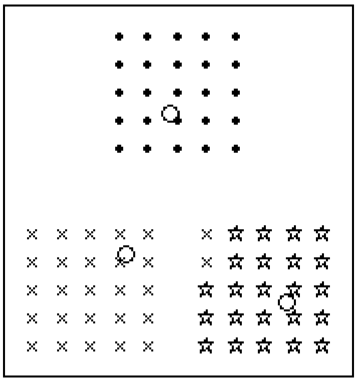


Figure 1. Mass Action Based Clustering method for butterfly data set.

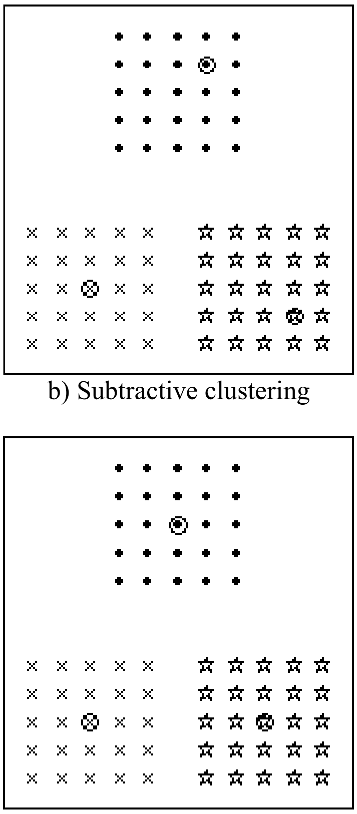
Table 1. Membership values of butterfly data set and cluster centers.

	Cluster 1 (0.83, 0.50)	Cluster 2 (0.17, 0.50)
1.	0.22727	0.77273
2.	0.038462	0.96154
3.	0.22727	0.77273
4.	0.10976	0.89024
5.	0	1
6.	0.10976	0.89024
7.	0.1	0.9
8.	0.5	0.5
9.	0.9	0.1
10.	0.89024	0.10976
11.	1	0
12.	0.89024	0.10976
13.	0.77273	0.22727
14.	0.96154	0.038462
15.	0.77273	0.22727

Example 2: We prepared 75 two dimensional vectors with three clusters for another example. Figure 2 displays this artificial data set which consists of three identical square lattices each containing 25 data points. Our result for this data set is summerized in Table 2. It shows cluster centers obtained by Fuzzy, Subtractive and new MAB methods.



a) Fuzzy clustering



b) Subtractive clustering

c) MAB clustering

Figure 2. Results of clustering techniques of an artificial symmetric data set.

Table 2. Fuzzy, Subtractive and MAB cluster centers for artificial data set.

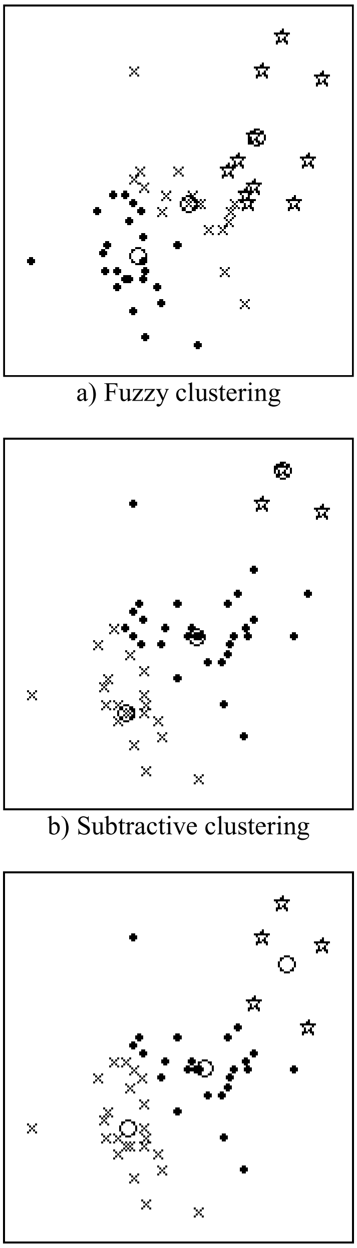
	Cluster 1		Cluster 2		Cluster 3	
Fuzzy	0.33	0.30	0.48	0.75	0.87	0.14
Subtractive	0.20	0.18	0.60	0.91	0.90	0.09
MAB	0.20	0.18	0.50	0.82	0.80	0.18

Example 3: Anderson Iris Data set [10] consists of 150 five dimensional vectors. This data set has often been used as a standard for testing clustering algorithms [6, 11]. The components of a vector are the measurement of the petal length and width, and sepal length and width of a particular iris plant and last fifth vector is added to describe the plant types. There are 50 plants in each of three classes of iris represented in the data. Our results for these data sets are summerized in Table 3.

Table 3. Membership values between every data points and cluster centers.

Fuzzy Clustering					
Cluster 1	0.20	0.58	0.08	0.05	0.01
Cluster 2	0.63	0.42	0.75	0.82	0.99
Cluster 3	0.48	0.37	0.58	0.53	0.50
Subtractive Clustering					
Cluster 1	0.22	0.58	0.08	0.04	0
Cluster 2	0.47	0.42	0.64	0.71	1
Cluster 3	0.39	0.37	0.54	0.5	0.5
MAB Clustering					
Cluster 1	0.20	0.59	0.08	0.06	0
Cluster 2	0.64	0.40	0.77	0.80	1
Cluster 3	0.45	0.32	0.55	0.51	0.5

Example 4: Blood pressure data set [10] consists of 53 two dimensional vectors that have three clusters. This data set has often been used as a standard for testing clustering algorithms [6, 11]. Table 3 summarizes the results for this data set. It shows the optimal positions of cluster centers.



a) Fuzzy clustering

b) Subtractive clustering

c) MAB clustering

Figure 3. Results of clustering techniques of blood pressure data set.

Table 3. Cluster centers obtained by using fuzzy, subtractive and MAB clustering techniques for blood pressure data set.

	Cluster 1		Cluster 2		Cluster 3	
Fuzzy	0.30	0.18	0.50	0.46	0.70	0.62
Subtractive	0.32	0.22	0.56	0.45	0.86	1.00
MAB	0.33	0.27	0.59	0.47	0.88	0.81

Table 3 shows the cluster centers obtained by fuzzy clustering, subtractive clustering techniques and a new method based mass action.

V. CONCLUSION

In this study, we focus on improving of the positions of cluster centers. Mass Action Based (MAB) clustering method is a new approach to finding the best cluster centers positions and allowing being reduced computation times, significantly. It uses both of k-means and subtractive clustering methods which are affected by mass action. While the performance of common clustering methods depends on the initial membership matrix values, this new approach removes the problem of initial condition. Also, for determining all data belonging to each cluster, we proposed a new fuzzification method called weighted fuzzification, which is the independent on data density, and by using it; we obtained the optimum position of cluster centers and membership values.

REFERENCES

1. J. S. R. Jang, C. T. Sun, and E. Mizutani, Neuro-Fuzzy and Soft Computing – a Computational Approach to Learning and Machine Intellence, Prentice Hall, 1997.
2. J. Yu, General C-Means Clustering Model, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.27, No.8, pp.1197-1211, August 2005.
3. E. Xei. and G. Beni, A Validity Measure for Fuzzy Clustering, IEEE Trans. on Pattern Analysis Machine Intelligence, vol. PAMI-13, no. 8, 1991.
4. M. Hekim, U. Orhan, A Validity Measure for a New Hybrid Data Clustering. International Symposium on Innovations in Intelligent Systems and Applications Istanbul, 2007.
5. A. Baraldi, and P. Blonda, A Survey of Fuzzy Clustering Algorithms for Pattern Recognetion-Part I and II, IEEE Trans. Systems, Man, and Cybernetics, Part B, vol. 29, no. 6, pp. 778–801, 1999.
6. J. Han and M. Kamber, Data mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000.
7. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, NewYork,1981.
8. T.J. Ross, Fuzzy Logic with Engineering Applications, McGraw-Hil., 1995.
9. S. Chopra, R. Mitra and V. Kumar, Identification of Rules Using Subtractive Clustering with Application to Fuzzy Controllers, Proceedings of the Third International Conference on Machine Learning and Cybernetics, Shanghai, 26-29 August 2004, pp. 4125-4130.
10. D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway, and E. Ostrowski, A Handbook of Small Data Sets, Chapman and Hall, London, 1994.
11. Y.T. Chein, Interactive Pattern Recognition, Marcel Dekker , New York, 1978.