ICIC Express Letters
Volume **10**, Number **12**, December **2016**
 ISSN 1881-803X
pp. **2845–2851**

DATA FIELD MODEL AND ITS APPLICATIONS OF CLUSTERING ANALYSIS

Tong Li

School of Information Engineering
Beijing Institute of Graphic Communication
No. 1, Xinghua Street, Daxing District, Beijing 102600, P. R. China
zqbbb2016@163.com

Received March 2016; accepted June 2016

Abstract. *Clustering analysis is an unsupervised method to find out the hidden structures in datasets. Most partition clustering algorithms are sensitive to the selection of the initial exemplars, noise and outliers. In this study, exemplar competition algorithm is proposed based on data field model to avoid the bad influence of the random initialization and the outliers. The provided experimental results show that the proposed approach, exemplar competition, is efficient.*
Keywords: Data mining, Clustering analysis, Partitional clustering

1. **Introduction.** Clustering is a technique that tries to discover underlying substructures in a set of unlabeled objects. Clustering analysis has been widely used for data analysis and been an active subject in several research fields [1-4]. There exist various types of approaches, such as spectral clustering [5], hierarchical clustering [6], K-means (KM) [7] and affinity propagation (AP) algorithm [8].

The most popular type of clustering analysis is partitional clustering. The objective of the partitional clustering is to decompose directly the dataset into a set of disjoint clusters, obtaining a partition which should optimize a certain criterion. KM algorithm is unstable because the initial centers are random. AP clustering is able to avoid the impact of the initial centers, but the method is somewhat complex and slow [9].

In our study, exemplar competition (EC) algorithm is proposed. The experimental results show that EC algorithm is simple and efficient. And it is able to avoid the influence of the outliers and the noise.

The rest of the paper is organized as follows. Section 2 gives the summary of data field model. Section 3 is devoted to giving the detailed steps of EC. Section 4 presents the main results. Section 5 gives some conclusions and the future work.

2. **Description of Data Field Model.** Data points of a dataset will be got together or separated by the interaction force between them. And the description of relationship between the data points is called data field model (DFM) [10].

Aggregation energy (AE) is represented as the minus square sum between x_i and its top M nearest neighbors in Formula (1) where E_i is the energy of x_i, x_j is a neighbor of x_i and $1 \leq j \leq M$ in DFM,

$$E_i = -\sum_{j=1}^{M} (x_j - x_i)^2 \tag{1}$$

In DFM, AE of a point is different from others. So, the points in a dataset can be divided into two types which are named exemplar, and member.

exemplar: An exemplar is a point which has the maximum AE in a given cluster.

member: A member is a point whose AE is less than its exemplar in a cluster.

3. EC Algorithm.

3.1. **Description of partitional clustering.** Maybe some points are very close to each other. There are two special situations in the dataset, and each situation will lead to a different competition. One of the situations is that the distance between two members of one cluster may be shorter than that of each member to its exemplar. The other situation is that the distance between two members of different clusters may also be shorter than that of the member to its exemplar. Each member cannot be served as an exemplar of these two cases, because the exemplar represents the energy which mostly focuses on the exemplar of a cluster according to DFM. That is to say, the energy the member owns, is less than that of exemplar. So, we can identify exemplars and members based on DFM. Then, we can divide the members into the corresponding clusters.

3.2. **Competition rule of EC method.** The competition rule, which is used to identify members, must indicate how well the data point is suited for other points to be an exemplar. It is supposed that there is a given dataset $X = \{x_1, x_2, \ldots, x_n\}$. D is the distance matrix and S is the similarity matrix which is set to be the minus D. That is, s_{ij} which is described as the similarity between x_i and x_j should be set to the negative Euclidean distance as Equation (2).

$$s_{ij} = -(d_{ij})^2 = -\|x_i - x_j\|^2 \tag{2}$$

According to Equations (1) and (2), AE can be described into another form as shown in Equation (3) where E_i represents the energy of x_i, x_m represents the mth nearest neighbor of x_i, s'_{im} is the similarity between x_i and x_m, p represents the number of the nearest neighbors of x_i and p is $1 \leq p \leq n-1$.

$$E_i = -\sum_{m=1}^{p}(x_m - x_i)^2 = \sum_{m=1}^{p} -{d_{mi}}^2 = \sum_{m=1}^{p} s'_{im} \tag{3}$$

We can easily draw a conclusion that the point which has the bigger AE than others will be more suitable to be an exemplar. This means that, the exemplars will be reserved at last after some steps. That is, AE as described in Equation (3) is considered as the competition rule in EC method.

3.3. **The strategy of finding out the exemplars of EC algorithm.** A conclusion is that a member could possibly be found between two nearer points through the competition with each other. So, identifying the member point is equivalent to finding out two nearer points.

If D is the distance matrix of X where $X = \{x_1, x_2, \ldots, x_n\}$, D' is obtained by sorting each row of D in ascending order. In this case, the first column of D' is a zero vector.

So, the rth locally minimum distance (LMD) of the lth round of the data competition, i.e., d'_{lr}, is defined as Equation (4).

$$d_{LMD}^{lr} = \min_r \left\{ \underset{1 \leq i \leq n}{D'(i, l+1)} \middle| 1 \leq r \leq n; 1 \leq l \leq q \right\} \tag{4}$$

where q is the rounder number of data competition.

In this case, two points of the first competition in the first round of competition is selected by d_{LMD}^{11}. Similarly, the two points of the rth competition in the lth round of data competition are selected by d_{LMD}^{lr}.

3.4. **The convergence condition of EC algorithm.** It is supposed that EC algorithm will stop after q rounds of competition and l_i members will be identified in each round. When $\sum_{i=1}^{q} l_i = n - K$, the clustering process is over.

3.5. **The description of EC algorithm.** The key idea of EC is to find out the exemplars or identify the members through the data competition. EC algorithm takes as input a collection of real-valued similarities between data points, where the similarity s_{ij} is shown in Equation (2) above.

And EC algorithm consists of the following steps as shown in Figure 1.

begin
1. input a dataset with K clusters and n objects.
2. initialize p=K.
3. calculate the distance matrix D and the similarity matrix S.
4. obtain a new matrix D'' by sorting each row of S in ascending order.
5. obtain a new matrix S' by sorting each row of S in descending order.
6. **do**
7. {
8. choose two nearer points x_i and x_j from D''.
9. calculate AE of x_i, $E_i = \sum_{m=1}^{p} s'_{im}$.
10. calculate AE of x_j, $E_j = \sum_{m=1}^{p} s'_{jm}$.
11. **if** $E_i > E_j$
12. {
13. the point x_j is considered as a member.
14. calculate the number of the identified members.
15. }
16. **else**
17. {
18. the point x_i is considered as a member.
19. calculate the number of the identified members.
20. }
21. }
22. **while** the number of the identified members is not equal to (n-K).
23. obtain the final K exemplars according to the identified members.
24. partition the members into the K clusters.
25. output the label of each object.

end

FIGURE 1. The steps of EC algorithm

In Step 24, there are some methods to divide members into different clusters. In our study, the rule is introduced and described in Figure 2.

4. Experimental Results.

4.1. **The clustering results on the artificial datasets.** In our study, EC is run on some artificial datasets. The dataset with size-even and density-even clusters is shown in Figure 3(a). It contains 300 objects distributed in six spherical clusters where the clusters are well separated and evenly distributed. Figure 3(c) shows that there is a dataset with density-diverse and size-even clusters. There are 1200 objects distributed in eight spherical clusters. These clusters are also well separated and evenly distributed.

```
begin
1.  input the members matrix M, the scale of the
    members n_m, and the exemplars matrix C;
2.  for  i=1 to n_m
3.       TC=C;
4.       x_i is chosen in M where the distance
         between x_i and C is minimum;
5.       x_i is assigned to a cluster through the
         minimum distance of x_i and TC;
6.       add x_i into TC and update TC;
7.       delete x_i from M and update M;
8.  end
end
```

FIGURE 2. The partitional method in EC algorithm

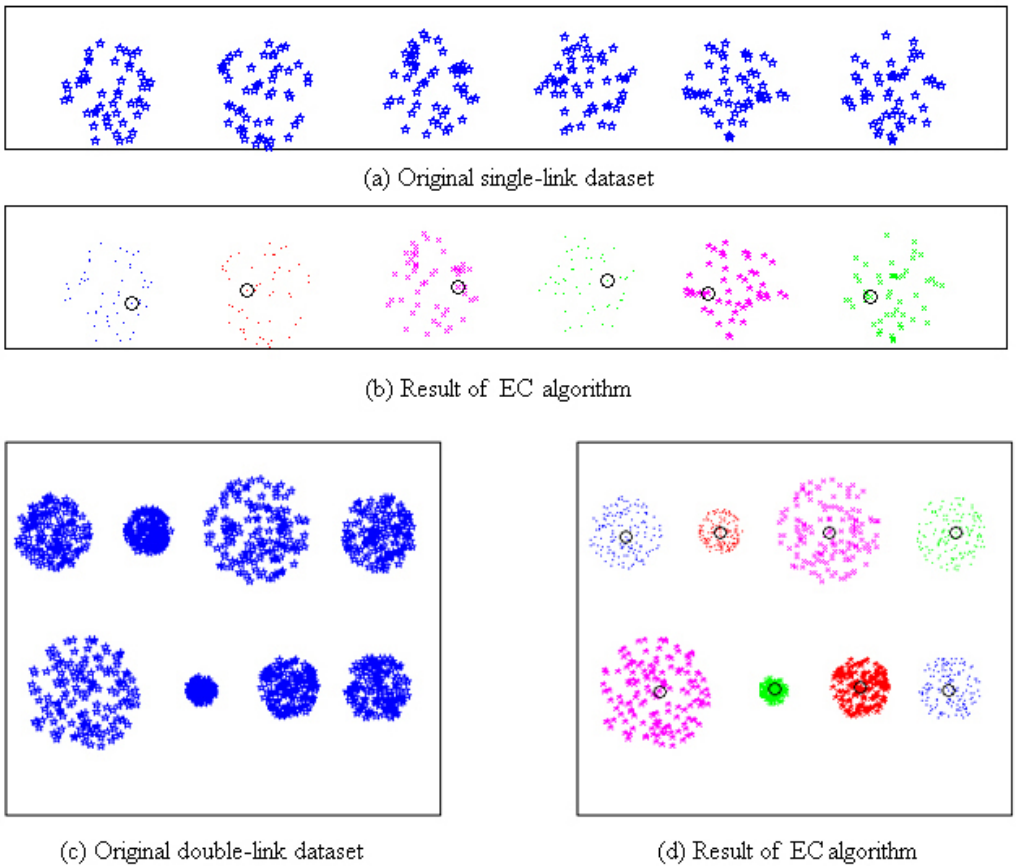

FIGURE 3. The clustering results on size-even datasets

The black circles in Figure 3(b) and Figure 3(d) represent the identified exemplars by EC algorithm. It can be seen from Figure 3(b) and Figure 3(d) that each identified exemplar is distributed evenly in each cluster. That is, if the clusters of a dataset are all size-even, EC algorithm can successfully find out the suitable exemplars and correctly partition.

A dataset with density-diverse and size-diverse clusters is shown in Figure 4(a). It contains 1690 data objects distributed in eight spherical clusters. In our study, the identified exemplars which are represented as red circle are off-center. In this case, the minimum distance rule cannot work well on the complex clusters and the clustering result is shown in Figure 4(b). It can be seen that the partitional result is not good. Three smaller

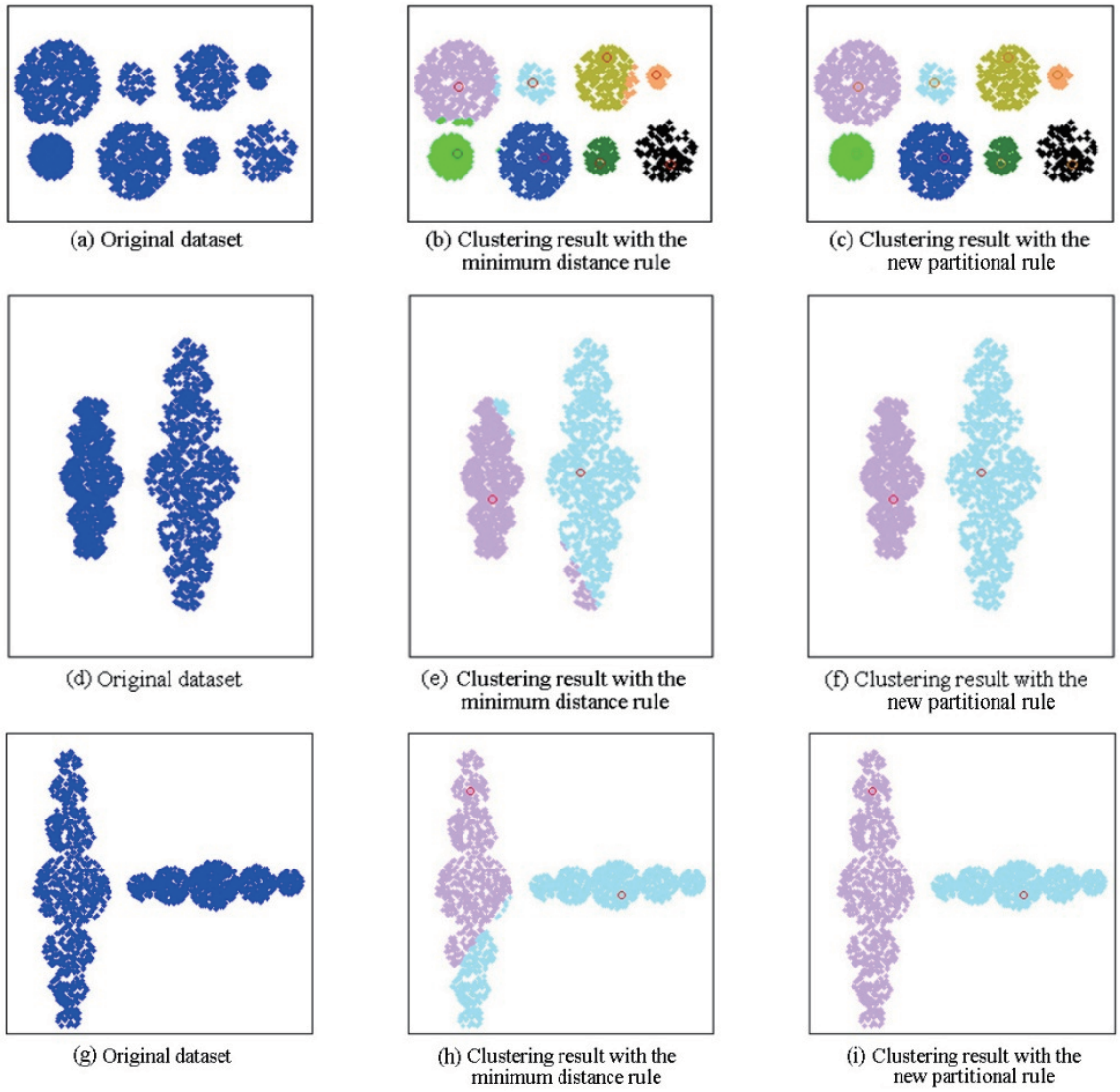

FIGURE 4. The clustering results on complex datasets

clusters eat some objects of bigger clusters. The distance between an object on the edge of the bigger cluster and the exemplar of a smaller cluster is less than that between the object and its exemplar. So, the object is assigned to the smaller cluster. Figure 4(c) shows the clustering result by using the new partitional rule described in Section 3.5. It can be seen that though the identified exemplars are not very good and off-center, the clustering result is still correct.

The datasets with density-diverse and size-even clusters are shown in Figure 4(d) and Figure 4(g). These datasets both contain 2000 objects. In Figure 4(d) two clusters are parallel and strangely distributed and there are two complex and vertical clusters shown in Figure 4(g). For the two parallel clusters, the exemplars represented by red circle are both off-center in Figure 4(e) and Figure 4(f). So, the partitional result with minimum distance rule is not good. There are some objects assigned to the wrong clusters in Figure 4(e). However, if the new partitional rule is used, the result is very good. All the objects are assigned to the right clusters in Figure 4(f). And for the two vertical clusters, the labels of each objects are correct.

The new rule of partitioning members is not the same as the minimum distance rule. As you know, the similarity of a member and a cluster is the distance between the member and the exemplar of the cluster. However, the proximity of them is defined as the minimum of the distance between the member and any one point of the cluster in the new rule. So,

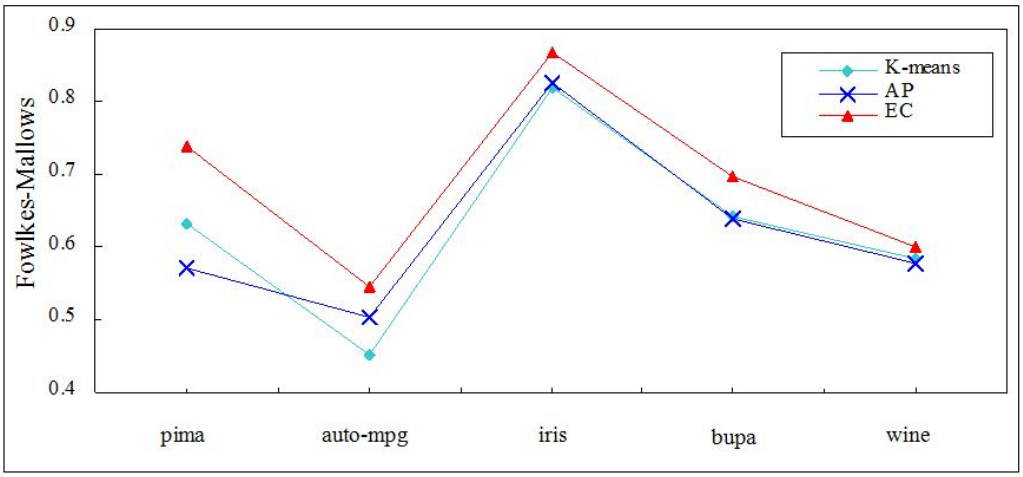

FIGURE 5. Clustering results of five datasets judged by Fowlkes-Mallows

it can find out the structure of the clusters and deal with a variety of clusters. That is, EC algorithm is more robust and efficient.

4.2. **The clustering results on the real datasets.** In this study, KM, AP and EC algorithms are compared on five datasets named pima, auto-mpg, iris, bupa and wine datasets, where datasets can be obtained from UC Irvine Machine Learning Repository [11]. And the clustering results judged by Fowlkes-Mallows (FM) [12] are shown in Figure 5.

Firstly, EC is compared with KM. The FM of EC is greater than that of KM on all datasets in Figure 5. That is, the performance of EC is better than KM. Though EC algorithm will spend more time than KM on the same dataset, EC algorithm can mostly obtain better result than KM. In a word, the conclusion is that the clustering result of EC is better than the best solution of KM. So, EC can outperform KM. Then, EC is compared with AP. It can be seen in Figure 5 that the Fowlkes-Mallows of EC is always greater than that of AP. It is easy to reach a conclusion that EC is better than AP.

In a word, the experimental results suggest that EC algorithm can outperform KM and AP. That is, EC algorithm is more stable and more efficient than others.

5. **Conclusion.** EC method based on DFM is designed which has several advantages. It can not only avoid the unwanted initialization and the bad influence of the outliers, but also reduce the sensibility of result. The experiments show that EC algorithm is efficient. In the future, EC algorithm will be used in color image segmentation.

Acknowledgment. This work is partially supported by Scientific Research Project of Beijing Municipal Education Commission (No. KM201410015005) and Key Project of Beijing Institute of Graphic Communication (No. Ea201507). The author also gratefully acknowledges the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] B. Peng, L. Zhang and D. Zhang, Automatic image segmentation by dynamic region merging, *IEEE Trans. Image Processing*, vol.20, no.12, pp.3592-3605, 2011.

[2] J. G. Sun, J. Liu and L. Y. Zhao, Clustering algorithms research, *Journal of Software*, vol.19, pp.48-61, 2008.

[3] Y. Sui, Yi, Z. M. Lu and P. Yang, An improved spectral clustering algorithm based on low rank approximation for image segmentation, *Journal of Computational Information Systems*, vol.9, no.24, pp.9809-9816, 2013.

[4] P. Gupta, S. Saxena and S. Singh, Color image segmentation: A state of the art survey, *International Journal of Computational Intelligence Research*, vol.8, no.1, pp.17-25, 2012.
[5] W. Y. Chen, Y. Q. Song and H. J. Bai, Parallel spectral clustering in distributed systems, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.33, no.3, pp.568-586, 2011.
[6] A. Fernandez and S. Gomez, Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms, *Journal of Classification*, vol.25, pp.43-65, 2008.
[7] J. B. MacQueen, Some methods for classification and analysis of multivariate observations, *Proc. of the 5th Berkeley Symp. Math. Stat. Prob.*, Berkeley, pp.281-297, 1967.
[8] B. J. Frey and D. Dueck, Clustering by passing message between data points, *Science*, vol.315, pp.972-976, 2007.
[9] S. J. Kiddle, O. P. Windram and S. Mchattie, Temporal clustering by affinity propagation reveals transcriptional modules in Arabidopsis thaliana, *Bioinformatics*, vol.26, pp.355-362, 2010.
[10] Q. Zhang, *Data Competition Algorithm and Its Application Research Based on Data Field Model*, Harbin Engineering University, 2013.
[11] *UCI Machine Learning Repositpory*, http://archive.ics.uci.edu/ml/, 2016.
[12] S. Dudoit and J. Fridlyand, A prediction-based resampling method for estimating the number of clusters in a dataset, *Genome Biology*, vol.3, no.3, pp.1-21, 2002.