International Journal of Fuzzy Logic and Intelligent Systems, vol. 4, no. 3, December 2004 pp. 337-340

Nearest neighbor and validity-based clustering

Seo H. Son*, Suk T. Seo, Soon H. Kwon

***Top Engineering Co., Ltd.**

Department of Electrical Engineering, Yeungnam University

Abstract

The clustering problem can be formulated as the problem to find the number of clusters and a partition matrix from a given data set using the iterative or non-iterative algorithms. The author proposes a nearest neighbor and validity-based clustering algorithm where each data point in the data set is linked with the nearest neighbor data point to form initial clusters and then a cluster in the initial clusters is linked with the nearest neighbor cluster to form a new cluster. The linking between clusters is continued until no more linking is possible. An optimal set of clusters is identified by using the conventional cluster validity index. Experimental results on well-known data sets are provided to show the effectiveness of the proposed clustering algorithm.

Key words : Clustering; Nearest neighbor; Cluster validity

I. Introduction

Clustering is an important research topic which has practical applications in many fields [1-9]. The classical or fuzzy cluster analysis is a technique for grouping data into clusters or classes, where similar data are assigned to the same cluster by assignments of membership grades ranging between zero and one to every data. The clustering problem can be formulated as the problem to find the number of clusters and a partition matrix, which has membership grades as its elements, from a given data set using the iterative or non-iterative algorithms.

A number of clustering algorithms such as hard c-means (or k-means) algorithm, fuzzy c-means (FCM) algorithm [4, 10] and its variants (e. g., Gustafson-Kessel algorithm providing elliptical clusters [11], the possibilistic clustering algorithm [12], the mountain method (MM) [13], the sum of all normalized determinants algorithm [14], the mixed c-means clustering model [15], the volume criteria-based clustering algorithm [16], the fuzzy compactness and separation algorithm(FCS) [17] have been proposed in the literature. Among the clustering algorithms including the hard c-means (or k-means), fuzzy c-means (FCM) algorithm and its variants, the FCM algorithm proposed by Bezdek [5] is the most widely used clustering algorithm.

Most of the clustering algorithms including the FCM are unsupervised algorithms based on the principle of minimizing the within cluster scatter matrix trace. Thus, in most clustering algorithms, the number of clusters reflecting the structure of the given data set should be assumed to be a user-defined value that is so hard to be set in real applications. Due to the reason, the clustering results including the partition matrix and the number of clusters need to be validated by using cluster validation indexes. The cluster validation indexes involve measuring how well the clustering results reflect the structure of the given data set. Many cluster validation indexes including Bezdek's partition coefficient (PC) [18] and partition entropy (PE) [19], Fukuyama and Sugeno's cluster validity criterion [20], Xie-Beni index [21], Bezdek and Pal's index [22], Kwon's index [23], and Kim-Lee-Lee index [24] have been proposed. The main disadvantage of the PC and PE is the lack of direct connection to the geometrical properties of the data and their tendency toward being monotonic with the number of clusters. Others of those take into account simultaneously the properties of the fuzzy membership degrees and the structure of the given data. In particular, Kwon's index provides a solution to overcome tendency toward being monotonic with the number of clusters.

From the above discussion, a clustering algorithm by which not only the partition matrix but also the number of clusters can be obtained without user's pre-assignment to the number of clusters is desirable. Bensaid and et al.'s study [25], that is, validity-guided (re)clustering method is a result toward the direction.

As discussed previously, the basic idea of the clustering is to group the given data into clusters by assigning similar data to the same cluster. If the similarity between data points is measured by distance between data points, we can make a conclusion that each data point should be grouped with the nearest neighbor data point under the predetermined guidance rules or indexes. This motivates the development of the nearest neighbor and validity-based clustering algorithm.

In this paper, we propose a nearest neighbor and validity-based clustering algorithm where each data point in the data set is linked with the nearest neighbor data point to form initial clusters and then a cluster in the initial clusters is linked with the nearest neighbor cluster to form a new cluster. The linking between clusters is continued until no more linking is possible, i.e., just one cluster is found. For the set of clusters obtained by each linking process, the cluster validity is checked by using the conventional cluster validity index. An optimal set of clusters is identified by selecting a

Manuscript received Nov. 18, 2004; Revised Dec. 9, 2004.

set of clusters optimizing the cluster validity index. Experimental results on well-known data sets are provided to show the effectiveness of the proposed clustering algorithm.

2. Nearest neighbor and validity-based clustering

Let a given data set $X=\{x_1, \ldots, x_n\}$ R^p which are n points in the p-dimensional space. A nearest neighbor and validity-based clustering algorithm can be summarized as follow:

Step 1: Calculate distances between data points in $X=\{x_1, \ldots, x_n\}$ R^p.

Step 2: Link each data point with the nearest neighbor data point to form initial clusters.
Set c be the number of initial clusters.

Step 3: Calculate the value of cluster validity index for the c clusters.

Step 4: If $c > 2$, then merge two clusters with the minimum distance into a new cluster, $c \leftarrow c-1$ and go to Step 3; otherwise halt and find the optimal cluster as a set of clusters with an optimal value of the cluster validity index.

In Step 2, the initial clusters are formed by grouping a set of data pairs linked with each other into a cluster. For checking of the cluster validity in Step 3, conventional cluster validity indexes such as the Kwon's index [23] and the Xie and Beni's index [21] can be used. In this paper, we use the Kwon's cluster validity index, which eliminates monotonic decreasing tendency of the Xie and Beni's index by introducing a punishing function, shown in Eq. 1.

$$v_K(U,V;X)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^2\|x_j-v_i\|^2+\frac{1}{c}\sum_{i=1}^{c}\|v_i-\bar{v}\|^2}{\min_{i\neq k}(\|v_i-v_k\|^2)} \quad (1)$$

where $u_{ij}\in\{0,1\}$, $\bar{v}=\frac{1}{n}\sum_{j=1}^{n}x_j$, $1<c<n$, and v_i is i-th cluster center.

In Step 4, the distance between two clusters $d_{c_1c_2}$ is defined as the minimum of distances between a data point contained in a cluster C_1 and a data point in the other cluster C_2, that is,

$$d_{C_1,C_2}=\min_{x_i\in C_1, x_k\in C_2}(\|x_i-x_k\|^2)$$

3. Illustrative example and simulation results

To show the effectiveness of the proposed clustering algorithm, experiments to three well-known data sets given in the two dimensional space are conducted. First, we consider an extended Yager and Filev data set of 15 data points in R2 space listed in the second and third columns of Table 1 [13]. After calculating the distances between data points, we link each data point with the nearest neighbor data point to form initial clusters. The data pairs obtained by the process are shown in the fourth column in Table 1. By grouping a set of data pairs linked with each other into a cluster, we can form the initial clusters, i.e., C_1 = (1, 2, 3, 4, 5), C_2 = (6, 7, 8, 9, 10), and C_3 = (11, 12, 13, 14, 15) shown in the fifth column of the Table 1. The next step is to calculate the value of cluster validity index for the set of clusters, $C=\{C_1, C_2, C_3\}$, by using the cluster validity index given in eqn. (1). The value of the validity index is 2.0195. Step 4 is to merge two clusters with the minimum distance, i.e., the nearest neighbor clusters into a new cluster. In this case, the distance between C_1 and C_3 (i.e., distance between data points 3 and 12) is smaller than that between C_1 and C_2 (i.e., distance between data points 3 and 9) and that between C_2 and C_3 (i.e., distance between data points 9 and 14). Thus C1 and C_3 should be merged into a new cluster. As like as the above, we repeat processes for merging of clusters and evaluation of the cluster set obtained by the merging process until the only one cluster exists. The sixth column of the Table 1 shows a set of clusters C'_1 and C'_2 obtained by merging C_1 = (1, 2, 3, 4, 5) and C_3 = (11, 12, 13, 14, 15) in the fifth column in Table 1. into C'_1= (1, 2, 3, 4, 5, 11, 12, 13, 14, 15) and while preserving $C'_2 = C_2$ = (6, 7, 8, 9, 10). The next step is to calculate the value of cluster validity index for the set of clusters, i.e., $C' = \{C'_1, C'_2\}$. The value of the validity index is 4.2946. And the last step is to select the optimal cluster as a set of clusters with an optimal value of the cluster validity index. In this example, the $C = \{C_1, C_2, C_3\}$ is selected as the optimal set of clusters to the given data set. Fig. 1 scatter plots the clustered data points where symbols 'O', '□' and 'X' denote data points contained in the C_1, C_2 and C_3, respectively. From the experimental results, we see that the proposed clustering algorithm correctly classify data points into three clusters.

Table 1 Clustering results on the extended Yager and Filev's data set in the R^2 space

Data number	x	y	Data pair with minimum distance	Initial clusters	Clusters after merging 1 and 3
1	0.36	0.85	(1,5)	C_1	C'_1
2	0.65	0.89	(2,4)	C_1	C'_1
3	0.62	0.55	(3,4)	C_1	C'_1
4	0.50	0.75	(4,1)	C_1	C'_1
5	0.35	1.00	(5,1)	C_1	C'_1
6	0.90	0.35	(6,9)	C_2	C'_2
7	1.00	0.24	(7.6)	C_2	C'_2
8	0.99	0.55	(8,10)	C_2	C'_2
9	0.83	0.36	(9,6)	C_2	C'_2
10	0.88	0.43	(10,6)	C_2	C'_2
11	0.40	0.28	(11,15)	C_3	C'_1
12	0.51	0.32	(12,15)	C_3	C'_1
13	0.33	0.48	(13,15)	C_3	C'_1
14	0.60	0.22	(14,12)	C_3	C'_1
15	0.45	0.35	(15,12)	C_3	C'_1

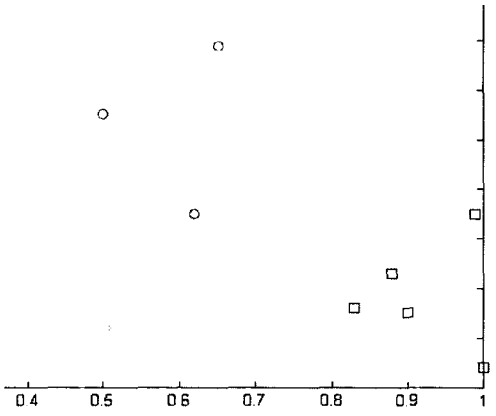


Fig. 1. Extended Yager and Filev data set (optimal number of cluster is three)

The second data set described by Bensaid et al. [25] includes 49 data points given in the R^2space and consists of three clusters. Fig. 2 shows a scattered plot of the data points clustered by the proposed clustering algorithm. From this experimental results, we see that the proposed clustering algorithm correctly classify data points into three clusters.

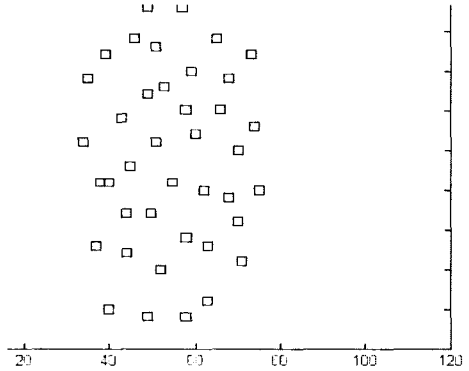


Fig. 2. Bensaid data set (optimal number of cluster is three).

The last data set is the superset of the starfield data set described by Xie and Beni [21]. The data set contains 66 data points known as a data set with eight or nine clusters. Fig. 3 scatter plots the data points clustered by the proposed clustering algorithm. From this experimental results, we see that the proposed clustering algorithm correctly classify data points into eight clusters.

4. Conclusion

We have proposed a nearest neighbor and validity-based clustering algorithm which is simple and effective. In the proposed algorithm without necessity of assumptions on the number of clusters and initial partition matrix, each data point in the data set is linked with the nearest neighbor data point to form initial clusters and then a cluster in the initial clusters is linked with the nearest neighbor cluster to form new clusters. The linking between clusters is continued until no more linking is possible. An optimal set of clusters is identified by using the conventional cluster validity index. Experimental results on well-known data sets were provided to show the effectiveness of the proposed clustering algorithm.

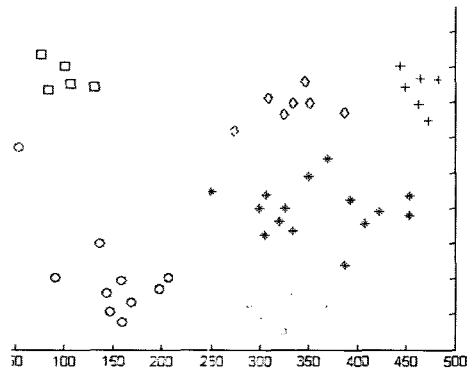


Fig. 3. Extended Starfield data set (optimal number of cluster is eight or nine).

References

[1] M. R. Anderberg, *Cluster Analysis for Application*, Academic Press, New York, 1973.

[2] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[3] J. A. Hartigan, *Clustering Algorithms*, Wiley, New York, 1975.

[4] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[5] A. K. Jain, R. C. Dubes, *Algorithms for clustering*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[6] L. Kaufmann P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York, 1990.

[7] B. S. Everitt, *Cluster Analysis*, 3rd Ed., Edward Arnold, London, 1993.

[8] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, *Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*, Wiley, New York, 1999.

[9] S. H. Kwon, "Threshold selection based on cluster analysis," *Pattern Recognition Letters*, Vol. 25, pp. 1045-1050, 2004.

[10] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," *J. Cybern.* Vol. 3, pp. 32-57, 1973.

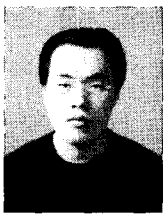
[11] D. E. Gustafson and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. of the IEEE Conf. on Decision Control*, San Diego, CA, pp. 761-766, 1979.

[12] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," *IEEE Trans. Fuzzy Syst.*, Vol. 1, No. 2, pp.98-110, 1993.

[13] R. Yager and D. P. Filev, "Approximate clustering via the mountain method," *IEEE Trans. Syst., Man, and Cybertn.*, Vol. 24, No. 8, pp. 1279-1284, 1994.
[14] P. J. Rouseeuw, L. Kaufmann, and E. Trauwaert, "Fuzzy clustering using scatter matrices," *Comput. Statist. Data Anal.*, Vol. 23, pp. 135-151, 1996.
[15] N. R. Pal, N. K. Pal and J. C. Bezdek, "A Mixed c-Means Clustering Model," *in Proc. FUZZ-IEEE'97*, pp. 11-21, 1997.
[16] R. Krishnapuram and J. Kim, "Clustering algorithms based on volume criteria," *IEEE Trans. Fuzzy Systems*, Vol. 8, pp. 228-236, 2000.
[17] K. L. Wu, J. Yu, and M. S. Yang, "A novel fuzzy clustering algorithm based on a fuzzy scatter matrix with optimality tests," *Pattern Recognition Letters*, (to be published).
[18] J. C. Bezdek, "Cluster validity with fuzzy sets," *J. Cybernet.* Vol. 3, No. 3, pp. 58-72, 1974.
[19] J. C. Bezdek, "Mathematical models for systematics and taxonomy," *in Proc. 8th Int. Conf. Numerical Taxonomy*, G. Estabrook, Ed., Freeman, San Franscisco, CA, pp. 143-166, 1975.
[20] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," *in Proc. 5th Fuzzy Syst. Symp.*, pp. 247-250, 1989 (in Japanese).
[21] X. L. Xie and G. A. Beni, "Validity measure for fuzzy clustering," *IEEE Trans. Pattern and Machine Intell.*, Vol. 13, No. 8, pp. 841-846, 1991.
[22] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *IEEE Trans. Syst. Man Cybern.*, Vol. 28, No. 3, pp. 301-315, 1998.
[23] S. H. Kwon, "Cluster validity index for fuzzy clustering," *Electronics Letters*, Vol. 34, No. 22, pp. 2176-2177, 1998.
[24] D. W. Kim, K. H. Lee and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, Vol. 37, pp. 2009-2025, 2004.
[25] A. M. Bensaid, L. O. Hall, J. C. Bezdek, L. P. Clarke, M. L. Silbiger, J. A. Arrington, and R. F. Murtagh, "Validity-guided (re)clustering with application to image segmentation," *IEEE Trans. Fuzzy Systems*, Vol. 4, No. 2, pp. 112-123, 1996.

Seo H. Son
He received the B.S. and M.S. degree in electrical and computer science from Yeungnam University, Daegu, Korea, in 2000 and 2002. He has completed Ph.D in electrical and computer science from Yeungnam University, Daegu, Korea in 2004. He is currently working with Top Engineering Co, Korea. His research interests include intelligent systems and control, Vision.

Phone : +82-54-480-0446
Fax : +82-54-482-0345
E-mail : shson@topengnet.com



Suk T. Seo
He received the B.S. degree in electrical and computer science from Yeungnam University, Daegu, Korea, in 2004. He is currently in M.S. course in electrical and computer science from Yeungnam University, Daegu, Korea in 2004. His research interests include intelligent systems and control.

Phone : +82-53-810-3932
Fax : +82-53-810-4629
E-mail : kenneth78@yumail.ac.kr

Soon. H. Kwon
He received the B.S. and M.S. degree in Control instrumentation engineering from Seoul University, Seoul, Korea, in 1983 and 1985. He received the Ph.D. in system science from Tokyo institute of Technology, Tokyo, Japan, in1995. He is currently with electrical and computer science in Yeungnam University, Daegu, Korea, as a professor since 1996. His research interests include knowledge-based intelligent systems and control.

Phone : +82-53-810-3514
Fax : +82-53-810-4629
E-mail : shkwon@yu.ac.kr