IEIE Transactions on Smart Processing and Computing, vol. 4, no. 4, August 2015
http://dx.doi.org/10.5573/IEIESPC.2015.4.4.209

A Simulation Study on The Behavior Analysis of The Degree of Membership in Fuzzy c-means Method

Takeo Okazaki[1], Ukyo Aibara[1] and Lina Setiyani[2]

[1] Department of Information Engineering, University of the Ryukyus / Okinawa, Japan
okazaki@ie.u-ryukyu.ac.jp, ukyo@ms.ie.u-ryukyu.ac.jp
[2] Department of Information Engineering, Graduate school of Engineering and Science, University of the Ryukyus / Okinawa, Japantya.sachi@ms.ie.u-ryukyu.ac.jp

***** Corresponding Author: Takeo Okazaki

Received July 15, 2015; Revised August 5, 2015; Accepted August 24, 2015; Published August 31, 2015

** Short Paper*

** Extended from a Conference: Preliminary results of this paper were presented at the ITC-CSCC 2015. This paper has been accepted by the editorial board through the regular reviewing process that confirms the original contribution.*

***Abstract*:** Fuzzy c-means method is typical soft clustering, and requires a degree of membership that indicates the degree of belonging to each cluster at the time of clustering. Parameter values greater than 1 and less than 2 have been used by convention. According to the proposed data-generation scheme and the simulation results, some behaviors in the degree of "fuzziness" was derived.

***Keywords*:** Fuzzy c-means, Degree of membership, Numerical simulation, Correct ratio, Incorrect ratio

1. Introduction

Soft clustering is clustering that permits belonging to more than one cluster, whereas hard clustering requires belonging to just one cluster to provide crisp classification. Fuzzy c-means (FCM) method [1, 2] is typical soft clustering, which is achieved to estimate a membership value that indicates the degree of belonging to each cluster. Since the parameters for the degree of "fuzziness" are included, it is necessary to provide a parameter value at the time of clustering. In most of the traditional research, parameter values greater than 1 and less than 2 have been used with little theoretical explanation.

In this study, we analyzed some behaviors of the degree of fuzziness by numerical simulations.

2. Fuzzy *c*-means Method

Given a finite set of n objects $\mathbf{X}=\{\mathbf{x}_1,\cdots,\mathbf{x}_n\}$ and the number of clusters c, we consider partitioning $\mathbf{X}$ to c clusters while allowing duplicate belonging. With the belonging coefficient u_{ki} (k:cluster_id, i:object_id), FCM aims to minimize the objective function $\mathrm{Err}(u,\boldsymbol{\mu})$.

$$\mathrm{Err}(u,\boldsymbol{\mu})=\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ki})^m\left\|\mathbf{x}_i-\boldsymbol{\mu}_k\right\|^2 \tag{1}$$

$$u_{ki}=\frac{1}{\sum_{j}^{c}\left(\frac{\left\|\mathbf{x}_i-\boldsymbol{\mu}_k\right\|}{\left\|\mathbf{x}_i-\boldsymbol{\mu}_j\right\|}\right)^{\frac{2}{m-1}}} \tag{2}$$

$$\boldsymbol{\mu}_k=\sum_{i}^{n}(u_{ki})^m\mathbf{x}_i \Big/ \sum_{i}^{n}(u_{ki})^m \tag{3}$$

$\boldsymbol{\mu}_k$ denotes each cluster center and m means the degree of fuzziness, with $m>1$.

The degree m corresponds to the level of cluster fuzziness. A larger m causes fuzzier clusters; in contrast, $m=1$ indicates crisp classification. We need to determine the value of m at the time of clustering, and $m=2$ has been applied in the absence of domain knowledge by convention.

3. Approach to Finding Properties of The Degree of Membership

Although we would like to find the universal or

mathematical properties of m, it is difficult to avoid the relation to data-specific characters. They may be found when considering a suitable m applied to various constructed data by designing a generation model to create various data and values to estimate an optimal m.

In order to design the data model, the following indexes, which concern cluster relationships, were picked up.

- Distance between clusters (cluster placement)
- Number of clusters
- Shape of clusters
- Number of objects in each cluster

The meaning of *cluster* in distance and shape is the set of objects to give the initial objects placement for our experiments, but it is not the target cluster. For distance, regular intervals give the typical placement, and we can arrange the number of cluster overlaps. For shape, the circle type is easy to handle because of density; however, the oval type requires consideration of bias.

The procedure for data generation and cluster assignment is as follows.

[Step.1] Decide the center vector v_i for each cluster.

$$v_i = \left(d \times \cos\left(\frac{2\pi}{c} i\right), d \times \sin\left(\frac{2\pi}{c} i\right) \right) \tag{4}$$

d: distance from the origin

[Step.2] Generate normal random numbers for each cluster with mean vector v_i and covariance matrix $\boldsymbol{E}$.

[Step.3] Calculate the coefficient p_{ik} that indicates object $\mathbf{x}_i$ belongs to cluster i.

$$p_{ik} = \frac{\frac{1}{d_{ik}+1}}{\sum_{j=1}^{c} \frac{1}{d_{jk}+1}} \tag{5}$$

d_{ik} : distance between object and cluster

[Step.4] Calculate the mean $\bar{p}_{ik}$ of the normal random numbers with mean p_{ik} and standard deviation $1/10c$ for all objects. If $\bar{p}_{ik} \geq \frac{1}{c}$ then object $\mathbf{x}_k$ is deemed to belong to cluster i.

To obtain an optimal m, we need some indexes to evaluate the FCM results. Assuming that $C_{n_i}^*$ is the number of objects belonging to cluster i in the input data, C_{n_i} is the number of objects belonging to cluster i in the results, $C_{n_i}^+$ is the number of correct objects belonging to cluster i in the results, and the correct ratio is used for overall suitability.

$$CR = \frac{\sum_{i=1}^{c} C_{n_i}^+}{\sum_{i=1}^{c} C_{n_i}^*} \tag{6}$$

In order to analyze the accuracy of each cluster, the correct ratio inside of a cluster denotes the rate by which objects should belong to it.

$$CR_i^{inside} = C_{n_i}^+ / C_{n_i}^* \tag{7}$$

The correct ratio outside of a cluster denotes the rate at which objects that do not belong to that cluster should not belong to it.

$$CR_i^{outside} = (n - C_{n_i}) / (n - C_{n_i}^*) \tag{8}$$

On the other hand, the incorrect ratio for each cluster can be the evaluation indexes in a similar manner.

$$IR_i^{inside} = (C_{n_i} - C_{n_i}^+) / C_{n_i} \tag{9}$$

$$IR_i^{outside} = (C_{n_i}^* - C_{n_i}^+) / C_{n_i}^* \tag{10}$$

4. Evaluation Experiments

According to the strategy in Section 3, we designed the evaluation experiment scheme as follows.

The results of the basic case with a regular interval and a circle shape at $c = 5, C_{n_i}^* = 50$ are shown in Fig. 1 to Fig. 6.

Table 1. Experimental conditions.

Parameters	Values
m	1.1 ~ 7
c : number of clusters	5 ~ 7
$C_{n_i}^*$	50 ~ 100
Distance between clusters	regular interval or biased placement
Shape of clusters	circle or oval

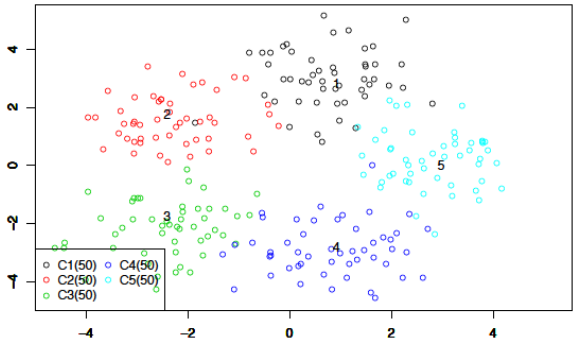


Fig. 1. A case of input data with regular placement, $c = 5, C_{n_i}^* = 50$

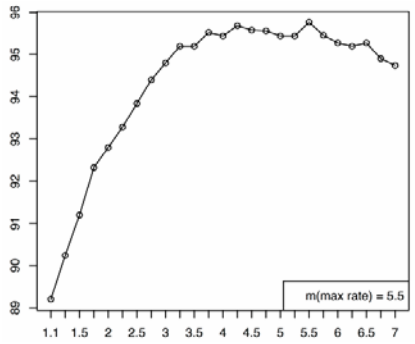


Fig. 2. CR **: correct ratio overall with regular placement,** $c = 5, C_{n_i}^* = 50$ **.**

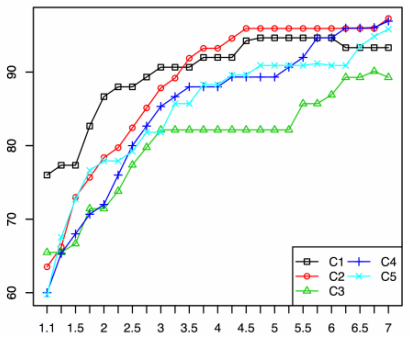


Fig. 3. CR_i^{inside} **: correct ratio inside a cluster with regular placement,** $c = 5, C_{n_i}^* = 50$ **.**

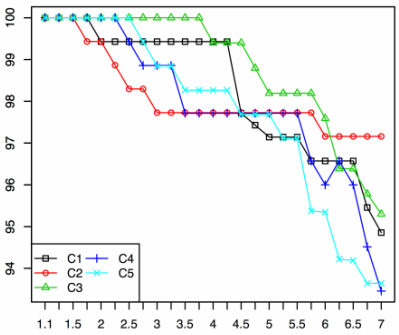


Fig. 4. $CR_i^{outside}$ **: correct ratio outside a cluster with regular placement,** $c = 5, C_{n_i}^* = 50$ **.**

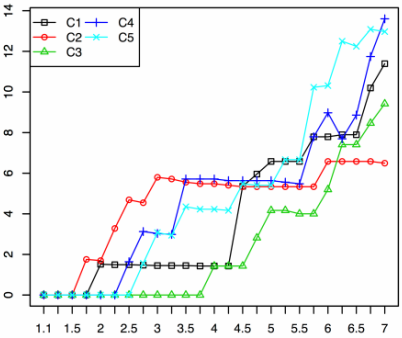


Fig. 5. IR_i^{inside} **: incorrect ratio inside a cluster with regular placement,** $c = 5, C_{n_i}^* = 50$ **.**

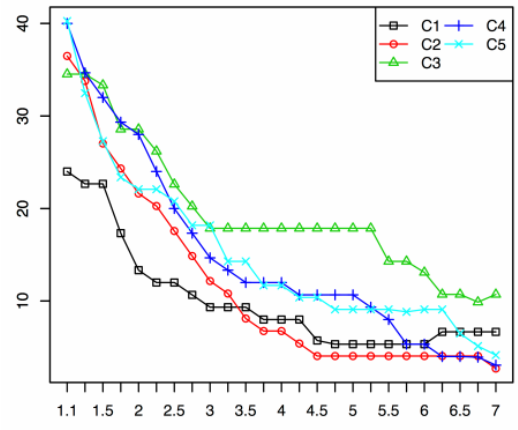


Fig. 6. $IR_i^{outside}$ **: incorrect ratio outside a cluster with regular placement,** $c = 5, C_{n_i}^* = 50$ **.**

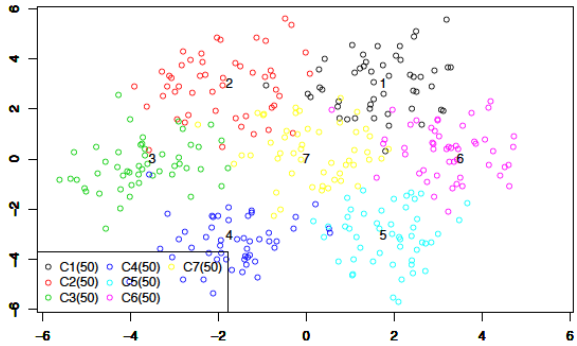


Fig. 7. A case of input data with regular placement, $c = 7, C_{n_i}^* = 50$

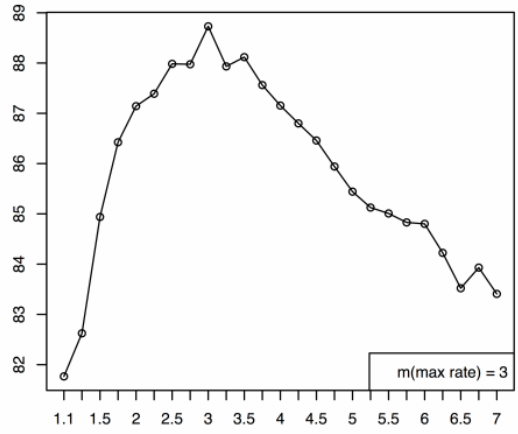


Fig. 8. CR **: correct ratio overall with regular placement,** $c = 7, C_{n_i}^* = 50$ **.**

The correct ratio overall had a peak from m = 3.5 to $m = 5.5$, and a clear inflection point could be seen around $m = 4$ for the each cluster evaluation index.

The results of the basic case with a regular interval and circle shape at $c = 7, C_{n_i}^* = 50$ are shown in Fig. 7 to Fig. 12.

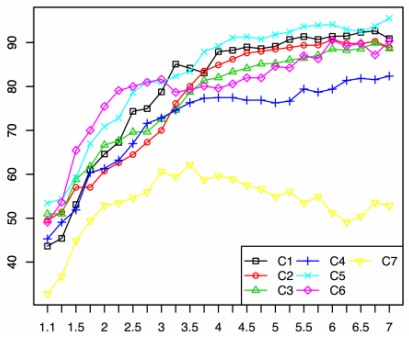


Fig. 9. CR_i^{inside} **: correct ratio inside a cluster with regular placement,** $c = 7, C_{n_i}^* = 50$ **.**

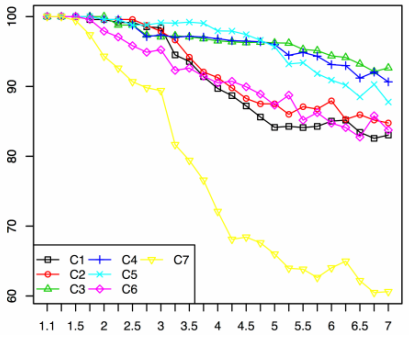


Fig. 10. $CR_i^{outside}$ **: correct ratio outside a cluster with regular placement,** $c = 7, C_{n_i}^* = 50$ **.**

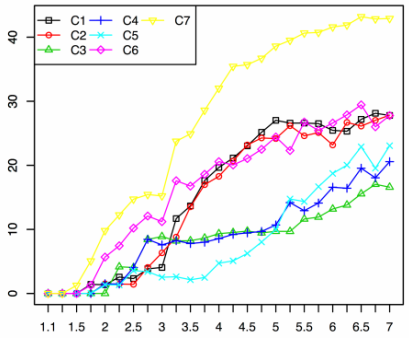


Fig. 11. IR_i^{inside} **: incorrect ratio inside a cluster with regular placement,** $c = 7, C_{n_i}^* = 50$ **.**

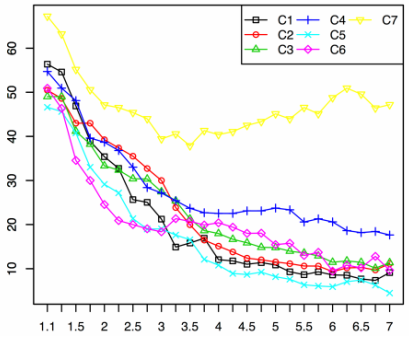


Fig. 12. $IR_i^{outside}$ **: incorrect ratio outside a cluster with regular placement,** $c = 7, C_{n_i}^* = 50$ **.**

The correct ratio overall had a peak at m = 3, and a clear inflection point could be seen around m = 3 for each cluster evaluation index. Cluster C_7 was located at the center of the objects, therefore C_7 was error prone, and its evaluation values were bad compared with the other six clusters.

The results of the modified case with biased placement and a circle shape at $c = 5, C_{n_i}^* = 100$ are shown in Fig. 13 to Fig. 18.

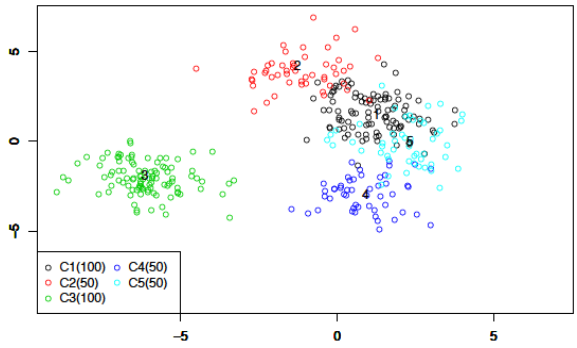


Fig. 13. A case of input data with biased placement, $c = 5, C_{n_i}^* = 100$ **.**

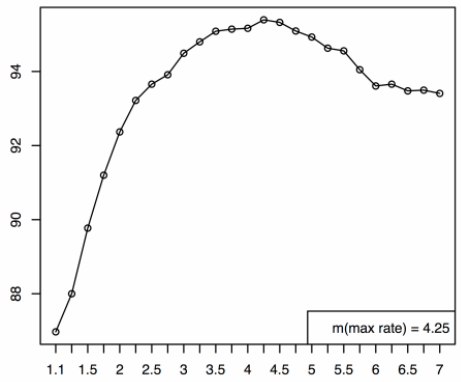


Fig. 14. CR **: correct ratio overall with biased placement,** $c = 5, C_{n_i}^* = 100$ **.**

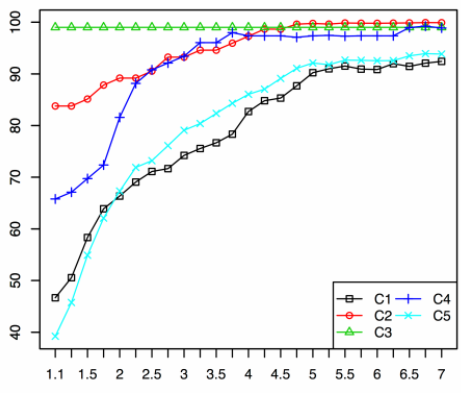


Fig. 15. CR_i^{inside} **: correct ratio inside a cluster with biased placement,** $c = 5, C_{n_i}^* = 100$ **.**

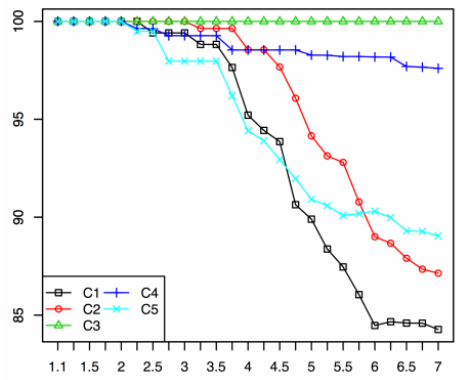


Fig. 16. $CR_i^{outside}$ **: correct ratio outside a cluster with biased placement,** $c=5, C_{n_i}^{*}=100$ **.**

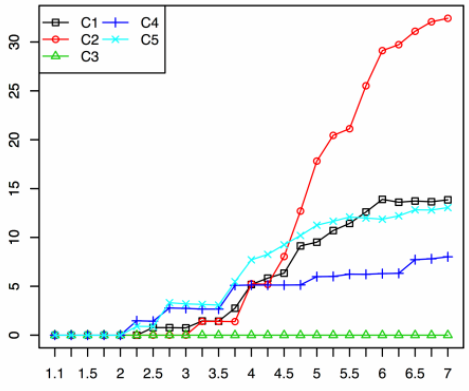


Fig. 17. IR_i^{inside} **: incorrect ratio inside a cluster with biased placement,** $c=5, C_{n_i}^{*}=100$ **.**

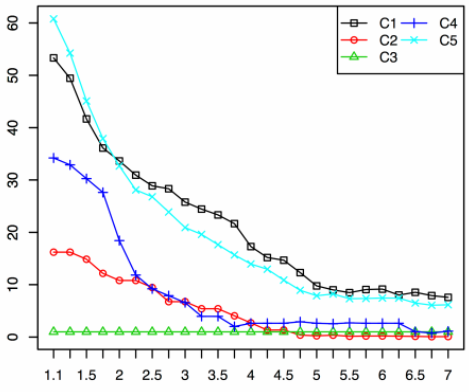


Fig. 18. $IR_i^{outside}$ **: incorrect ratio outside a cluster with biased placement,** $c=5, C_{n_i}^{*}=100$ **.**

The correct ratio overall had a peak at m = 4.25, and a inflection area could be seen from m = 3 to m = 4 for each cluster evaluation index. Cluster C_3 was located in isolation, therefore it could be distinguished stably.

The results of the modified case with biased placement and a circle shape with $c=7, C_{n_i}^{*}=100$ are shown in Fig. 19 to Fig. 24.

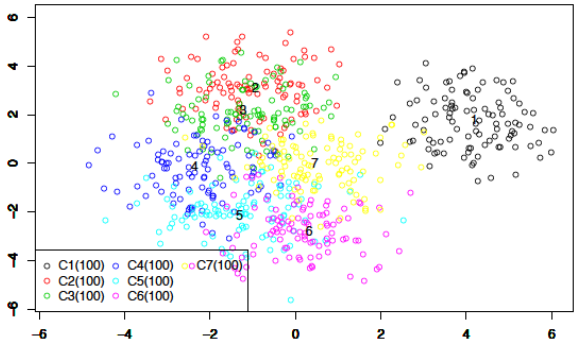


Fig. 19. A case of input data with biased placement, $c=7, C_{n_i}^{*}=100$ **.**

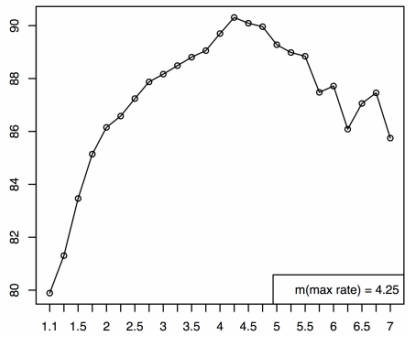


Fig. 20. CR **: correct ratio overall with biased placement,** $c=7, C_{n_i}^{*}=100$ **.**

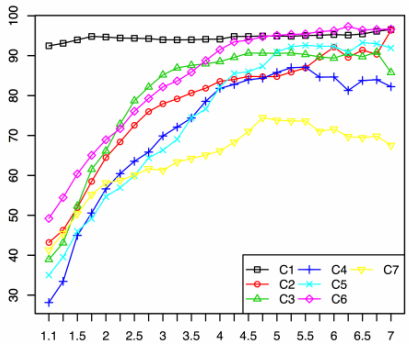


Fig. 21. CR_i^{inside} **: correct ratio inside a cluster with biased placement,** $c=7, C_{n_i}^{*}=100$

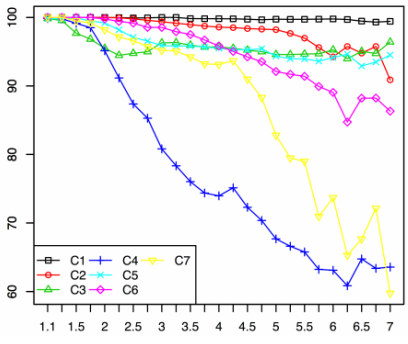


Fig. 22. $CR_i^{outside}$ **: correct ratio outside a cluster with biased placement,** $c=7, C_{n_i}^{*}=100$ **.**

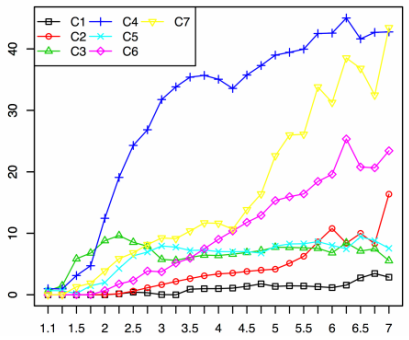


Fig. 23. IR_i^{inside} **: incorrect ratio inside a cluster with biased placement,** $c = 7, C_{n_i}^* = 100$ **.**

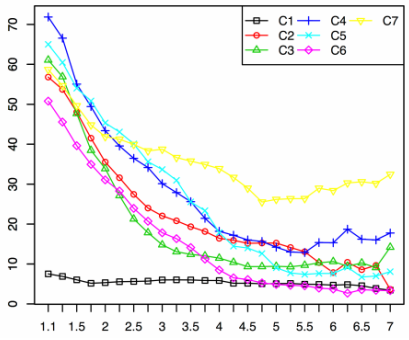


Fig. 24. $IR_i^{outside}$ **: incorrect ratio outside a cluster with biased placement,** $c = 7, C_{n_i}^* = 100$ **.**

The correct ratio overall had a peak at m = 4.25, and a clear inflection point could be seen around m = 4 for each cluster evaluation index. These values were larger than those from regular placement. Cluster C_4 and C_7 were located at the center of the objects; therefore, these were error prone. Cluster C_1 was located apart from other objects, and could be distinguished stably.

In a limited number of clusters, both regular and biased placement cases showed that the optimal m was larger. The optimal m for biased placement was larger than those for regular placement. A value of 3 or more for m was valid when the number of clusters was 7 or less.

5. Application to Motor Car Type Classification

We confirmed the validity of the experimental results through application of actual data from motor car road tests [4]. The 32 cars had 5 variables, such as fuel consumption, amount of emissions, horsepower, vehicle weight and 1/4 mile time. Because of the data description, we assumed four clusters: big sedan, midsize sedan, small sedan and sports car.

The results of FCM for the conventional $m = 2$ and proposal m = 4 are shown in Table 2 and Fig. 25. Blue lines correspond to m = 2, and red lines correspond to m = 4. Black line categories have no difference between m = 2 and m = 4.

Table 2. Comparison of clustering results.

Category	$m = 2$	$m = 4$
C1	Datsun 710 Merc 240D Merc 230 Fiat 128 Honda Civic Toyota Corolla Toyota Corona Fiat X1-9 Porsche 914-2 Lotus Europa Volvo 142E	Datsun 710 Merc 240D Merc 230 Fiat 128 Honda Civic Toyota Corolla Toyota Corona Fiat X1-9 Porsche 914-2 Lotus Europa Volvo 142E
C2	Hornet Sportabout Duster 360 Cadillac Fleetwood Lincoln Continental Chrysler Imperial Camaro Z28 Pontiac Firebird Ford Pantera L Maserati Bora	Hornet Sportabout Duster 360 Cadillac Fleetwood Lincoln Continental Chrysler Imperial Camaro Z28 Pontiac Firebird Ford Pantera L Maserati Bora
C3	Hornet 4 Drive Hornet Sportabout Merc 450SE Merc 450SL Merc 450SLC Dodge Challenger AMC Javelin Camaro Z28 Ford Pantera L Maserati Bora	Hornet 4 Drive Hornet Sportabout **Valiant** **Duster 360** Merc 450SE Merc 450SL Merc 450SLC Dodge Challenger AMC Javelin Camaro Z28 **Pontiac Firebird** Ford Pantera L Maserati Bora
C4	Mazda RX4 Mazda RX4 Wag Hornet 4 Drive Valiant Merc 240D Merc 230 Merc 280 Merc 280C Toyota Corona Ferrari Dino Volvo 142E	Mazda RX4 Mazda RX4 Wag **Datsun 710** Hornet 4 Drive Valiant Merc 240D Merc 230 Merc 280 Merc 280C Toyota Corona **Porsche 914-2** **Lotus Europa** Ferrari Dino Volvo 142E

The cluster placement was biased, and the results for m = 4 gave a more appropriate classification because of the original car descriptions.

6. Conclusion

For typical soft clustering (FCM), the degree of membership has an important role. Parameter values greater than 1 and less than 2 have been used by convention with little theoretical explanation. We analyzed the behavior of the parameter with simulation studies. The results showed the relations between the optimum value

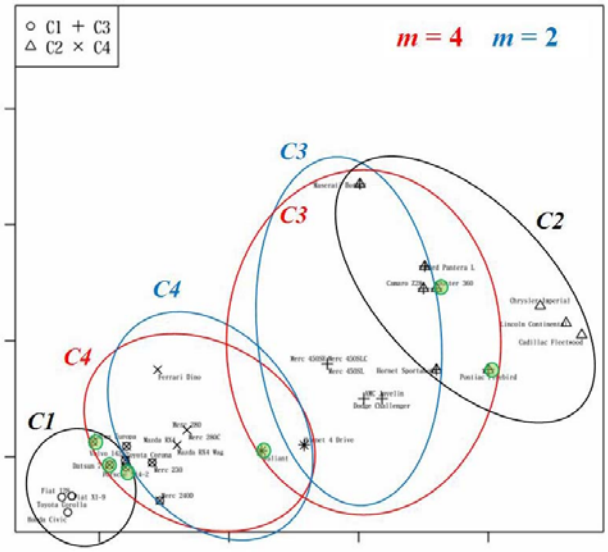


Fig. 25. Mapping of clustering results

and cluster placements or the number of clusters. We mentioned that at least a larger value than that used by convention was suitable. It is clear that a lower *m* provides a conservative decision that does not allow too much overlap among the clusters. For the correct ratio inside the cluster and the incorrect ratio outside the cluster, a smaller *m* is desirable. However a larger *m* is desirable for the correct ratio outside the cluster and the incorrect ratio inside the cluster. With judgment from a multi-faceted perspective, the optimal *m* should be larger than the conventional value.

As a future issue for research, we need to know more sophisticated features of the parameter by using a greater variety of data generation.

References

[1] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", *Journal. of Cybernetics*, Vol.3, pp.32–57, 1974. Article (CrossRef Link)

[2] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York, 1981. Article (CrossRef Link)

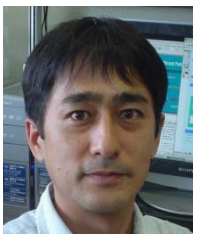
[3] S. Miyamoto, K. Umayahara and M. Mukaidono, "Fuzzy Classification Functions in the Methods of Fuzzy c-Means and Regularization by Entropy", *Journal. of Japan Society for Fuzzy Theory and Intelligent Informatics*, Vol.10, No.3, pp.548–557, 1998. Article (CrossRef Link)

[4] H. Henderson and P. Velleman, "Building multiple regression models interactively". *Biometrics*, vol.37, pp.391–411, 1981. Article (CrossRef Link)

[5] S. Hotta and K. Urahama, "Retrieval of Videos by Fuzzy Clustering", *Image Information and Television Engineers Journals*, Vol.53, No.12, pp.1750–1755, 1999. Article (CrossRef Link)

[6] L. Bobrowski and J. C. Bezdek, "c-means clustering with the L_1 and $L\infty$ norms", *IEEE Transactions on Systems Man and Cybernetics*, Vol.21, No.3, pp.545–554, 1991. Article (CrossRef Link)

[7] R. J. Hathaway, J. C. Bezdek and W. Pedrycz, "A parametric model for fusing heterogeneous fuzzy data", *IEEE Transactions on Fuzzy Systems*, Vol.4, No.3, pp.270–281, 1996. Article (CrossRef Link)



Takeo Okazaki is Associate Professor of Information Engineering at University of the Ryukyus, Japan. He received his B.Sci. and M.Sci. degrees in Algebra and Mathematical Statistics from Kyushu University, Japan, in 1987 and 1989, respectively and his Ph.D. in Information Engineering from University of the Ryukyus in 2014. He was a research assistant at Kyushu University from 1989 to 1995. He has been an assistant professor at University of the Ryukyus since 1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis. He is a member of JSCS, IEICE, JSS, GISA, and BSJ Japan.

Ukyo Aibara received his B.Eng. degree in Information Engineering from University of the Ryukyus, Japan, in 2015. In 2013, he graduated from National Institute of Technology, Kumamoto College. He discussed the evaluation of the performance and character for the soft clustering in his graduate research and thesis. Especially he investigated a variety of applications for fuzzy c-means method, and developed an enhancement package with Statistical Language R.

Lina Setiyani is Master Course Student of Information Engineering at University of the Ryukyus, Japan. She received her B.Comp. degree in Information Engineering from Janabadra University, Yogyakarta, Indonesia, in 2007. Her graduation research thema was SMS Based Information System of Janabadra University Student Admission. Her research interest is about finding optimal solution of a problem with genetic algorithm using Java language program.