SAHN 모델의 부분적 패턴 추정 방법에 대한 연구

A Study on Partial Pattern Estimation for Sequential Agglomerative Hierarchical Nested Model

장경원*, 안태천**
Kyung Won Jang*, Tae Chon Ahn**

Abstract - In this paper, an empirical study result on pattern estimation method is devoted to reveal underlying data patterns with a relatively reduced computational cost. Presented method performs crisp type clustering with given n number of data samples by means of the sequential agglomerative hierarchical nested model (SAHN). Conventional SAHN based clustering requires large computation time in the initial step of algorithm. To deal with this concern, we modified overall process with a partial approach. In the beginning of this method, we divide given data set to several sub groups with uniform sampling and then each divided sub data group is applied to SAHN based method. The advantage of this method reduces computation time of original process and gives similar results. Proposed is applied to several test data set and simulation result with conceptual analysis is presented.

Key Words : data clustering, pattern estimation, system identification, SAHN, hierarchical clustering

1. Introduction

In the rule based system rule based system modeling, data clustering algorithm has been applied to system structure identification, and its methodological efficiency has been proved by numerous previous researches and applications[1]. the benefit of this algorithm is it enables efficient rule based model design with the relatively minimized number of rules than conventional grid partitioning[1][2].

HCM(Hard C-means Clustering) and FCM(Fuzzy C-means Clustering) method widely applied to rule based system modeling as a representative method. However, this method requires priori knowledge about data set and concern, numerous researches are devoted to determine proper number of cluster and this concern is also called cluster validation[3][4]. Common aspect of cluster validations are frequently resulted in a certain number of clusters that is called *'the proper number of clusters'* form tested range of possible candidates of data distribution. However, selected number of cluster may not give satisfactory result because data clustering is frequently deployed in a application dependant issues, so clustering result or its resulted validity may re-evaluated by a certain object function by their own purpose or requirement[4]. This tendency often appears in the rule based system modeling. However, information about given dataset is still important condition to system modeler. To deal with this issues, we used sequential agglomerative hierarchical nested model based method to estimate underlying data patterns in given dataset. This method gives effective ranges of the number of cluster with less difficulties of dimensionality. The SAHN based method performs clustering form n-1 to 1(n: number of data point) with similarity measure with a sequential hierarchical manner and gives distribution result with *'knee curve'*. However, In the initial stage of this method, this method requires huge computation time[5][6].

In this paper, we presented partial approach to reduce the computational cost of the conventional method. The proposed method divide the entire data set to several sub data set with uniform sampling and performs sequential hierarchical clustering. Experimental study shows quite noteworthy result with the proposed approach.

저자 소개
*圓光大學校 制御計測工學科 博士課程
**圓光大學校 電氣電子 및 情報工學部 正敎授 · 工博

2. Clustering Algorithm and Proposed Approach

2.1 Clustering and pattern estimation parameter

The SAHN algorithm is graph-theoretic model that uses local connectivity criterion instead of objective function such as HCM. In the system identification of rule based system modeling clustering algorithm often deals with compromise two conflicted facts that is efficiency and

accuracy of identified model. Therefore, cluster validity is less crucial than pattern recognition. However, proper selection of the number of cluster is still important. SAHN based algorithm and pattern estimation parameter is described as follows[6]:

1) Similarity measure with nearest pair detection

In the initial stage, set the each datum to be cluster center. With initial clusters, calculate the euclidean distance between clusters for distance matrix $\boldsymbol{D}$ by equation 1 and find minimum distance pair in distance matrix $\boldsymbol{D}$. The minimum distance pair become a new cluster and calculate centroid by equation 2.

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

where, n: number of data
m: dimension of data
D: n×n, distance matrix

$$v_s=\frac{1}{|c_s|}\sum_{k,x_k\in c_s} x_k \tag{2}$$

where, v_s: centroid

2) calculation of the estimation parameter P_s and ΔP_s by following equation.

$$d_s=\sum_{k,x_k\in c_s}\| x_k-v_s\|^2 \tag{3}$$

$$p_s=\frac{d_s}{|c_s|} \tag{4}$$

where, d_s: sum of distance of the inter cluster
p_s: average distance of d_s in $s-th$ selected cluster
$|c_s|$: cardinality of the selected data pair

$$P_s=\frac{1}{C_s}p_s \tag{5}$$

$$\Delta P_s=P_s-P_{s-1}$$

where, C_s: total number of clsuters in $s-th$ eration
P_s: average of p_s
ΔP_s: variation of P_s

This method iterate this sequence until the number of cluster is reached to 1 as shown in figure 1.

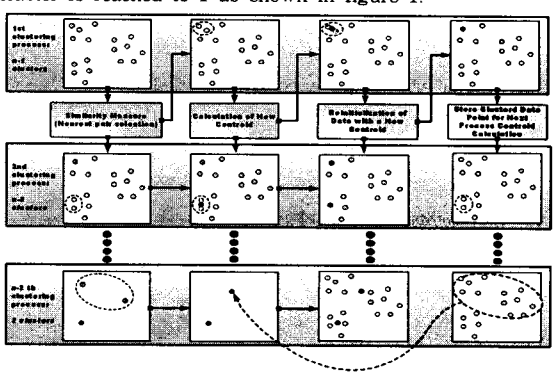
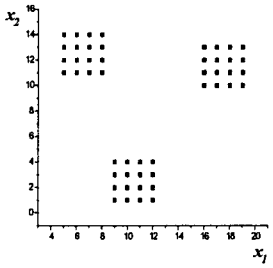
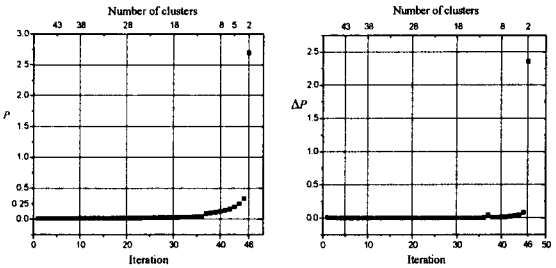


Figure 1. Clustering process

Figure 2. shows numerical example of this method. The given data set in figure 2.(a). shows distinct distribution with 3 data group with clear distribution.



(a) Scatter plot of example data set

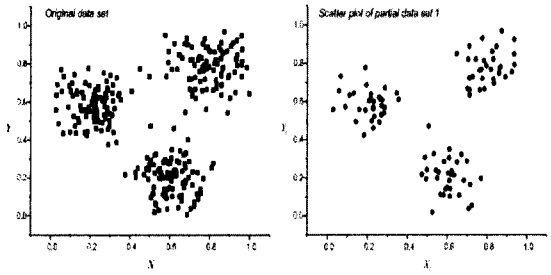


(b) P_s (c) ΔP_s

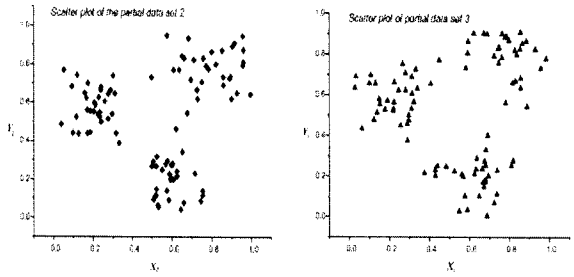
Figure 2. Numerical example

2.2 Proposed approach

In this paper, a partial estimation of the SAHN based method is proposed. The conventional method requires large computational cost in the initial stage of clustering because of distance matrix $\boldsymbol{D}$. To deal with this concern, we divide the given data set to several sub groups with uniform sampling. This approach reduces the initial entry of data so total computation time can be reduced. The purpose of the uniform sampling is to maintain data patterns in the original data set as well as possible. The original data set and divided sub groups are shown in figure 3 for the simulation. The experiment data set have 290 data points and about 3 cluster is observed.



(a) Original data set (b) sub data set 1



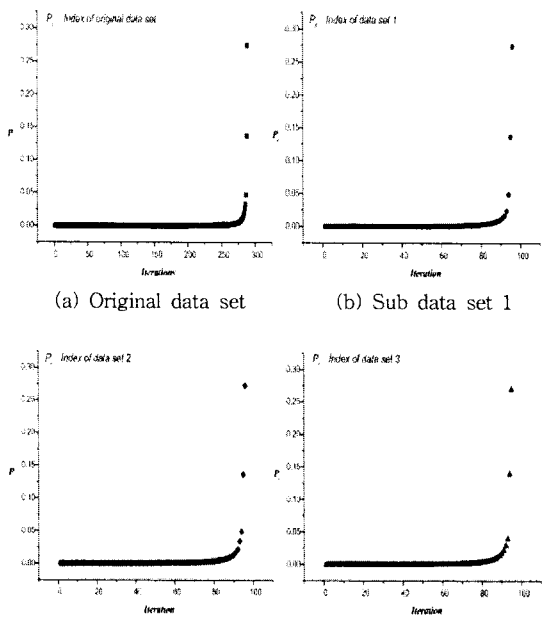
(c) Sub data set 2 (d) Sub data set 3

Figure 3. Scatter plot of data set

2.3 Simulation results

For the experimental study, 290 data points of original data set is applied to this simulation in the figure 3(a). Figure 3(b), (c) and (d) shows uniformly sampled sub data set are shown with scatter plot. Each of data set has 97(sub data set 1), 97(sub data set 2) and 96(sub data set 3) data points respectively.

Simulation results are shown in figure 4. Figure 4(a) is P_s index of original data set with conventional method and figure 4(b), (c) and (d) is P_s result of each sub data set. As shown in the figure 4, P_s results shows similar result as original data set result. However, computation time of conventional method and proposed approach shows quite different result as shown in table 1. The computation time of the original data set is about 45 seconds and the overall computation time of proposed method is about 10 seconds.



(a) Original data set (b) Sub data set 1

(c) Sub data set 2 (d) Sub data set 3

Figure 4. Simulation result of experiment data

Table 1. Computation Time

Method		Computation Time(sec)	
Conventional		45.2985	
Proposed	Sub data set 1	3.7606	10.023
	Sub data set 2	3.1203	
	Sub data set 3	3.1422	

3. Conclusion

In this paper, we presented a partial pattern estimation for SAHN based method. Proposed approach shows quite satisfactory result for the computational cost with a similar result of the original data set. In the future research, we will concentrate on the more accurate cluster validity and practical rule-based model application.

Acknowledgement

This work has been supported by KESRI(R-2004-B-133), which in funded by MOCIE(Ministry of Commerce, Industry and Energy)

References

[1] Sugeno, M. Yasukawa, T.," A Fuzzy Logic Based Approach to Qualitative Modeling", IEEE Trans. on Fuzzy Systems, vol. 1, no. 1, pp. 7-31, 1993.

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

[3] Nikhil, R., Pal, K., Bezdek, J. C., Runkler, T. A., "Some Issues in System Identification using Clustering", Int. Conference on Neural Networks, vol. 4, pp. 2524-2529, 1997.

[4] Sun, H., Wang, S., Jiang, Q.,"A New Validation Index for Determining the Number of Clusters in a Data Set", Proc. IJCNN '01, Neural Networks, vol. 3, pp. 31582-1857, 2001.

[5] Jane, A., Dubes, R.,"A Algorithm for Clustering Data", Prentice-Hall, Englewood Cliffs, 1988.

[6] Jang, K. W., Song, Y. J., Kang, J. H., Ahn, T. C., "Data Pattern Estimation with Movement of the Center of Gravity", Proc., of IEEK Summer Conference, vol. 26, no. 1, pp. 1541-1544, July 2003.