ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

규칙 생성 시스템을 위한 새로운 연속 클러스터링 조합☆

New Sequential Clustering Combination for Rule Generation System

김 승 석* 최 호 진**
Sung Suk Kim Ho Jin Chio

요 약

본 논문에서는 수치적 데이터를 이용하여 규칙을 생성하는 시스템에 대해 순차적인 클러스터링 방법을 제안한다. 단일 클러스터링 기법은 방대하고 복잡한 공간 내에서는 원하는 결과를 얻지 못할 수 있다. 이런 문제점을 해결하기 위해 제안된 방법은 서로 다른 클러스터링 기법을 순차적으로 수행하여 장점들은 활용하고 단점들은 보안하는 형태를 제안하였다. Mountain 클러스터링과 Chen 클러스터링을 이용하여 non-parametric 공간에서 자율적으로 클러스터를 구성하였고, global 공간과 local 공간으로 역할을 분담하여 클러스터를 추정한다. 추정된 클러스터들은 신경회로망이나 퍼지 시스템과 같은 지능 시스템의 구조와 초기 파라미터 결정에 활용될 수 있으며, 확장하여 헬스케어와 의료 분야에서의 결정 제공 시스템의 학습에 도움을 줄 수 있다. 제안된 방법을 유용성을 시뮬레이션을 통해 보이고자 한다.

ABSTRACT

In this paper, we propose a new clustering combination based on numerical data driven for rule generation mechanism. In large and complicated space, a clustering method can obtain limited performance results. To overcome the single clustering method problem, hybrid combined methods can solve problem to divided simple cluster estimation. Fundamental structure of the proposed method is combined by mountain clustering and modified Chen clustering to extract detail cluster information in complicated data distribution of non-parametric space. It has automatic rule generation ability with advanced density based operation when intelligent systems including neural networks and fuzzy inference systems can be generated by clustering results. Also, results of the mechanism will be served to information of decision support system to infer the useful knowledge. It can extend to healthcare and medical decision support system to help experts or specialists. We show and explain the usefulness of the proposed method using simulation and results.

☞ keyword : Sequential clustering, Mountain clustering, Local clustering, Decision rule

1. Introduction

In a decision support system, fuzzy inference system and intelligent system modeling, constructing and managing rules are important to evaluate the system efficiency [1-4]. An expert system is one of approaches to generate a useful decision support mechanism to help human's decision with expert knowledge [1]. A decision support system which depends on a data driven approach generates an unbiased rule mechanism in order to construct an inference engine of decision without human's effort. Compared with the expert system approach, data-driven rule generation has several advantages including the reuse of collected data and unbiased results [1,3,5]. Rule generation is the pre-processing step before executing the inference engine of a decision support system [2,4]. The clustering approach can generate cluster information such as cluster centers which relates relevant rules of inference system [1,3,5]. So, clustering results are important to construct successful overall system.

Clustering algorithm can largely classify two types in parametric operating features as parametric [1] and

* 정 회 원 : Research assistant professor, Computer Science, Korea Advanced Institute of Science and Technology
powerkimss@kaist.ac.kr
** 정 회 원 : Associate professor, Computer Science, Korea Advanced Institute of Science and Technology
hojinc@kaist.ac.kr
[2012/04/26 투고 – 2012/05/06 심사 – 2012/08/02 심사완료]
☆ This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST)(No. 2012-0001001).
☆ A preliminary version of this paper appeared in ICONI 2011, Dec 15-19, Sepang, Malaysia. This version is improved considerably from the previous version by including new results and features.

http://dx.doi.org/10.7472/jksii.2012.13.5.1

non-parametric method [1][6-8].

IMC-2 method [5] can try to improve better quality clustering of mountain method and compares performance using Global Silhouette Value and Separation Index. Chatzis et al. [6] proposed possibility formulation of mixture model in clustering method instead probabilistic formulation.

In this paper, we focus on non-parametric clustering to improve the clustering results for rule generation using modified mountain and Chen clustering techniques [1,9,10]. Although having good characteristics, mountain clustering can suffer information loss during cluster generations and destroy operation. We concentrate on preventing information loss during the destroy operation which more considers neighbor clusters with high density fields. To find neighbor cluster in local fields, we use a modified Chen clustering method to obtain more meaningful clusters in the destroyed spaces. Finally, we propose a rule generation method which can detect neighbor clusters in high density cluster groups. To simplify and clear the work, we generate a clustered data space using intended random data sets and apply to the proposal. In the simulations, we show usefulness of proposed clustering to obtain rules.

In Section 2, we review the characteristics of mountain and Chen clustering. In Section 3, we explain proposed rule generation mechanism using proposed clustering combination. Then, Section 4, we address the usefulness of the proposed method using simulations and results. In Section 5, we conclude the proposed method and future works.

2. Related Work

2.1 Mountain Clustering

In non-parametric clustering, a mountain clustering [1] is cumulated density based approach.

Cluster candidate $m(\mathbf{v})$ is estimated by following density.

$$m(\mathbf{v}) = \sum_{i=1}^{N} \exp\left(-\frac{\|\mathbf{v} - \mathbf{x}_i\|^2}{2\sigma^2} \right) \quad (1)$$

where $m(\mathbf{v})$ is considered as cluster center candidate, N is data size, $\mathbf{x}_i$ is i th data, and σ is construction parameter of mountain.

Especially, the σ is very important to generate shape and height of mountain. If it is large, then each $m(\mathbf{v})$ also gets large value and generated mountain shapes are more smooth than smaller σ.

The resolution of grid partition for center candidate vcanhandleaccuracyofinferenceresults.Thelowresolutionofgrid partitioncanmakeoffastcomputationspeedbutaccuracyshouldbel ow.Theoppositecasehashugecomputationalloadandobtainsmore accuratecenters. Two variables, resolution of vand σ, are major consideration of constructing the mountains.

After the constructing the mountain densities, an algorithm performs destroy operation to obtain next cluster in information as follows.

$$m_{new}(\mathbf{v}) = m(\mathbf{v}) - m(\mathbf{c}_1)\exp\left(-\frac{\|\mathbf{v} - \mathbf{c}_1\|^2}{2\beta^2} \right) \quad (2)$$

where $m(\mathbf{c}_1)$ is a center of chosen maximum cumulative density and β is destroying parameter to the mountain.

The value of β determines destroy radius and depth which affected estimated cluster number and relevant important parameters.

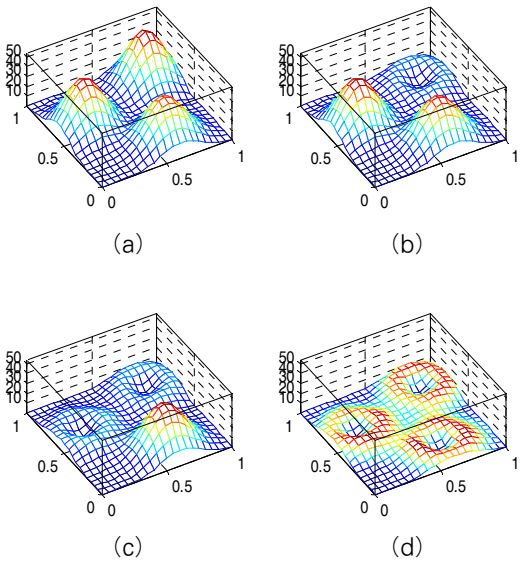
The small destroy mountains will be generated by small β values.

The small destroy mountains sometimes cannot remove useless mountains and can remain undestroyed field to bad effect next step.

In this case, no meaningful clusters can be generated by less enough destroy mountains. Or, if larger β can also generate other problems. The large destroy mountain can remove essential cluster centers.

More specifically, in the complex data distribution, determining of scale parameter σ and β are important step. However, mountain clustering cannot determine relevant mountain in heavy bias density distributions.

The destroy progress of mountains are shown in Fig. 1. As shown Fig. 1, mountain clustering has unique characteristics by cumulative densities and destroy operation.



(a) constructing mountain, (b) first destroy operation, (c) second destroy operation, (d) final results.

(Fig. 1) Mountain clustering.

Maximum top value is considered as a cluster center and then will be destroyed for obtaining the next cluster center using finding maximum cumulative density. Although having good performances, destroy operation of mountain cluster can eliminate neighbor clusters during destroy operation which its cumulative density was dramatically reduced.

The destroy function generally uses Gaussian function. If meaningful neighbor cluster is place in effective field, then its cumulative density information can be damaged.

When predetermined criterion is satisfied, mountain clustering method is terminated. Designer can select appropriate termination conditions.

2.2 Chen Clustering

In the density based clustering, Chen clustering [8,9] has a unique characteristics.

General clustering almost considers each center's relation such as a membership grade or a cumulative density and so on.

However, Chen clustering only considers updated cluster center itself. General similarity measure with centers uses general Gaussian method as follows.

$$r_{ij} = \exp\left(-\frac{1}{2}\frac{\|v_i - v_j\|^2}{\sigma^2}\right) \tag{3}$$

where v is cluster center candidates and algorithm only calculates each center candidates without original data set. The data set only use initial setting of center candidates. Algorithm arranges similarity using pre-determined value as follows.

$$r_{ij} = \begin{cases} 0, & \text{If } r_{ij} < \zeta \\ r_{ij}, & \text{Otherwise} \end{cases} \tag{4}$$

where ζ is a cutting threshold in the Chen clustering. If specific relation or similarity r_{ij} has low value, its similarity is excluded in the estimation. New center of the clustering calculates as follows.

$$v_i^{'} = \frac{\sum_{i=1}^{n} r_{ij} v_j}{\sum_{i=1}^{n} r_{ij}}, i = 1,2,\ldots,n \tag{5}$$

where v is cluster center candidate and unfortunately, centers have same number of data sizes. The computational load of the algorithm can exponentially increase when data set are large.

Chen clustering has unique characteristics like equation (4). Restricted similar measure eliminates small similarity or data having large distance. Also, updated centers did not measure to data, measure process only updated centers themselves as shown equation (3). In algorithm, data is only used at allocating centers of initialization step. Although usefulness of algorithm, a computational load of Chen clustering rapidly is increased if number of the given data sets are increased. Most algorithms iteratively calculates restricted parameters. However, Chen clustering repeatedly calculates number of data size per a step. It is weakness of Chen clustering. However, without data set during learning

progress, Chen clustering has benefit of some bias density environment.

3. Proposed Clustering as Rule Generation

To improve limited performance abilities of single clustering method, we sequentially combine two clustering methods to obtain more precision clustering results. It performs more detailed clustering process in local fields when whole space dividing process using another clustering method is finished

The first step of the proposed method performs mountain clustering for overall data set. In this paper, the sigma which is mountain parameter has a relatively large scale value and the detailed tuning procedure will be handled by Chen clustering. In fact, relatively small value of sigma induces many clusters including many meaningless cluster candidates in mountain clustering.

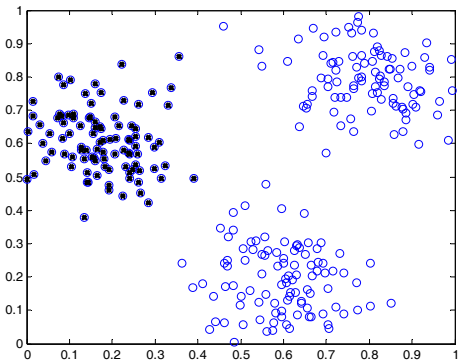
In the second step, Chen clustering performs the estimation of the cluster centers in restricted local fields. Local field selection was based on a similarity measure with selected center and data.

A variable called second sigma which is Chen clustering parameter uses the calculation measurement in Gaussian similarity and no relevant data can be missed if the similarity is lower than a threshold value. Fig. 2 shows selected local field in the entire data set having 3 clusters (groups).

The local field approach can reduce heavy computation load problem of Chen clustering.

After selecting local fields, Chen clustering searches each local space and then can obtain sub clusters. In Chen clustering, a cluster center does not consider other cluster centers to estimate a selected center. Although data states alone or few distributions in rare field, Chen clustering also allocates center with low density. This characteristic is unique to Chen clustering. We do not want to increase the size of cluster centers which have low density distributions.

After the Chen clustering step, a deletion process operates to find meaningless centers with low densities and excludes final cluster center candidates.



(Fig. 2) Candidates of local field.

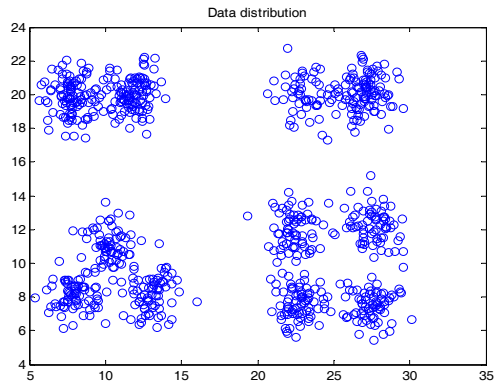
Rule generations in decision system using mathematical computation methods are heavily depend on numerical data distributions. Density distribution of input-output data can explain structures or causalities of the system. It means that rules can be induced through data density combination in given data spaces. So, high intensive density fields in data space can consider the candidates of the rules or system characteristics in generating decision system. The clustering method can find and induce meaningful density information.

Especially, mountain clustering cannot extract neighbor clusters in large data density fields when inappropriate threshold values are determined. It is structural disadvantage because of destroy mountain has radial basis field. Also, Chen clustering has exponentially increased computation load if data set has large. Two non-parametric based clustering methods have each own disadvantages. However, hybrid combination, including global search of mountain clustering and local detail search of Chen clustering, will help overall clustering performance and enhance weakness each other.

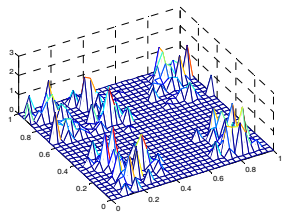
4. Experimental Results and Discussion

4.1 Simulation and Results

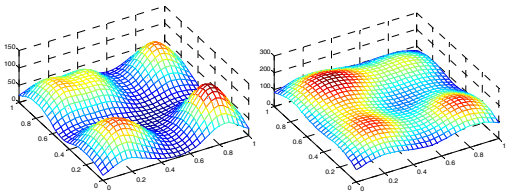
In simulation, we sequentially illustrate the progress of proposed method to demonstrate usefulness. For the purpose of experiment, we have generated intended random data



(Fig. 3) Total data distribution.



(a) sigma=0.005



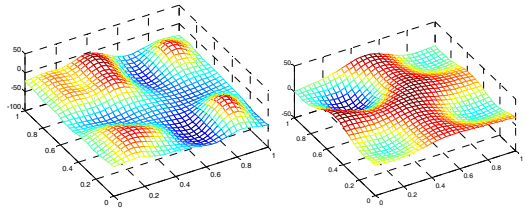
(b) sigma=1 (c) sigma=2

(Fig. 4) Generating Mountain using cumulative density

sets. And, the data sets were spread to several fields of whole space. Through the spreading and grouping process, data sets were divided and clustered. Total data distribution is shown in Fig. 3.

As shown in Fig. 3, four large clusters existed and each large cluster has 2 to 4 sub-clusters with similar or different density distributions. For the sake of convenience, we have normalized the data values from 0 to 1.

Fig. 4 shows generating mountains in the mountain clustering step when sigma is 0.005, 0.1 and 0.2. Choosing appropriated sigma is important to obtain correct structures



(a) sigma = 1 (b) sigma = 2

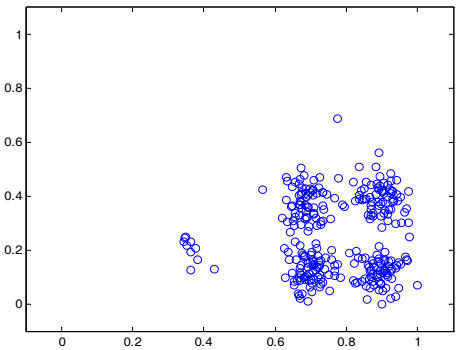
(Fig. 5). Destroy operation.

to generate rules. We can see various smoothness or sharpness of mountain.

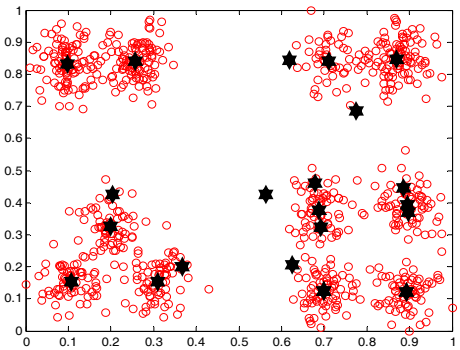
Changing sigma generated different shapes of mountains using cumulative density. Relatively small sigma value influences more than sharp many mountains and large sigma influences smoothing shape few mountains. As shown figures, through small sigma, algorithm can also obtain meaningless cluster centers. Alternatively, in large sigma, algorithm cannot obtain essential cluster centers because strong smoothness can covers neighbor essential cluster candidates.

After five destroy operations in mountain clustering when beta is 0.2, new destroyed data distribution is shown in Fig. 5 at five destroy iterations. Compared with Fig. 4, high density regions are dramatically decreased. Especially scale (height) of z-axis is rapidly reduced.

Next, in the Chen clustering step, new data of the relevant local fields are extracted from mountain clustering. A center with best cumulative density is chosen, then relevant neighbor data are chosen using similarity threshold as shown in Fig. 6. This data field is a part of total data distribution and chosen local field data set have reduced data numbers. Compared with whole data distribution in Fig. 3, it can easily confirm that intensive distribution fields with high densities were selected in Fig. 6. Local field data selection also depends on Gaussian similarity. As shown in Fig. 6, a selected local field may include some data of other cluster because some data with near distance can be within threshold. If sigma of mountain clustering is small case as sharp condition, many meaningless centers can be generated. However, proposed method uses mountain clustering to detect sub-local fields for extracting individual centers using Chen clustering. Proposed method obtains four divided local fields. Chen clustering finally extracts cluster centers in



(Fig. 6) Selected first local field.



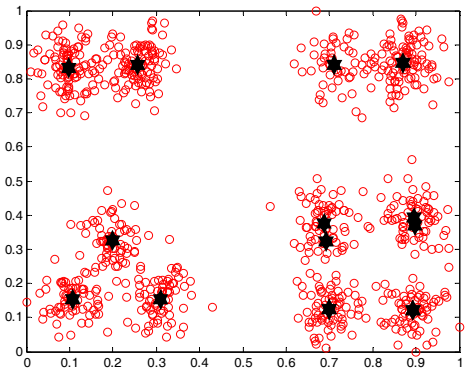
(Fig. 7) Final results without refinement.

reduced spaces or local fields. The results of the estimated centers and data distribution are shown in Fig. 7.

Due to unique characteristic of Chen clustering, some estimated centers have low data distribution. In an extreme case, a center may have only one data relation with sparse distribution. It is one of disadvantages of Chen clustering. To remedy this characteristic, proposed method eliminates meaningless centers with lower densities than pre-determined threshold.

Final inferred cluster centers as rules are shown in Fig. 8.

We display that centers with meaningless position are eliminated in Fig. 8. Rule generation results are summarized in Table 1. We show that rule generation detects 6 rules in class 1 which has 4 sub classes. If the elimination process, which compared with Fig. 7 and 8, is not performed, detected cluster numbers of each column in Table. 1 will be increased and meaningless information can also corrupted final results.



(Fig. 8) Final results with refinement

(Table 1) Rule generation results.

Class	Sub class	Detecting	Remark
1	4	6	
2	3	3	
3	2	2	
4	2	2	Different density

4.2 Discussion

The proposed method has combined two non-parametric clustering approaches to obtain rules in the numerical data spaces. Realistic considerations are pre-determining parameters. To perform overall process, we have to set sigma and zeta variable of mountain clustering, and sigma, similarity and cutting threshold of Chen clustering. However, both sigmas have some predictable relation and threshold values also relate Gaussian similarity. In complex distribution environment, a single clustering approach cannot obtain desired performance results because its own disadvantages. In this case, clustering combined methods can easily solve the difficulties. Extended future researches will be considered.

5. Conclusions

In this paper, we proposed a new rule generation mechanism using sequentially combined clustering methods. As illustrated, the second clustering method compensates the

information loss of the first clustering. Because only performs at local cluster field, computational load of Chen clustering method is reduced. Each clustering has weak points. We intend to obtain advanced performance using combined clustering structure to recover and enhance other weak points and advantages. And, we show and explain the simulation results in dividing local fields and extracting cluster centers.

To get satisfied decision procedure, it is one of the essential approaches that the intention of inference engine can serve useful information to clustering procedure. Integrated model with constructing clustering and optimizing inference mechanism is also a favor and important issue. Our future research in the clustering will modified learning rule by the intention of the overall decision system.

References

[1] J.S.R. Jang, C. T. Sun, E. Mizutani, Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence, Prentice Hall, 1997.

[2] M.G. Tsipouras, T.P. Exarchos, D.I. Fotiadis, A. Kotsia, A. Naka, L.K. Michalis, "*A Decision Support System for the Diagnosis of Coronary Artery Disease*,"19th IEEE international Symposium on Computer-based Medical System 2006 (CBMS 2006), pp. 279-284, 2006.

[3] S. Thedoridis, K.Koutroumbas, Pattern Recognition Third edition, Academic Press, 2006.

[4] J. He, H.J. Hu, R. Harrision, P. C. Tai, Y. Pan, "*Rule Generation for Protein Secondary Structure Prediction With Support Vector Machines and Decision Tree*," IEEE Trans on. NanoBioscience, Vol. 5, Issue. 1, pp. 46-53, 2006.

[5] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice hall, 1999.

[6] P. Agrawal, N. K. Verma, S. Agrawal, S. Vasikarla, "*Color Segmentation Using Improved Mountain Clustering Technique Version-2*," 2011 Eighth International Conference on Information Technology: New Generation, pp. 536-542, 2011.

[7] S. P. Chatzis, G. Tsechpenakis, "*A possibilistic clustering approach toward generative mixture models," Pattern Recognition*, Vol. 45, Issue. 5, pp. 1819-1825, 2012.

[8] J. T. Rickard, R. R. Yager, W. Miller, "*Mountain Clustering on Nonuniform Grids*," International Symposiumon Information Theory 2004 Proceddings, pp. 106-111, 2004.

[9] C.C Wong, C.C. Chen, "*A Hybrid Clustering and Gradient Descent Approach to Fuzzy Modeling*," IEEE Trans on Systems, Man, and Cybernetics-Part B: Cybernetics, Vol. 29, No. 6, pp. 686-693, 1999.

[10] C.C. Wong, C.C. Chen, M.C Su, "*A novel algorithm for data clustering*," Pattern Recognition, Vol. 34, Issue. 2, pp. 425-442, 2001.

◐ 저 자 소 개 ◑

Sung Suk Kim

BS, Electrical Engineering, Chungju National University, 1997
MS, Electrical Engineering, Chungbuk National University, 2002
PhD, Electrical Engineering, Chungbuk National University, 2005
2010 – Present, Research assistant professor, Computer Science, Korea Advanced Institute of Science and Technology
Interest: Artificial Intelligence, Intelligent System, Data Mining

Ho Jin Choi

BS, Computer Engineering, Seoul National University, 1982
MS, Software System Design, University of Newcastle, England, 1985
PhD, Imperial College London, England, 1995
1997-2002, Assistant professor, Korea Aerospace University
2002-2009, Associate professor, Korea Information & Communications University
2009-Present, Associate professor, Computer Science, Korea Advanced Institute of Science and Technology
Interest: Artificial Intelligence, Software Engineering, Knowledge Engineering