General Formulation of Rough C-Means Clustering

Seiki Ubukata, Akira Notsu, and Katsuhiro Honda

Osaka Prefecture University, Sakai, Osaka, JAPAN

Summary
Hard C-means (HCM) clustering and fuzzy C-means (FCM) clustering, a fuzzy extension of HCM, are widely used non-hierarchical clustering techniques. Rough C-means (RCM), on the other hand, is a rough set-based extension of HCM that introduces the lower and upper areas of clusters representing the positive and possible memberships of objects to the clusters, respectively. In the context of RCM clustering, the problem exists of selecting one out of two counterbalancing methods, namely, Lingras and West's RCM (LRCM) and Peters' RCM (PRCM). In this paper, we propose generalized rough C-means (GRCM) clustering by re-organizing notations of RCM and unifying LRCM and PRCM. GRCM is formulated as a hybrid model based on LRCM and PRCM. Therefore, GRCM can represent not only the conventional LRCM and PRCM, but also their intermediate mixed states by adjusting some parameters. We performed numerical experiments to compare the performances of the proposed method using various parameters. We observed the trade-off between the classification accuracy in the lower areas and the fraction of objects classified as the lower areas. Through this research, we experimentally conclude that GRCM enables to observe advantages and disadvantages of LRCM and PRCM. Furthermore, it provides good results by combining them.

Keywords:
Clustering, Rough Clustering, Hard C-Means, Rough C-Means, Rough Set Theory

1. Introduction

Hard C-means (HCM) clustering, also known as k-means clustering, is a renowned, widely used non-hierarchical clustering technique [1]. HCM assigns each object to one unique cluster using the crisp membership value. However, real-life data often include objects whose belongingness to clusters is ambiguous. To address ambiguous cluster memberships, soft computing approaches, such as fuzzy theory and rough set theory, are utilized. Fuzzy C-means (FCM) clustering was proposed as a fuzzy extension of HCM by relaxing the domain of membership values to the unit interval. It has been utilized as a flexible and robust method [2, 3].

In addition to fuzzy theory, rough set theory is a promising soft computing approach that can handle the vagueness, uncertainty, inconsistency, and incompleteness inherent in data by considering rough approximations [4, 5, 6]. Rough set theory considers the certainty and uncertainty of belongingness by introducing the lower and upper approximations that represent the positive and possible memberships in a set of interest, respectively. Rough set-based HCM is called rough C-means (RCM) clustering. In the k-means context, it is referred to as rough k-means (RKM) clustering. Lingras and West first established RCM (LRCM) by introducing the rough set theory to HCM by using the lower and upper areas, which are analogous regions of the lower and upper approximations, respectively [7]. Moreover, Peters proposed a refined version of LRCM (PRCM) by modifying the object assignment using the ratio of distances instead of the difference of distances. In addition, this approach employs a calculation method of cluster centers using the upper area instead of the boundary area [8].

Furthermore, RCM has been extended in various ways. For example, it has been combined with fuzzy theory [9, 10, 11]. The above RCM-type methods are basically formulated algorithmically by following the HCM iterative procedure without considering explicit objective functions. Meanwhile, Endo et al. investigated various types of objective functions of RCM-type methods [12].

In the context of RCM clustering, the problem exists of selecting one out of two counterbalancing methods, namely, LRCM and PRCM. In this paper, we thus propose a generalized rough C-means (GRCM) clustering method by re-organizing notations of RCM and unifying LRCM and PRCM. GRCM is formulated as a hybrid model based on LRCM and PRCM. Therefore, GRCM can represent not only the conventional LRCM and PRCM, but also their intermediate mixed states by adjusting some parameters. We conducted numerical experiments to compare the performances of the proposed method. We observed the trade-off of the classification accuracy in the lower areas and the fraction of objects classified as the lower areas.

The remainder of this paper is organized as follows. Section 2 provides preliminaries, including the conventional HCM, LRCM, and PRCM. In Section 3, a general formulation of RCM is proposed by re-organizing notations of RCM and unifying LRCM and PRCM. Section 4 provides numerical experiments and discussion of their results. Finally, Section 5 presents conclusions.

2. Preliminaries

In this section, HCM, LRCM, and PRCM are explained as conventional methods. First, symbols that appear in this

Manuscript received September 5, 2017
Manuscript revised September 20, 2017

paper are summarized as follows. Let U be a set of n objects to be classified:

$$U = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_n\}, \tag{1}$$

where each object is a point in an m-dimensional vector space:

$$\boldsymbol{x}_i = (x_{i1}, \dots, x_{ij}, \dots, x_{im}) \in \mathbb{R}^m. \tag{2}$$

Let C be the number of clusters and $\boldsymbol{b}_c$ be the representative point of cluster c called the cluster center:

$$\boldsymbol{b}_c = (b_{c1}, \dots, b_{cj}, \dots, b_{cm}) \in \mathbb{R}^m. \tag{3}$$

The distance between an object, i, and a cluster center, $\boldsymbol{b}_c$, is abbreviated as follows:

$$d_{ci} = ||\boldsymbol{x}_i - \boldsymbol{b}_c||. \tag{4}$$

The minimum distance between object i to cluster centers is denoted by:

$$d_i^{min} = \min_{1 \le l \le C} d_{li}. \tag{5}$$

2.1 Hard *C*-Means

As mentioned above, HCM is one of the most renowned non-hierarchical clustering techniques [1]. Let $u_{ci} \in \{0, 1\}$ be the crisp membership value of object i in cluster c. After randomly initializing the cluster centers, HCM iteratively updates the assignments of the objects to their nearest clusters and the cluster centers by the following rules:

$$u_{ci} = \begin{cases} 1 & \left(c = \arg\min_{1 \le l \le C} d_{li}\right), \\ 0 & (\text{otherwise}). \end{cases} \tag{6}$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci} \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}. \tag{7}$$

2.2 Lingras and West's Rough *C*-Means

HCM represents the object memberships to the clusters by using crisp membership values in the Boolean index $\{0, 1\}$. It cannot represent uncertain memberships to the clusters. Rough set theory introduces two types of memberships to a subset, $X \subseteq U$, namely, the lower and upper approximations of X, which represent the positive and possible belongingness to X, respectively [4, 5, 6]. Lingras and West firstly established LRCM as a rough extension of HCM by introducing the lower and upper areas of clusters, which are analogous regions of the lower and upper approximations, respectively [7]. LRCM is executed based on an HCM-like alternative update procedure that includes the object assignments to the approximate areas and the calculation of the cluster centers based on these areas. Let $\underline{A}_c, \overline{A}_c$, and $\hat{A}_c$ be the lower, upper, and boundary areas of cluster c. In each iteration step, HCM assigns each object to one cluster (its nearest cluster), whereas LRCM assigns it to the lower and upper areas of the cluster to deal with the certainty and uncertainty of belongingness. In the object assignment process, the set T of clusters to which object i possibly belongs, other than its nearest cluster h, is generated as follows:

$$h = \arg\min_{1 \le l \le C} d_{li}, \tag{8}$$

$$T = \{\, c \mid d_{ci} - d_{hi} \le \beta \wedge c \ne h\}, \tag{9}$$

where $\beta \ge 0$ is a threshold of the difference of distances. The memberships to the lower and upper areas of the clusters are determined in the following approach using T:

1. If $T \ne \emptyset$, then $\boldsymbol{x}_i \in \overline{A}_h$ and $\boldsymbol{x}_i \in \overline{A}_c, \forall c \in T$.

2. Otherwise, if $T = \emptyset$, $\boldsymbol{x}_i \in \underline{A}_h$ and $\boldsymbol{x}_i \in \overline{A}_h$.

Boundary area $\hat{A}_c$ is calculated by $\overline{A}_c \setminus \underline{A}_c$. LRCM calculates the cluster center, $\boldsymbol{b}_c$, by the convex combination of the centers of the lower and boundary areas of cluster c:

$$\boldsymbol{b}_c = \begin{cases} \frac{\sum_{\boldsymbol{x}_i \in \underline{A}_c} \boldsymbol{x}_i}{|\underline{A}_c|} & \left(|\hat{A}_c| = 0\right), \\ \frac{\sum_{\boldsymbol{x}_i \in \hat{A}_c} \boldsymbol{x}_i}{|\hat{A}_c|} & \left(|\underline{A}_c| = 0\right), \\ \underline{w} \frac{\sum_{\boldsymbol{x}_i \in \underline{A}_c} \boldsymbol{x}_i}{|\underline{A}_c|} + \hat{w} \frac{\sum_{\boldsymbol{x}_i \in \hat{A}_c} \boldsymbol{x}_i}{|\hat{A}_c|} & (\text{otherwise}), \end{cases} \tag{10}$$

where $\underline{w}, \hat{w} \in [0, 1], s.t.\ \underline{w} + \hat{w} = 1$ are the priority weights of the lower and boundary areas, respectively. LRCM proceeds with the iterative updates of $\underline{A}_c, \overline{A}_c$, and $\boldsymbol{b}_c$, as in HCM.

2.3 Peters' Rough *C*-Means

Peters proposed a refined version of LRCM (PRCM) by modifying the generation process of the set, T, of possible clusters, other than the nearest cluster, and the calculation method of cluster centers [8]. PRCM generates T in a similar manner as LRCM; however, it uses the ratio of

distances and its threshold $\alpha \geq 1$ instead of the difference of distances:

$$T = \{\, c \mid \frac{d_{ci}}{d_{hi}} \leq \alpha \wedge c \neq h\}. \tag{11}$$

By using the ratio of distances, the dependence on the data scale can be reduced. PRCM calculates cluster center $\boldsymbol{b}_c$ by the convex combination of the centers of the lower and upper areas:

$$\boldsymbol{b}_c = \begin{cases} \frac{\sum_{x_i \in \overline{A}_c} \boldsymbol{x}_i}{|\overline{A}_c|} & \left(|\underline{A}_c| = 0\right), \\ \underline{w} \frac{\sum_{x_i \in \underline{A}_c} \boldsymbol{x}_i}{|\underline{A}_c|} + \overline{w} \frac{\sum_{x_i \in \overline{A}_c} \boldsymbol{x}_i}{|\overline{A}_c|} & (\text{otherwise}), \end{cases} \tag{12}$$

where $\underline{w}, \overline{w} \in [0,1], s.t.\ \underline{w} + \overline{w} = 1$ are the priority weights of the lower and upper areas, respectively. Peters additionally suggested an object assignment strategy in which each lower area has at least one object. This approach provides computational stability and can avoid the case dividing in Eq. (12). In this paper, we do not consider this strategy because it is considered just an implementation technique.

3. General Formulation of Rough *C*-Means

In the context of RCM clustering, the problem exists of selecting one out of two counterbalancing methods, namely, LRCM and PRCM. In this paper, we thus propose a generalized rough *C*-means (GRCM) clustering method that re-organizes notations of RCM and unifies LRCM and PRCM. In the context of HCM-type clustering, it is more convenient to use matrix-element forms of set structures. Therefore, we introduce matrix-element forms to represent the membership to areas of interest. Let $\underline{u}_{ci}$, $\overline{u}_{ci}$, and $\hat{u}_{ci}$ be the memberships of object i to the lower, upper, and boundary areas of cluster c, respectively. We rewrite LRCM and PRCM by using the matrix-element forms to improve the prospect of RCM.

3.1 Unification of upper-area constructions

First, we consider procedures of the object assignment to the upper area. LRCM determines the membership, $\overline{u}_{ci}$, of object i to the upper area of cluster c by using the difference of distances with its threshold $\beta \geq 1$. The membership of the upper area can be derived by:

$$\overline{u}_{ci} = \begin{cases} 1 \left(d_{ci} - d_i^{min} \leq \beta\right), \\ 0 \qquad (\text{otherwise}). \end{cases} \tag{13}$$

By a transposition, this can be rewritten as:

$$\overline{u}_{ci} = \begin{cases} 1 \left(d_{ci} \leq d_i^{min} + \beta\right), \\ 0 \qquad (\text{otherwise}). \end{cases} \tag{14}$$

We can determine that the upper area is constructed so that an object is assigned not only to the nearest cluster, but also to relatively close clusters with reference to an allowable level increased by adding β. On the other hand, PRCM uses the ratio of distances with its threshold $\alpha \geq 1$:

$$\overline{u}_{ci} = \begin{cases} 1 \left(\frac{d_{ci}}{d_i^{min}} \leq \alpha\right), \\ 0 \ (\text{otherwise}). \end{cases} \tag{15}$$

Clearing the fraction, this can be rewritten as:

$$\overline{u}_{ci} = \begin{cases} 1 \left(d_{ci} \leq \alpha d_i^{min}\right), \\ 0 \quad (\text{otherwise}). \end{cases} \tag{16}$$

Eq. (16) is a stable form since it can avoid division by zero; i.e., it allows the case $d_i^{min} = 0$. Then, object i is a member of the upper area of the nearest cluster, which is at the same point with the object. Additionally, we can determine that the upper area is constructed, including relatively close clusters, with reference to an allowable level increased by multiplying α. Considering that these upper areas are obtained with reference to allowable levels increased by addition or multiplication to the minimum distance to the cluster, we propose a hybrid version of the calculation of the upper area membership by unifying Eqs. (14) and (16) as follows:

$$\overline{u}_{ci} = \begin{cases} 1 \left(d_{ci} \leq \alpha d_i^{min} + \beta\right), \\ 0 \qquad (\text{otherwise}). \end{cases} \tag{17}$$

The allowable level is increased by both $\alpha \geq 1$ and $\beta \geq 0$. This approach is a general constructing method built on LRCM and PRCM. Obviously, it becomes LRCM's construction when $\alpha = 1$, whereas it becomes PRCM's construction when $\beta = 0$. Furthermore, it approaches HCM's construction when $\alpha = 1$ and $\beta = 0$. In both LRCM and PRCM, if object i is uniquely assigned to the upper area of cluster c, it should also be a member of the lower area of the cluster. Hence, the membership $\underline{u}_{ci}$ of object i to the lower area of cluster c can be derived by:

$$\underline{u}_{ci} = \begin{cases} 1 \left(\overline{u}_{ci} = 1 \wedge \sum_{l=1}^{C} \overline{u}_{li} = 1\right), \\ 0 \qquad (\text{otherwise}). \end{cases} \tag{18}$$

This represents the detection of uniquely assigned objects to the upper area. The membership of the boundary area is

calculated by subtracting the membership of the lower area from the membership of the upper area:

$$\hat{u}_{ci} = \overline{u}_{ci} - \underline{u}_{ci}. \quad (19)$$

3.2 Unification of cluster-center calculations

Second, we consider cluster-center calculations. LRCM calculates the cluster center by the convex combination of the centers of the lower and boundary areas:

$$\boldsymbol{b}_c = \underline{w}\frac{\sum_{i=1}^n \underline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \underline{u}_{ci}} + \hat{w}\frac{\sum_{i=1}^n \hat{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \hat{u}_{ci}}, \quad (20)$$

where $\underline{w}, \hat{w} \in [0,1], s.t.\ \underline{w} + \hat{w} = 1$ are the priority weights of the lower and boundary areas, respectively. On the other hand, PRCM calculates it using the upper area instead of the boundary area:

$$\boldsymbol{b}_c = \underline{w}\frac{\sum_{i=1}^n \underline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \underline{u}_{ci}} + \overline{w}\frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}}, \quad (21)$$

where $\underline{w}, \overline{w} \in [0,1], s.t.\ \underline{w} + \overline{w} = 1$ are the priority weights of the lower and upper areas, respectively. Unifying Eqs. (20) and (21), we propose a hybrid version of the calculation of the cluster center as the convex combination of the centers of the three areas:

$$\boldsymbol{b}_c = \underline{w}\frac{\sum_{i=1}^n \underline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \underline{u}_{ci}} + \overline{w}\frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} + \hat{w}\frac{\sum_{i=1}^n \hat{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \hat{u}_{ci}}. \quad (22)$$

where $\underline{w}, \overline{w}, \hat{w} \in [0,1], s.t.\ \underline{w} + \overline{w} + \hat{w} = 1$ are the priority weights of the lower, upper, and boundary areas, respectively. If the boundary or lower area becomes empty, i.e., $\sum_{i=1}^n \hat{u}_{ci} = 0$ or $\sum_{i=1}^n \underline{u}_{ci} = 0$, we must remove the related terms and adjust the remaining weights so that their sum equals one and is the convex combination as follows, respectively:

$$\boldsymbol{b}_c = \frac{1}{\underline{w} + \overline{w}}\left(\underline{w}\frac{\sum_{i=1}^n \underline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \underline{u}_{ci}} + \overline{w}\frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}}\right) \quad \left(\sum_{i=1}^n \hat{u}_{ci} = 0\right), \quad (23)$$

$$\boldsymbol{b}_c = \frac{1}{\overline{w} + \hat{w}}\left(\overline{w}\frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} + \hat{w}\frac{\sum_{i=1}^n \hat{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \hat{u}_{ci}}\right) \quad \left(\sum_{i=1}^n \underline{u}_{ci} = 0\right). \quad (24)$$

Here, Eq. (19) implies that, if $\sum_{i=1}^n \hat{u}_{ci} = 0$, then $\underline{u}_{ci} = \overline{u}_{ci}$. Additionally, if $\sum_{i=1}^n \underline{u}_{ci} = 0$, then $\hat{u}_{ci} = \overline{u}_{ci}$. Hence, Eqs. (23) and (24) are reduced to the same equation, that is, the center of the upper area:

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} \quad \left(\sum_{i=1}^n \hat{u}_{ci} = 0\right), \quad (25)$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} \quad \left(\sum_{i=1}^n \underline{u}_{ci} = 0\right). \quad (26)$$

This fact implies that upper areas are stable and important for cluster-center calculations. The upper area is considered the most fundamental structure in RCM since it is normally non-empty and the lower and boundary areas are derived based on it.

Summarizing the above discussion, the proposed calculation can be derived as follows:

$$\boldsymbol{b}_c = \begin{cases} \dfrac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} & \left(\sum_{i=1}^n \hat{u}_{ci} = 0 \vee \sum_{i=1}^n \underline{u}_{ci} = 0\right), \\ \underline{w}\dfrac{\sum_{i=1}^n \underline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \underline{u}_{ci}} + \overline{w}\dfrac{\sum_{i=1}^n \overline{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \overline{u}_{ci}} + \hat{w}\dfrac{\sum_{i=1}^n \hat{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \hat{u}_{ci}} & (\text{otherwise}). \end{cases} \quad (27)$$

$$\underline{w}, \overline{w}, \hat{w} \geq 0, \quad (28)$$

$$\text{s.t. } \underline{w} + \overline{w} + \hat{w} = 1. \quad (29)$$

The above calculation becomes that of LRCM when $\overline{w} = 0$, whereas it becomes PRCM's calculation when $\hat{w} = 0$. Actually, the top two equations of Eq. (10) can be reduced to the center of the upper area.

Note that, in our opinion, since the incidence of the empty lower area can be regarded as an unexpected cluster disappearance, the setting of the roughness or the number of clusters should be reviewed after forced termination.

3.3 Generalized Rough *C*-Means

Finally, we propose the generalized rough *C*-means (GRCM) method by summarizing the above discussions. An algorithm of GRCM can be described as follows:

Algorithm: Generalized Rough *C*-Means

Step 1. Determine the number C of clusters, roughness parameters $\alpha \geq 1$ and $\beta \geq 0$, and priority weights $\underline{w}, \overline{w} \in [0,1], s.t.\ \underline{w} + \overline{w} \leq 1$ of the lower and upper areas, respectively.

Step 2. Initialize cluster centers by random sampling without replacement from U.

Step 3. Calculate $\overline{u}_{ci}$ by Eq. (17).

Step 4. Calculate $\underline{u}_{ci}$ by Eq. (18).

Step 5. Calculate $\hat{u}_{ci}$ by Eq. (19).

Step 6. Calculate $\boldsymbol{b}_c$ by Eq. (27).

Step 7. Repeat Steps 3 to 6 until the cluster assignments do not change.

GRCM can be regarded as a general formulation and also a hybrid model based on LRCM and PRCM. GRCM becomes LRCM when $\alpha = 1$ and $\overline{w} = 0$, whereas it becomes PRCM when $\beta = 0$ and $\hat{w} = 0$. Furthermore, it approaches HCM when $\alpha = 1$ and $\beta = 0$. In this case, the membership generator in Eq. (17) becomes $d_{ci} \leq d_i^{min}$. Moreover $d_{ci} = d_i^{min}$, i.e., the assignment only to the nearest cluster. Note

that it becomes very close to HCM; however, it is slightly different. This is because if there exist two more clusters whose distances from object i respectively match d_i^{min}, their respective memberships to the upper area simultaneously become one at the same time. That is, it allows $\sum_{l=1}^{C} \bar{u}_{li} \geq 2$. Furthermore, $\sum_{l=1}^{C} \underline{u}_{li} = 0$. On the other hand, since HCM must satisfy $\sum_{l=1}^{C} u_{li} = 1$, the tie-break strategy is typically imposed.

4. Numerical Experiments

GRCM is formulated as a hybrid model based on LRCM and PRCM. GRCM can represent not only the conventional LRCM and PRCM, but also their intermediate mixed states by adjusting the four parameters $\alpha \geq 1, \beta \geq 0$, $\underline{w}, \overline{w} \in [0,1], s.t. \underline{w} + \overline{w} \leq 1$. α and β are parameters that change the roughness of clustering. The larger α and β tend to engender the smaller lower areas and the larger upper and boundary areas. By detecting and rejecting the boundary area, the classification accuracy in the lower area is expected to be improved. On the other hand, positively classified objects are unexpectedly reduced at the same time. Therefore, there exists the trade-off between the classification accuracy in the lower areas and the fraction of objects classified as the lower areas. To assess this trade-off, we introduce two performance indicators, namely, the purity and quality. Let the purity be the classification accuracy within the lower areas of the clusters:

$$purity = \sum_{c=1}^{C} \frac{\max_{l \in L} \underline{u}_{ci} u_{li}^*}{\sum_{i=1}^{n} \underline{u}_{ci}}, \quad (30)$$

where u_{li}^* is the membership value of object i to a given class label, $l \in L$. Let the quality be the fraction of objects classified as the lower areas of the clusters:

$$quality = \sum_{c=1}^{C} \frac{\sum_{i=1}^{n} \underline{u}_{ci}}{n}. \quad (31)$$

We measure these indicators only if all lower areas are not empty. Moreover, we observe the trade-off by viewing scatter plots of the purity and quality by using the following three datasets retrieved from the UCI Machine Learning Repository [13].

1. Iris dataset: $n = 150$ objects are classified into $C = 3$ clusters.
2. Wine dataset: $n = 178$ objects are classified into $C = 3$ clusters.
3. Breast Cancer Wisconsin (BCW) dataset: $n = 683$ objects (excluding objects that contain missing values) are classified into $C = 2$ clusters.

In each dataset, all dimensions are standardized to have the average of zero and the standard deviation of one.

We comprehensively compared the effects of the patterns of the four parameters. We tested fifteen patterns of weights in combination with $\underline{w}, \overline{w} = \{0.0, 0.25, 0.5, 0.75, 1.0\}$ such that $\underline{w} + \overline{w} \leq 1$. $\widehat{w}$ is determined automatically by $1 - \overline{w} - \underline{w}$. In each pattern of weights, four patterns of $\alpha = \{1.0, 1.5, 2.0, 2.5\}$ are tested and represented by black cross marks, red circles, green triangles, and blue squares, respectively. Fixing the weights and α, the performance (the purity and the quality) was measured by shifting 100 patterns of $\beta \in [0.0, 4.0]$ (in the case of the BCW dataset, $\beta \in [0.0, 2.0]$) in the equal interval. The pair of the purity and the quality was calculated by the average values of 100 trials and plotted in the scatter plot. Figs. 1, 2, and 3 shows the results of Iris, Wine, and BCW, respectively. We found a similar tendency in all the figures as we describe below. We can grasp the trade-off by viewing the scatter plots. The larger α and β tend to engender the larger purity and smaller quality. Patterns that realize a higher purity, maintaining the high quality, are better; i.e., patterns located in the upper right corner of subfigures are better.

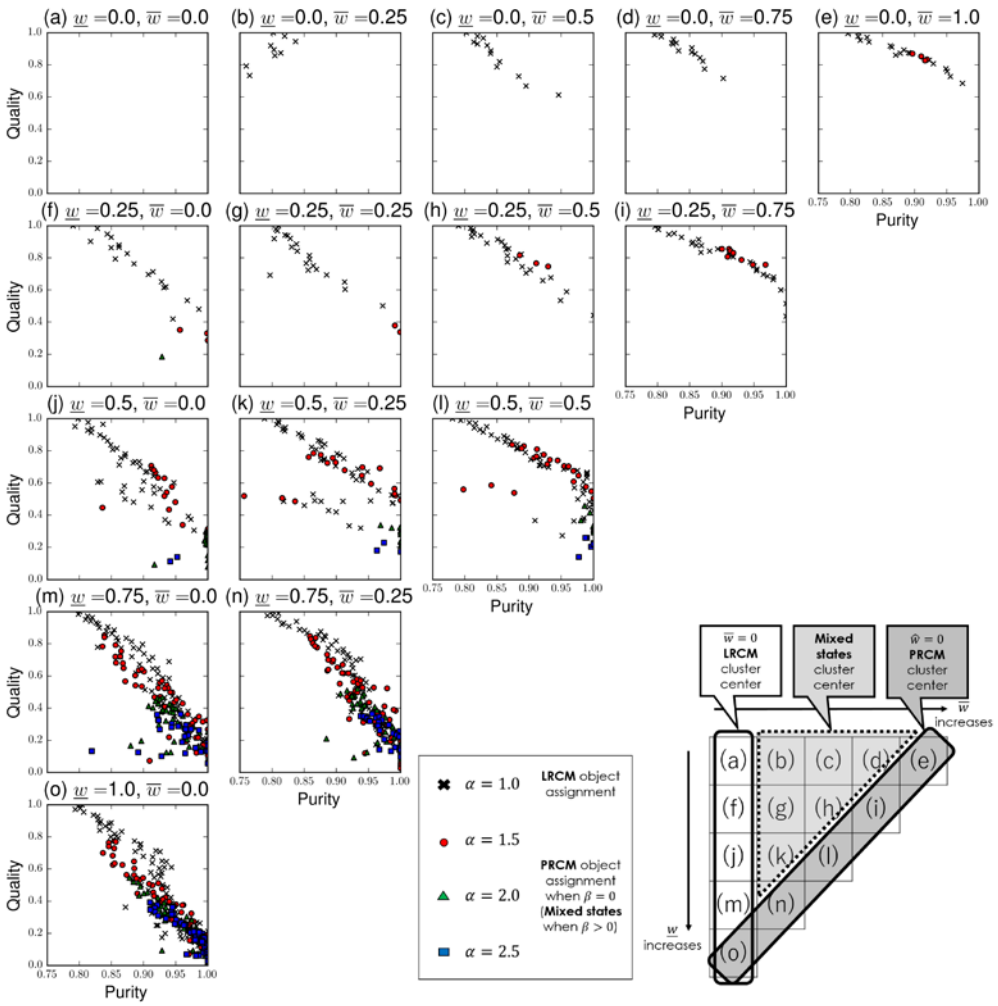


Figure 1: (Iris dataset) Comparison of the trade-off between the purity and quality with each pattern of parameters.

As shown in Figs. 1, 2, and 3, the left-most subfigures (a), (f), (j), (m), and (o) represent $\overline{w} = 0$, that is, LRCM's calculation of cluster centers. On the other hand, diagonal subfigures (e), (i), (l), (n), and (o) represent $\hat{w} = 0$, that is, PRCM's calculation of cluster centers. Proceeding from left to right, the impact $\overline{w}$ of the upper area increases and states gradually change from LRCM to PRCM. GRCM can represent intermediate mixed states between LRCM and PRCM, such as subfigures (b), (c), (d), (g), (h), and (k).

The top-left subfigure (a) represents clustering based on only the boundary region and shows no results. In the first place, the boundary areas often do not exist and are then incalculable. Even if the boundary areas exist, they tend to be located at a point greatly deviated from the original cluster centers. Furthermore, coincidences of the centers of the boundary areas of multiple clusters may occur and cause the empty lower areas and unexpected cluster disappearances. Therefore, impact $\hat{w}$ of the boundary areas

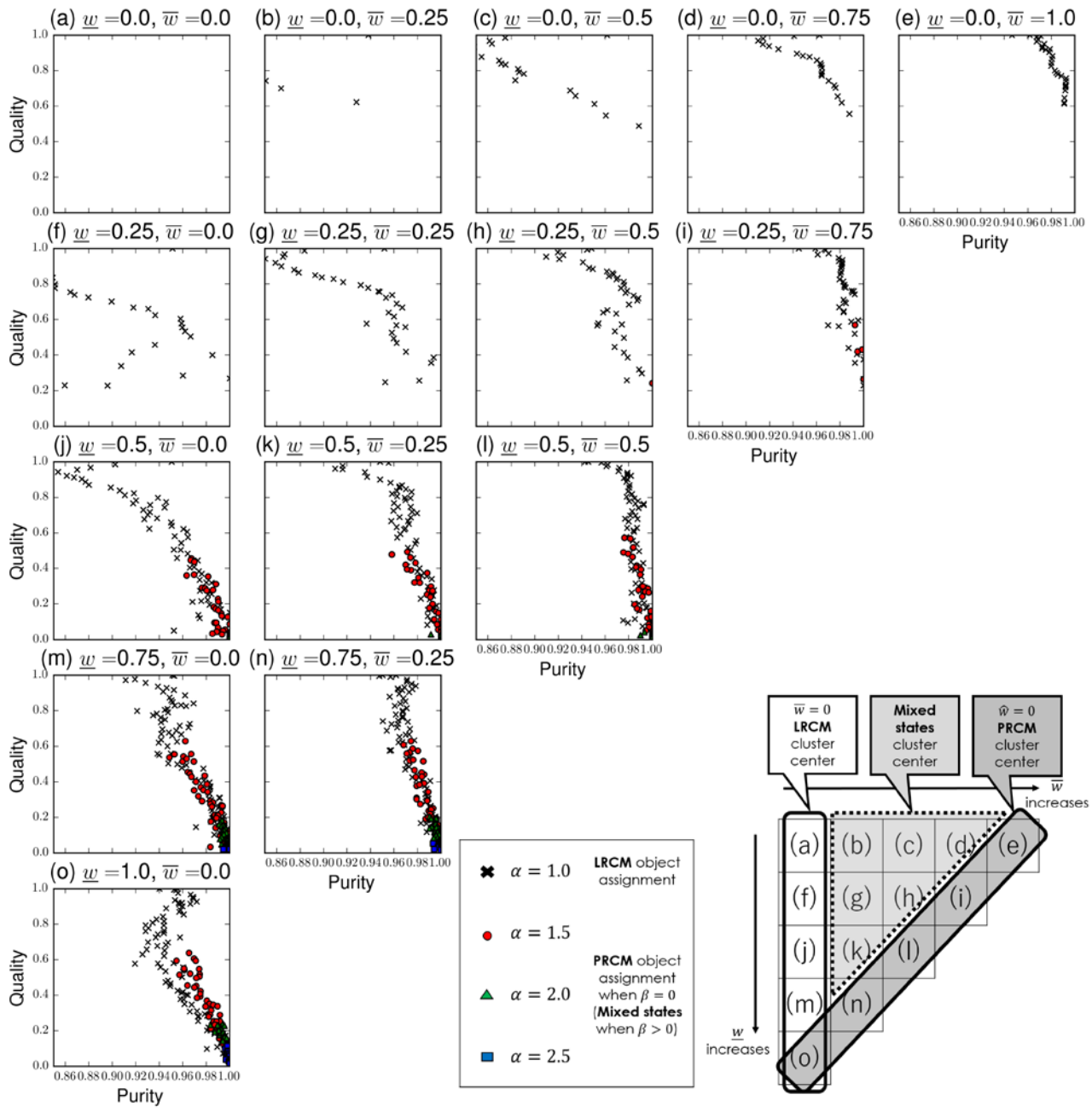


Figure 2: (Wine dataset) Comparison of the trade-off between the purity and quality with each pattern of parameters.

may cause unstable results. Owing to this instability, in the upper-left subfigures, a larger α produces no results.

Intensifying the impact $\overline{w}$ of the upper areas and approaching PRCM's calculation of the cluster centers (see the subfigures in the order of (a)-(e), (f)-(i), (j)-(l), (m)-(n), respectively), the overall performances and stability tend to improve with suppressing the variance.

As for the update of the cluster centers, we experimentally concluded that PRCM's calculation which uses the upper area is better than LRCM's calculation, which uses the boundary area. Therefore, we focused on PRCM's strategy of the cluster center and observe the results in the order of (o), (n), (l), (i), and (e). In (o), the cluster centers are calculated based on only the lower areas. In this case, the calculations of LRCM and PRCM coincide. Therefore, it is easy to compare the effect of the roughness parameters α

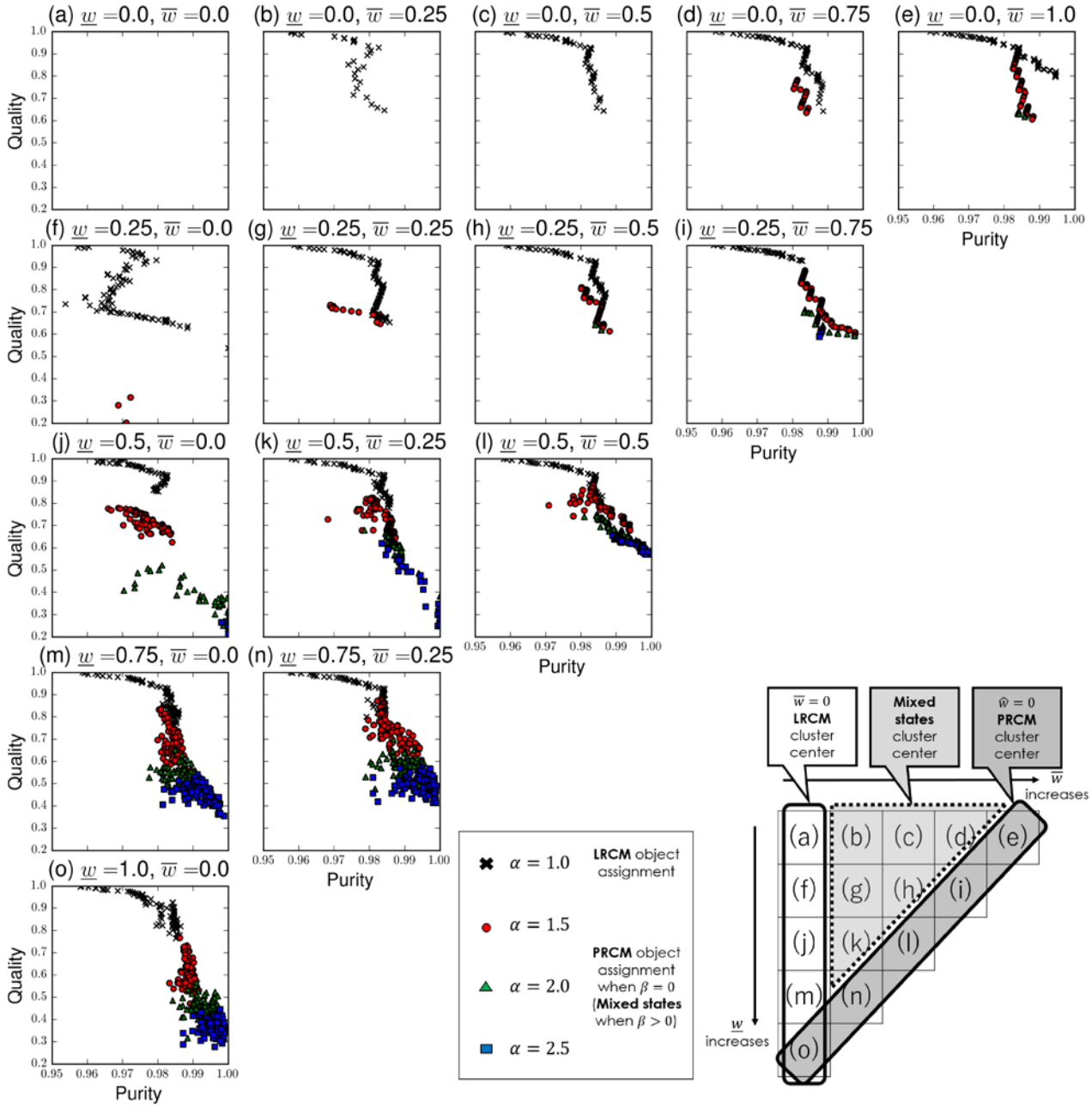


Figure 3: (BCW dataset): Comparison of the trade-off between the purity and quality with each pattern of parameters.

and β. A larger α tends to engender a greater purity and lower quality. Furthermore, in each α, such changes are promoted by increasing β. Proceeding from (o) to (e), the impact $\overline{w}$ of the upper areas increases, and the results slightly improve with suppressing the variance. On the other hand, it becomes unstable when $\alpha \geq 1.5$. Overall, smaller $\underline{w} \leq 2.5$ (upper subfigures) produces relatively unstable results if $\alpha \geq 1.5$. Moreover, smaller $\overline{w}$ (upper left subfigures) tends to produce more unstable results. In (e), which is based on the assignment by the difference and the cluster center by the lower and upper areas, $\alpha = 1$ provides fairly good results, although there is some instability in a larger α. This is a unique state in GRCM and not realized in the conventional LRCM and PRCM.

Experimentally, we concluded that PRCM's calculation of the cluster center ($\hat{w} = 0$) is better. Additionally, a larger $\overline{w} \geq 0.5$ and smaller $\alpha \leq 1.5$ tend to provide relatively stable and better results. That is, the black cross marks and red triangles in (l), (i), and (e) are relatively good. These

states are derived by GRCM as a hybrid of the LRCM assignment and PRCM cluster center. Therefore, GRCM contributed to the discovery of the combined merits of LRCM and PRCM. Furthermore, GRCM enabled observation of the robustness of the upper area and the instability of the boundary area by generating the intermediate mixed states.

5. Conclusion

In this paper, we proposed a general formulation of RCM clustering by introducing matrix-element forms of the memberships of the areas and a hybrid version of the membership assignment to the upper area, as well as the calculation of the cluster center. The proposed GRCM can represent two counterbalanced methods, namely, Lingras and West's RCM (LRCM) and Peters' RCM (PRCM) by adjusting the parameters. Furthermore, GRCM can represent their intermediate mixed state because it is a hybrid model based on the other two methods. We carried out numerical experiments to compare the performance relating to the trade-off between the purity and quality. Through this research, we conclude the following:

(1) GRCM can represent LRCM and PRCM, which are two counterbalanced RCM methods, in one unified method and thereby improve the prospect of RCM.
(2) GRCM can make it easier to observe advantages and disadvantages of LRCM and PRCM.
(3) GRCM may provide good results by combining the merits of LRCM and PRCM.

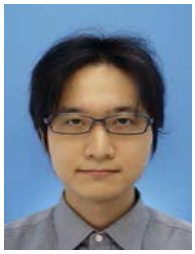
While GRCM has more expressiveness for RCM principles, we must determine additional parameters. For future work, we intend to investigate the automatic determination of appropriate parameters in the given context.

Acknowledgments

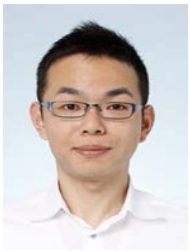
This work was partly supported by JSPS KAKENHI (Grant No. JP17K12753) and the Program to Disseminate Tenure Tracking System, MEXT, Japan.

References

[1] J. B. MacQueen, "Some Methods of Classification and Analysis of Multivariate Observations," in Proc. 5th Berkeley Symp. Math. Stat. Prob., pp. 281-297, 1967.
[2] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, 1981.
[3] 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.
[4] Z. Pawlak, "Rough Sets," International Journal of Computer & Information Sciences, Vol. 11, Issue 5, pp. 341-356, 1982.
[5] Z. Pawlak, "Rough Classification," International Journal of Man-Machine Studies, Vol. 20, Issue 5, pp. 469-483, 1984.
[6] Z. Pawlak, "Rough Set Approach to Knowledge-Based Decision Support," European Journal of Operational Research, Vol. 99, Issue 1, pp. 48-57, 1997.
[7] P. Lingras and C. West, "Interval Set Clustering of Web Users with Rough K-Means," Journal of Intelligent Information Systems, Vol. 23, Issue 1, pp. 5-16, 2004.
[8] G. Peters, "Some Refinements of Rough k-Means Clustering," Pattern Recognition, Vol. 39, Issue 8, August 2006, pp. 1481-1491.
[9] S. Mitra, H. Banka, and W. Pedrycz, "Rough-Fuzzy Collaborative Clustering," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), Vol. 36, Issue 4, pp. 795-805, 2006.
[10] S. Mitra and B. Barman, "Rough-Fuzzy Clustering: An Application to Medical Imagery," Rough Sets and Knowledge Technology, Vol. 5009 of the series Lecture Notes in Computer Science, pp. 300-307, 2008.
[11] P. Maji and S. K. Pal, "RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets," Fundamenta Informaticae 80 pp. 475-496, 2007.
[12] Y. Endo and N. Kinoshita, "Various Types of Objective-Based Rough Clustering," Fuzzy Sets, Rough Sets, Multisets and Clustering, Vol. 671 of the Series Studies in Computational Intelligence, pp. 63-85, 2017.
[13] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/.



Seiki Ubukata received B.E., M.I.S., and Ph.D. degrees in Information Science from Hokkaido University in 2007, 2009, and 2014, respectively. From 2014 to 2015, he was an Assistant Professor at Osaka University. Since 2015, he has been an assistant professor in the Department of Computer Science and Intelligent Systems at Osaka Prefecture University. His research interests include fuzzy clustering, data mining, rough set theory and agent-based simulation.



Akira Notsu received B.E., M.I., and Ph.D. degrees in Informatics from Kyoto University in 2000, 2002, and 2005, respectively. From 2005 to 2016, he was a research associate, assistant professor, and associate professor in the Department of Computer Science and Intelligent Systems at Osaka Prefecture University. From 2016, he has been an associate professor in the Graduate School of Humanities and Sustainable System Sciences at Osaka Prefecture University. His research interests include agent-based social simulation, communication networks, game theory, human-machine interface, and cognitive engineering.

Katsuhiro Honda received B.E., M.E., and Ph.D. degrees in Industrial Engineering from Osaka Prefecture University, Osaka, Japan in 1997, 1999, and 2004, respectively. From 1999 to 2013, he was a research associate, assistant professor and associate professor at Osaka Prefecture University, where he is currently a professor in the Department of Computer Science and Intelligent Systems. His research interests include hybrid techniques of fuzzy clustering and multivariate analysis, data mining with fuzzy data analysis, and neural networks.