2003년도 대한전자공학회 하계종합학술대회 제26권 제1호

무게중심 이동을 이용한 데이터 패턴의 추정

장경원, 송윤재, 강진현, 안태천
원광대학교 전기 전자 및 정보공학부

Data Pattern Estimation with Movement of the Center of Gravity

Kyungwon Jang, Yunjae Song, Jinhyun Kang, Taechon Ahn
Dept. of Electrical Electronic and Information Engineering, Wonkwang
E-mail : jaang@wonkwang.ac.kr

Abstract

In This Paper, alternative method for data pattern estimation is proposed and its numerical experiment is carried out. Proposed method gives candidates cluster numbers of given data set between n-2 and 2 by means of movement of the center of gravity. To observe the performance of proposed method, Test sample data sets are offered. Finally, this method is applied to Box and Jenkins's gas furnace data to verify the performance with previous researches.

I. Introduction

In the rule base system modeling, data clustering algorithm has been applied to system identification and its efficiency has been proved in the numerous previous researches and applications [8,9]. The benefit of this algorithm is it could build efficient rule base model with the relatively minimized number of rules against conventional grid partitioning. K-means (or C-means) algorithm is well known as representative data clustering method, such as HCM (Hard C-means Clustering and other numerous k-means related algorithms [1]. Conventional K-means type algorithm requires a priori knowledge about data set and it is highly heuristic to have satisfactory result [1]. Many researchers pointed out these concerns and lots of studies carried out [4-11]. However it seems still controversial.

In this paper, alternative method proposed does not require priori knowledge and gives information about underlying patterns in given data set with less difficulties of dimensionality. Some issues on cluster validation (choice of the proper number of clusters) are addressed with a different point of view. Instead of only one result of the number of clusters, this method presents several candidates of the number of clusters in the given data set. We assume that there is no absolute choice of the number of clusters, which covers all application area. To observe the performance, several sample data offered and its simulation is carried out.

II. Basic concept and numerical example

The overall stream of the proposed method is decreasing the number of cluster from $n-1$ to 2 (**n**: total number of data) by similarity measure between data points. If nearest data pattern pair is detected, fuse the nearest pair as a new cluster and calculate its centroid. This centroid becomes a new point to be measured with rest of data point or cluster in next iteration.

In this section, proposed method is introduced with a simple example. For the example, artificially composed data set is given in table 1 and its scatter plot is also given in figure 1.

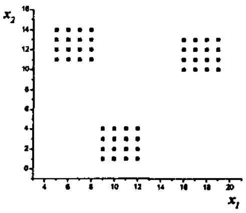


Figure 1. Scatter plot of the example data set

Table 1. Example data

No.	x_1	x_2	No.	x_1	x_2	No.	x_1	x_2
1	5	11	17	9	1	33	16	10
2	5	12	18	9	2	34	16	11
3	5	13	19	9	3	35	16	12
4	5	14	20	9	4	36	16	13
5	6	11	21	10	1	37	17	10
6	6	12	22	10	2	38	17	11
7	6	13	23	10	3	39	17	12
8	6	14	24	10	4	40	17	13
9	7	11	27	11	3	41	18	10
10	7	12	28	11	4	42	18	11
11	7	13	29	12	1	43	18	12
12	7	14	30	12	2	44	18	13
13	8	11	31	12	3	45	19	10
14	8	12	32	12	4	46	19	11
15	8	13	33	16	10	47	19	12
16	8	14	34	16	11	48	19	13

{ Step 1: Similarity measure and the nearest pair detection. Let the given data set $i = 1 \cdots n, j = 1 \cdots m$. Where n is number of data and m is dimension. Its distance matrix $D(n \times n)$ between data points is calculated as

$$d_{ik} = \sum_{j=1}^{m} (x_{ij} - x_{kj})^2, i = 1 \cdots n, k = i+1 \cdots n \qquad (1)$$

Where d_{ik} is component of distance matrix D . From the result of equation (1) find the nearest data pair to be clustered. If there is more than one nearest pair, then choose only one among them.

{ Step 2: Calculate the centroid v_i , and the distance between centroid and data point by equation (2) and (3) respectively. Afterward, calculate the mean value p_i, $i = 1,2,\cdots,n-s$ at the iteration s of each cluster by equation (4). At last, delete the selected pair from given data set X and insert the centroid of the selected pair instead of deleted data pair. Where $|C_i|$ is cardinality of $i-th$ cluster.

$$v_i = \frac{1}{|C_i|} \sum_{k,\, x_k \in C_i} x_k \qquad (2)$$

$$d_i = \sum_{k,\, x_k \in c_i} \|x_k - v_i\|^2 \qquad (3)$$

$$p_i = \frac{d_i}{|C_i|} \qquad (4)$$

{ Step 3: Calculate the mean value $P_s, s = 1 \cdots n-2$ of the p_i , and its increment ΔP by equation (5) and (6) respectively.

$$P_s = \sum_{i=1}^{n-s} \frac{p_i}{C_s}$$

$$= \frac{1}{C_s} p_1 + \frac{1}{C_s} p_2 + \cdots + \frac{1}{C_s} p_{selected} + \cdots + \frac{1}{C_s} p_{n-s} \qquad (5)$$

$$\Delta P_s = P_s - P_{s-1} \qquad (6)$$

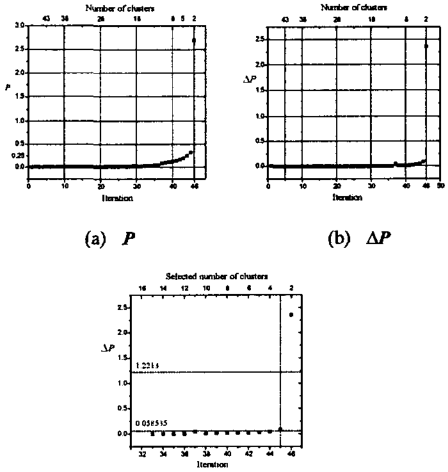
Where $C_s = n - s$ is the number of clusters at the $s-th$ iteration and total iteration of proposed method is $n-2$. Afterward, go to step 1 and iterate step until the number of cluster is reached to 2.

{ Step 4: Calculate the mean value P_{mean} of P_s and ΔP_{mean} by equation (7) and (8) respectively. When iteration is reached to $n-2$ then disregard the ΔP_s valves of ΔP_{mean} below. Afterward, calculate the new ΔP_{mean} and compare ΔP_{mean} values with the new ΔP_{mean} from iteration $n-2$. If ΔP_{mean} is over new ΔP_{mean} , disregard the number of clusters,

$$P_{mean} = \sum_{s_{n-2}}^{n-2} \frac{\Delta P_s}{s_{n-2}} \qquad (7)$$

$$\Delta P_{mean} = \sum_{s=1}^{n-2} \frac{\Delta P_s}{s_{n-2}} \qquad (8)$$

Simulation result of example data is shown in figure 2. As shown in figure (a) and (b), ΔP from iteration 45 to 46 shows sudden change. As a index of the possible number of clusters ΔP_{mean} is applied. ΔP_{mean} eliminate the range that shows almost no change and relatively big change. The possible candidates of the number of cluster is within ΔP_{mean} value 1.2213~0.058535 which the selected number of cluster is 3.



(a) P

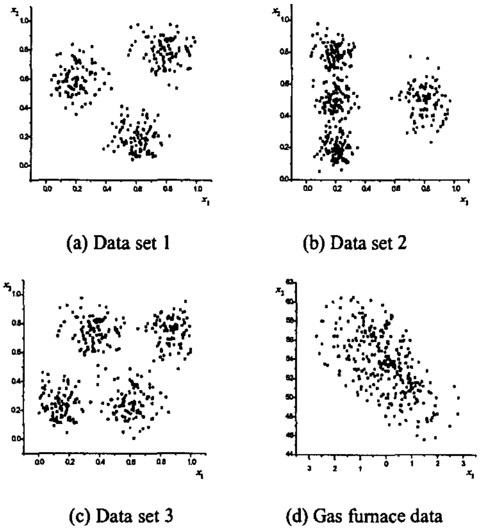
(b) ΔP

(c) Selected number of clusters

Figure 2. Simulation result of example data

III. Numerical Experiments

The given data samples are extracted from FCM demo of the MATLAB fuzzy logic toolbox and Box and Jenkins's gas furnace data [1,10].



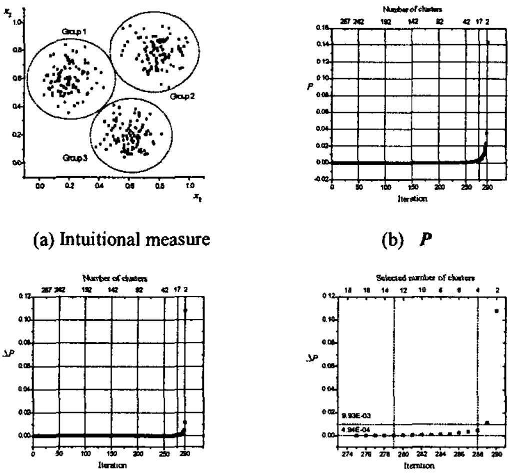
(a) Data set 1 (b) Data set 2

(c) Data set 3 (d) Gas furnace data

Figure 3. Test data samples

3.1 Simulation result of data set 1

As shown in figure 4(a), the given data set show relatively clear separation between data groups and intuitional measure is 3. However, each cluster has more than 3 sub groups in the data set. Figure 4(b) shows variation of *P* values. Figure 4(d) shows range of the possible number of clusters, which is 4～13.



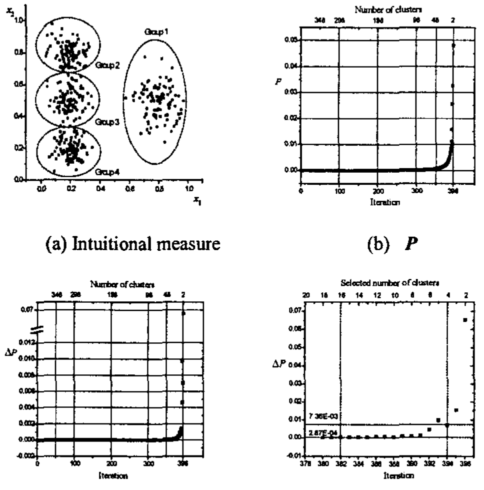
(a) Intuitional measure (b) *P*

(c) ΔP (d) Selected number of clusters

Figure 4. Simulation result of data set 1

3.2 Simulation result of data set 2

Simulation result for data set 2 is shown in figure 5. Intuitional measure of the given data set is 4. In a difference point of view, if we divide the group 1 into two groups, it can be 5. Figure 5 (c) and (d) is simulation result of data set 2. Given data set 2 can be clustered from 4 to 16.



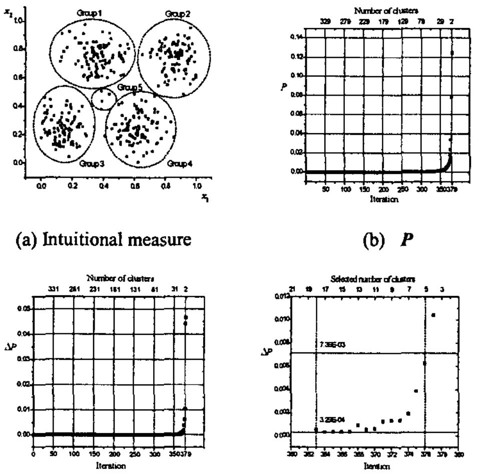
(a) Intuitional measure (b) *P*

(c) ΔP (d) Selected number of clusters

Figure 5. Simulation result of data set 2

3.3 Simulation result of data 3

In the figure 6(a), intuitional measure of the number of clusters is 5 but this data set also has several sub groups in data. As a consequence of simulation, possible range of the number of clusters is from 5 to 18.



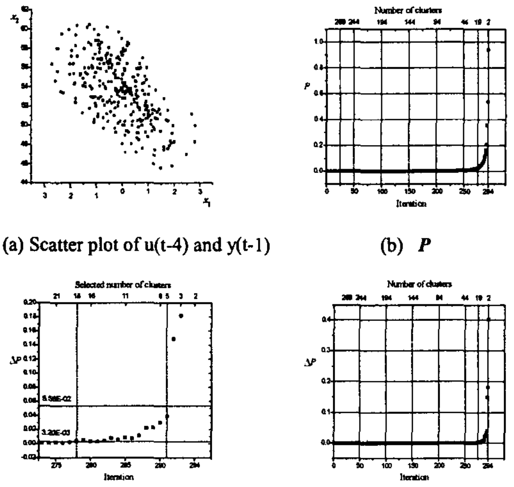
(a) Intuitional measure (b) *P*

(c) ΔP (d) Selected number of clusters

Figure 6. Simulation result of data set 3

3.4 Box and Jenkins's gas furnace data

In this section, the proposed method is applied to Box and Jenkins's gas furnace data *u (t-4) and y (t-1) to verify* performance of proposed algorithm [10]. Result of simulation is verified with reference [8]. Figure 7 (d) shows selected number of clusters. Possible range of the number of clusters is from 5 to 18.



(a) Scatter plot of u(t-4) and y(t-1) (b) ***P***

(c) Δ***P*** (d) Selected number of clusters

Figure 7. Simulation result of gas furnace data

IV. Discussion and Concluding Remarks

In this Paper, data pattern estimation method is proposed and its numerical experiment is carried out. Purpose of this method is detecting possible number of data patterns to avoid difficulties of conventional K-means algorithm. As it shown in simulation results, when data set have distinct number of clusters and shows clear separation, proposed method can detect exact number of clusters. On the other hand, If the cluster is located relatively quite close with adjacent clusters and each cluster has quite large amount of data points proposed method may not bring same result as intuitional measure because the ΔP_{mean} does not consider near boundaries of ΔP_{mean} . In the future research, the obscure values near ΔP_{mean} considered estimation method will be addressed.

References

[1] J-S. R. Jang, C. T. Sun, E. Mizutani, "Neuro-Fuzzy and Soft Computing", Prentice Hall, Inc., NJ., 1997.

[2] I. Gath, A. B. Geva, "Unsupervised Optimal Fuzzy Clurstering", IEEE Trans. on Pattern Alalysis and Machine Intelligence, Vol. 11, No. 7, pp. 773-781, July 1989.

[3] X. L. Xie and G. A. Beni, "Validity Measure for Fuzzy Clustering", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 3, No. 8, pp. 7-31, 1991.

[4] R. Nikhil, K. Pal, J. C. Bezdek, T. A. Runkler, "Some Issues in System Identification using Clustering", IEEE Trans. on Fuzzy Systems, pp. 2524-2529. 1997.

[5] S. H. Kwon, "Cluster Validity Index for Fuzzy Clustering", IEE Electronic Letters, Vol. 34, No. 22, pp. 2176-2177, 29th October 1998.

[6] A. O. Boudraa "Dynamic Estimation of Number of Clusters in Data Set", IEE Electronic Letters, Vol. 35, No. 19, pp. 1606-1607, 1999.

[7] A. F. Sintas, J. M. Cadenas, F. Martin, "Detecting Homogeneous Groups in Clustering using the Euclidean Distance", Fuzzy Sets and Systems, Vol. 120, pp. 213-215, 2001.

[8] Y. H. Joo, H. S. Hwang, K. B. Kim and K. B. Woo, "Linguistic Model Identification for Fuzzy System", IEE Electronic Letters, Vol. 31, No. 4, pp. 330-331, 16th Feburary 1995.

[9] M. Sugeno, T. Yasukawa, "A Fuzzy Logic Based Approachto Qualitative Modeling", IEEE Trans. on Fuzzy Systems, Vol. 1, No. 1, pp. 7-31, 1993.

[10] G. E. P Box and G. M. Jenkins, 'Time Series Analysis - Forecasting and Control', Holden-Day Inc., 1976.