Clustering Fusion Method Based on Multi-granularity Rough Sets: Postprint

Authors: Yu Peiqiu, Li Jinjin, Lin Guoping

Date: 2018-06-19T00:00:00+00:00

Abstract

Existing clustering ensemble algorithms approach from the perspective of clustering members; if all clustering members are utilized, the ensemble result is susceptible to influence from low-quality members, whereas selecting clustering members prior to ensemble introduces subjectivity into the selection strategy. To circumvent these two limitations to a certain extent, a novel clustering ensemble method can be proposed from the perspective of elements. By employing multi-granularity decision-inconsistent rough sets to select a portion of elements with determined categories, and subsequently utilizing these elements for clustering ensemble to generate a new partition, the multi-granularity decision-inconsistent rough set model can characterize the phenomenon where attributes are consistent while decisions are inconsistent during multi-granularity decision-making processes. A rough set model based on multi-granularity decision inconsistency is proposed, and a clustering ensemble method is presented. The specific procedure is as follows: First, the K-means clustering algorithm is applied multiple times on the dataset to generate multiple granular structures on the universe; second, the inter-granule inclusion degree is calculated between all pairs of granular structures to establish an inclusion degree matrix, the Otsu algorithm is applied to the matrix to compute a threshold, yielding multiple groups of information granules that satisfy the threshold condition, and the lower and upper approximations under multi-granularity decision inconsistency are solved; finally, the categories of elements in the lower approximation and boundary region are processed separately, thereby obtaining a fused clustering partition. Experimental results demonstrate that the proposed method can effectively improve clustering results, possesses high time efficiency, and exhibits good robustness.

Full Text

Preamble

Clustering Ensemble Algorithm Based on Multi-Granulation Rough Set

Yu Peiqiu[1,2], Li Jinjin[1]†, Lin Guoping[1,2]
(1. School of Mathematics & Statistics, Minnan Normal University, Zhangzhou, Fujian 363000, China;
2. Laboratory of Granular Computing, Zhangzhou, Fujian 363000, China)

Abstract: Existing clustering ensemble algorithms operate from the perspective of cluster members. Using all cluster members causes the ensemble result to be affected by inferior members, while selecting cluster members for fusion introduces subjectivity into the selection strategy. To avoid these limitations to some extent, this paper proposes a novel clustering fusion method from the perspective of elements. By employing multi-granulation rough sets with incongruous decisions, we select a portion of elements with determinate classes and utilize these elements to generate a new partition through clustering fusion. The multi-granulation rough set model with incongruous decisions can characterize phenomena where attributes are consistent but decisions differ during multi-granulation decision processes. This paper proposes a multi-granulation rough set model based on incongruous decisions and presents a corresponding clustering fusion method. Specifically, we first apply the K-means clustering algorithm multiple times on a dataset to generate multiple granulation structures on the universe. Next, we calculate pairwise inclusion degrees among all granulation structures to construct an inclusion degree matrix, apply Otsu' s algorithm to compute a threshold, identify multiple groups of information granules satisfying the threshold condition, and solve the lower and upper approximations under multi-granulation decision inconsistency. Finally, we process the classes of elements in the lower approximation and boundary region separately to obtain a fused clustering partition. Experimental results demonstrate that this method can effectively improve clustering outcomes, achieves high time efficiency, and exhibits good robustness.

Keywords: multi-granulation rough set; clustering ensemble; Otsu' s method; inclusion degree

---

0 Introduction

Clustering ensemble is a powerful tool that can significantly enhance the robustness and stability of unsupervised classification methods. Classical multi-granulation clustering analysis represents an important approach in exploratory data analysis, particularly in data mining and knowledge discovery, for revealing the true distribution of data. Rough set models determine different partitions

based on subsets of attribute sets, thereby forming multiple granulations. However, these models do not consider cases where attribute sets are identical while decisions differ. This paper proposes a multi-granulation decision-inconsistent rough set model that characterizes phenomena where attribute sets are identical but decisions differ, thereby enriching and developing multi-granulation rough set theory.

During the process of generating partitions using clustering algorithms, inconsistent class labels from clustering algorithms frequently occur. This situation constitutes a special case of the multi-granulation decision-inconsistent rough set model, which can be applied to clustering fusion. Clustering fusion is a fusion strategy based on clustering analysis results. Drawing on multi-granulation rough set theory, we define a multi-granulation decision-inconsistent rough set model.

For clustering fusion, some scholars adopt the analytical logic of fusing all existing clustering results. For instance, Li Feijiang et al. proposed a clustering fusion method combining rough sets and evidence theory, while Fred established a co-occurrence matrix based on similarity between data points and determined whether two points belong to the same class in the clustering result by setting thresholds. Additionally, Srehl and Ghosh proposed three hypergraph-based methods: MCLA, HGPA, and CSPA. These methods fuse all clustering results and cannot avoid the impact of inferior clustering members on fusion quality. Other scholars first evaluate clustering members, eliminate inferior ones, and then perform fusion. For example, Faceli et al. obtained optimal fusion results through genetic algorithm iteration; Hong et al. improved final clustering fusion quality by first selecting clustering members; and Yang et al. proposed a multi-granulation clustering fusion weighted iteration model based on rough set theory. While these methods select clustering members, the evaluation and selection process introduces strong subjectivity, causing certain biases in clustering fusion results. Using the multi-granulation decision-inconsistent rough set model to solve for the lower approximation of true clusters and determining boundary region element classes based on lower approximation element classes can mitigate such biases to some extent.

Regardless of clustering member quality, inferior and superior members can reach consensus on the classification of certain elements for the same true cluster. Leveraging the characteristic of multi-granulation rough sets to "seek common ground while reserving differences," this paper attempts for the first time to apply the method of computing multi-granulation lower approximations to identify consensus elements between inferior and superior clustering members. By running the K-means clustering algorithm multiple times in a complete information system to generate multiple partitions (i.e., multiple granulations), each clustering member in a partition is treated as an equivalence class. Using multi-granulation fusion methods, we compute the lower and upper approximations of these clustering members and determine boundary region element classification by examining relationships between lower approximation and boundary region

elements. Following the fundamental clustering principle of "large inter-class differences and small intra-class differences," we classify elements by finding the minimum average distance between an element and its nearest elements in each lower approximation, thereby reconstructing the partition.

---

1 Multi-Granulation Decision-Inconsistent Rough Set

In real-world decision-making, decisions made by experts involve subjectivity—different experts may provide different decisions based on the same conditions. This phenomenon is termed multi-granulation decision inconsistency in this paper. We first define the concept of a multi-granulation decision-inconsistent information system.

Definition 1 (Multi-Granulation Decision-Inconsistent Information System). Let $MS = \{IS_i | IS_i = (U, AT, f_i)\}(i \leq m)$ be a multi-granulation information system, where $IS_i = (U, AT, f_i)$ is a ternary information system, $U = \{x_1, x_2, \cdots, x_n\}$ is a non-empty finite universe; $AT = \{a_1, a_2, \cdots, a_{|AT|}\}$ is an attribute set; $f_i : U \times AT \longrightarrow V_c$ is a decision function, and V_c is a decision index set, i.e., $\forall x \in U$, $f(x, AT) \in V_c$. If $\exists x \in U$ such that when $1 \leq r \leq m$, $1 \leq s \leq m$, $f_r(x) \neq f_s(x)$, then $MIDS = \{IS_i | IS_i = (U, AT, f_i)\}(i \leq m)$ is called a multi-granulation decision-inconsistent information system.

Definition 2 (Multi-Granulation Decision-Inconsistent Rough Set). Let $MIDS = \{IS_i | IS_i = (U, AT, f_i), i = 1, 2, \cdots, m\}$ be a multi-granulation decision-inconsistent information system, $IS_i = (U, AT, f_i)$, $f_i : U \times AT \longrightarrow V_c$ is a decision function. The multi-granulation decision-inconsistent lower approximation is

$$ID\sum_{i=1}^{m} f_i(x) = \{(y \in U | f_1(x) = f_1(y) \wedge f_2(x) = f_2(y) \wedge \cdots \wedge f_m(x) = f_m(y)\}$$

The multi-granulation decision-inconsistent upper approximation is

$$ID\sum_{i=1}^{m} f_i(x) = \{y \in U | f_1(x) = f_1(y) \vee f_2(x) = f_2(y) \vee \cdots \vee f_m(x) = f_m(y)\}$$

The multi-granulation decision-inconsistent boundary is

$$BN\sum_{i=1}^{m} f_i(x) = ID\sum_{i=1}^{m} f_i(x) - ID\sum_{i=1}^{m} f_i(x)$$

Then $(ID\sum_{i=1}^{m} f_i(x), ID\sum_{i=1}^{m} f_i(x))$ is called a multi-granulation decision-inconsistent rough set.

Multi-granulation decision-inconsistent rough sets have the following properties:

1. $ID\sum_{i=1}^{m} f_i(x) \subset ID\sum_{i=1}^{m} f_i(x)$
2. $\bigcup ID\sum_{i=1}^{m} f_i(x) = U, \bigcup ID\sum_{i=1}^{m} f_i(x) = U$
3. $\forall u \in ID\sum_{i=1}^{m} f_i(x) \Leftrightarrow \forall i \leq m, f_i(u) = f_i(x) \Leftrightarrow ID\sum_{i=1}^{m} f_i(u) = ID\sum_{i=1}^{m} f_i(x)$
4. $ID\sum_{i=1}^{m} f_i(x) = \bigcap f_i(x), ID\sum_{i=1}^{m} f_i(x) = \bigcup f_i(x)$

Example 1 (Multi-Granulation Decision-Inconsistent Information System and Solution of Approximations). Let the universe $U = \{x_1, x_2, x_3, x_4, x_5\}$. Under two different granulations, there are decision tables as shown in Table 1 .

Clearly, $MIDS = \{IS_i | IS_i = (U, A, f_i), i = 1, 2\}$ is a multi-granulation decision-inconsistent information system. From Definition 1.2, we have: $ID\sum_{i=1}^{2} f_i(x_1) = ID\sum_{i=1}^{2} f_i(x_2) = \{x_1, x_2\}$, $ID\sum_{i=1}^{2} f_i(x_1) = ID\sum_{i=1}^{2} f_i(x_2) = \{x_1, x_2, x_5\}$, $BN\sum_{i=1}^{2} f_i(x_1) = BN\sum_{i=1}^{2} f_i(x_2) = \{x_5\}$.

---

2 Clustering Fusion Algorithm Based on Multi-Granulation Rough Set (MGIDA)

2.1 Noise Removal and Solution of Multi-Granulation Inconsistent Lower Approximation and Boundary Region

We first introduce the clustering algorithm used in this paper: the K-means clustering algorithm. K-means is a classical clustering algorithm still widely used today. The algorithm proceeds as follows: given the number of clusters k, select k initial cluster centers $K = \{\{x_1\}, \{x_2\}, \cdots, \{x_k\}\}$, representing k classes, where initially $K_1 = \{x_1\}, K_2 = \{x_2\}, \cdots, K_k = \{x_k\}$. Then repeat the following process until all cluster centers no longer change: (1) For each $x \in U - K$, compute the distance from x to the k cluster centers; if x is closest to cluster center K^*, then $K^* = K^* \cup \{x\}$; (2) Recompute the average value of each attribute for samples within K^* as the new cluster center for K^*. Finally, output all cluster centers and element classes, and the algorithm terminates.

Let U be a non-empty finite universe. After running a clustering algorithm multiple times on this universe, multiple partitions of U are generated. Each partition formed by a clustering algorithm run is viewed as a single granulation structure; multiple runs create multiple granulation structures, i.e., multiple granulation spaces. Using property (4) of Definition 2, we can conveniently compute the multi-granulation decision-inconsistent lower approximation in this granulation space. However, during computation, lower approximations may be generated by consistency noise. Before using the multi-granulation inconsistent

rough set model for fusion, these noises must be removed. The approach involves first computing inclusion degrees among all granules in the universe.

Definition 3 [11]. Let sets A and B be non-empty subsets of universe U. The inclusion degree between sets A and B is defined as

$$sim(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

The computed inclusion degrees are stored in an inclusion degree matrix $S(C)$.

Definition 4 (Inclusion Degree Matrix $S(C)$). Let $C = \{C_n : C_n \subseteq U\}$ be a family of subsets on universe U. Compute the compatibility degree between $C_i (i < n)$ and $C_j (j < n)$ according to Definition 2.1, and fill the compatibility degree between C_i and C_j into the i-th row and j-th column of matrix $S(C)$. The resulting matrix is the inclusion degree matrix.

Clearly, any value in $S(C)$ is between 0 and 1. After obtaining the inclusion degree matrix, Otsu' s algorithm [12] is used to compute the inclusion degree threshold. Otsu' s algorithm, also known as the Otsu method, was proposed by Japanese scholar Nobuyuki Otsu in 1979 as a thresholding method that maximizes inter-class variance between foreground and background for image segmentation. Let the mean of $S(C)_{r\times r}$ be m. There exists a threshold t that divides all elements in $S(C)$ into two classes: class A with values greater than t and class B with values less than t. Let the mean of class A be m_A and the mean of class B be m_B. Nobuyuki Otsu [12] defines the inter-class variance as

$$I_{AB} = r^2(m_A - m)^2 + r^2(m_B - m)^2$$

This method finds an optimal threshold $\hat{t}$ that minimizes misclassification probability when segmenting an image represented by gray values. This optimal threshold is obtained by traversing all possible values of t to maximize I_{AB}. This paper treats the inclusion degree matrix $S(C)$ as an image. After computing the threshold, inclusion degrees below the threshold are considered to be caused by consistency noise (treated as background). We then compute the multi-granulation decision-inconsistent lower approximation for granules with inclusion degrees above the threshold. By property (4) of Definition 2, let the information granules satisfying the threshold condition be $\{C_{s1}, C_{s2}\}$. Then $\forall x \in ID\sum_{i=1}^{m} f_i(x) \cap \bigcup C_{si}$. If there exists $ID\sum_{i=1}^{m} f_i(y) \neq \emptyset$, then merge these two lower approximations, making $ID\sum_{i=1}^{m} f_i(x) = ID\sum_{i=1}^{m} f_i(y) = ID\sum_{i=1}^{m} f_i(x) \cup ID\sum_{i=1}^{m} f_i(y)$. In subsequent discussions, let $BN = U - \bigcup$ denote the boundary region.

Example 2 (Noise Removal and Solution of Multi-Granulation Decision-Inconsistent Lower Approximation and Boundary Region). Let the universe $U = \{x_1, x_2, x_3, x_4, x_5\}$. Running a clustering algorithm twice generates partitions $C = \{C_1, C_2\}$, where $C_1 = \{\{x_1, x_2, x_5\}, \{x_3, x_4\}\}$ and

$C_2 = \{\{x_1, x_2\}, \{x_3, x_4, x_5\}\}$. Compute the compatibility degrees and fill them into the compatibility matrix to obtain $S(C)$:

$$S(C) = \begin{bmatrix} 1 & 0 & 2/3 & 1/5 \\ 2/3 & 0 & 1 & 0 \\ 1/5 & 2/3 & 0 & 1 \end{bmatrix}$$

Using Otsu' s algorithm to compute the threshold (in MATLAB, Otsu' s algorithm is implemented as the system function `graythresh`), we obtain a threshold of 0.4314. The information granules satisfying the threshold condition are $Gr_1 = \{\{x_1, x_2, x_5\}, \{x_1, x_2\}\}$ and $Gr_2 = \{\{x_3, x_4\}, \{x_3, x_4, x_5\}\}$. Inclusion degrees not satisfying the threshold condition are caused by consistency noise and are not processed. Solving the multi-granulation decision-inconsistent lower approximation for information granules meeting the threshold condition and computing the boundary region yields:

$$ID\sum_{i=1}^{2} f_i(x_1) = ID\sum_{i=1}^{2} f_i(x_2) = \{x_1, x_2, x_5\} \cap \{x_1, x_2\} = \{x_1, x_2\}$$

$$ID\sum_{i=1}^{2} f_i(x_3) = ID\sum_{i=1}^{2} f_i(x_4) = \{x_3, x_4\} \cap \{x_3, x_4, x_5\} = \{x_3, x_4\}$$

$$BN = U - ID\sum_{i=1}^{2} f_i(x_1) \cup ID\sum_{i=1}^{2} f_i(x_3) = \{x_5\}$$

By Definition 2, elements in the lower approximation belong to the same class in every clustering process, so lower approximation elements are definitively in the same class. However, boundary elements may not belong to the same class in every clustering process, so their classes are indeterminate and require algorithmic determination.

2.2 Processing of Boundary Region Elements

To facilitate processing of boundary region elements, we first provide a definition of clustering derived from the principle of "large inter-class distances and small intra-class distances."

Definition 5. Let (U, A, f) be a complete information system. Given a distance metric $d(x, y) : U \times U \to [0, +\infty)$, clustering establishes a partition $\mathcal{C} = \{\mathcal{C}_k : k = 1, 2, \cdots, m\}$ on U such that for any $x, y \in \mathcal{C}_i$, if for any $y' \in \mathcal{C}_i$, $d(x, y) \leq d(x, y')$, then for any $z \in \mathcal{C}_j (i \neq j)$, we have $d(x, y) < d(x, z)$.

Clustering fusion ultimately aims to generate a clustering. Using the multi-granulation decision-inconsistent rough set model inevitably produces a boundary region. How to process elements in this boundary region becomes a significant issue. If we can identify relationships between elements in the multi-granulation decision-inconsistent lower approximation and those in the boundary region, we can determine boundary region element classes through lower approximation element classes. Based on Definition 5, we present the following theorem:

Theorem 1. Let (U, A, F) be a complete information system, $\mathcal{C} = \{\mathcal{C}_k : k = 1, 2, \cdots, m\}$ be a clustering on U, and $\mathcal{C}' = \{\mathcal{C}'_k : \mathcal{C}'_k \subseteq \mathcal{C}_k, k = 1, 2, \cdots, m\}$, where $\mathcal{C}'_k$ is any non-empty subset of $\mathcal{C}_k$. If $x \in U - \bigcup_{i=1}^{m} \mathcal{C}'_i$ and $d(x, y) = \min\{D(x, \mathcal{C}_k) : k = 1, 2, \cdots, m\}$ holds for any $y \in \bigcup_{k=1}^{m} \mathcal{C}'_k$, then $x \in \mathcal{C}_i$ if and only if there exists $y \in \mathcal{C}'_i$ such that for any $z \in \mathcal{C}'_j (i \neq j)$, $d(x, y) < d(x, z)$, where $D(x, X) = \min\{d(x, t) : t \in X\}$.

Proof.
1) Sufficiency: If $d(x, y) = \min\{D(x, \mathcal{C}_k) : k \leq m\}$ and $y \in \bigcup_{k=1}^{m} \mathcal{C}'_k$, then $\exists \mathcal{C}_i$ such that $d(x, y) = D(x, \mathcal{C}_i)$. Since $\mathcal{C}'_i \subseteq \mathcal{C}_i$, we have $\forall z \in U - \mathcal{C}'_i$, $d(x, y) < d(x, z)$. Consequently, $\forall z \in U - \mathcal{C}_i \subseteq U - \mathcal{C}'_i$, $d(x, y) < d(x, z)$, which implies $x \in \mathcal{C}_i$.

2) Necessity: If $x \in \mathcal{C}_i$, $x \in U - \bigcup_{k=1}^{m} \mathcal{C}'_k$, and $d(x, y) = \min\{D(x, \mathcal{C}_k) : k = 1, 2, \cdots, m\}$ holds for any $y \in \bigcup_{k=1}^{m} \mathcal{C}'_k$, then $y \in \mathcal{C}'_i$ and $\forall z \in \mathcal{C}'_j (i \neq j)$, $d(x, y) < d(x, z)$. Otherwise, if $y \in \mathcal{C}'_j (i \neq j)$, there would exist $t \in \mathcal{C}'_i$ satisfying $d(x, t) = \min\{d(x, r) | r \in \mathcal{C}'_i\}$ such that $d(x, y) < d(x, t)$, which would imply $x \in \mathcal{C}_j$, leading to a contradiction.

Explanation: Theorem 1 provides a method for processing boundary region elements: the element with the minimum distance to all lower approximations (the smallest $D(x, \mathcal{C}_k)$, $k \leq m$) must belong to the lower approximation closest to it. We can gradually reduce the boundary region by finding the boundary element x closest to all lower approximations and merging it into the nearest lower approximation. Due to the complexity of real datasets, a more reliable approach is to replace the distance from x to the nearest element in a lower approximation with the average distance from x to the nearest N_0 elements (where N_0 is the number of elements) in that lower approximation. By comparing the minimum distances from x to all lower approximations, we can determine the 归属 of each element in the boundary region. In this paper, we set $N_0 = \min\{|\mathcal{C}_k| | k = 1, 2, \cdots, m\}$. Repeating this process until all elements in the boundary region are merged into lower approximations yields a new partition with no indeterminate elements.

Example 3 (Processing of Boundary Region Elements). Continuing from Example 2, $BN = U - ID\sum_{i=1}^{2} f_i(x_1) \cup ID\sum_{i=1}^{2} f_i(x_3) = \{x_5\}$. Let $N_0 = \min\{2, 2\} + 1 = 2$. Let Gr_1 represent the lower approximation obtained from Gr_1 and Gr_2 represent the lower approximation obtained from Gr_2. If $d(x_5, x_1) + d(x_5, x_2) < d(x_5, x_3) + d(x_5, x_4)$, then $Gr_1 = Gr_1 \cup \{x_5\}$; otherwise,

$Gr_2 = Gr_2 \cup \{x_5\}$.

Based on the above discussion, we present the clustering fusion algorithm based on multi-granulation decision-inconsistent rough sets:

Algorithm 1 (Clustering Fusion Algorithm Based on Multi-Granulation Decision-Inconsistent Rough Set, MGIDA).
Input: A family of partitions generated by running multiple clustering algorithms multiple times.
Output: A fused clustering partition.

Step 1: Compute the compatibility degree matrix.
Step 2: Use Otsu' s algorithm to compute the threshold for the compatibility degree matrix. Compatibility degrees between clustering members greater than the threshold form an information granule satisfying the threshold condition; find all such information granules.
Step 3: Use Definition 2 to solve for the lower approximation and compute the boundary region BN.
Step 4: Take $x \in BN$ satisfying $d(x,y) = \min\{D(x, ID\sum_{i=1}^{m} f_i(x)) : k = 1,2,\cdots,m; x \in U\}$, and reclassify x using the method described in Theorem 1.
Step 5: $BN \leftarrow BN - \{x\}$; when $BN \neq \emptyset$, return to Step 4.
Step 6: Output the classes of all elements.

---

3 Experimental Results and Analysis

To verify the algorithm' s effectiveness, we conduct experiments on 10 datasets. The dataset information is shown in Table 3 . To generate different weak partitions from the same dataset, we use Algorithm 2 to process the datasets. The specific process is illustrated in Example 4.

Example 4 (Generating Different Weak Partitions Using the Same Dataset). Given an information system as shown in Table 2 , generate two 1-dimensional random vectors with modulus 1: $r_1 = \langle 0.5030, 0.8406, 0.2007\rangle^T$, $r_2 = \langle 0.0979, 0.6985, 0.7089\rangle^T$. Compute $U \cdot r_1$ and $U \cdot r_2$ respectively, where '$\cdot$' represents matrix multiplication, to obtain:

$$Ur_1 = \langle 0.6967, 0.4586, 0.8946, 1.3101\rangle^T, \quad Ur_2 = \langle 0.6454, 0.4254, 0.8774, 0.9974\rangle^T$$

Applying K-means clustering to Ur_1 and Ur_2 yields two different weak partitions: $C_1 = \{\{x_1, x_2, x_3\}, \{x_4\}\}$ and $C_2 = \{\{x_1, x_2\}, \{x_3, x_4\}\}$. Thus, different weak partitions are generated from the same dataset.

Algorithm 2 [5] (Weak Partition Generation).
1. Generate a random d-dimensional vector u with $|u| = 1$.
2. $X' = X_{n\times d} \cdot u_{d\times 1}$.
3. $A_m \leftarrow KMeans(X')(m < n)$.

We use K-means to cluster the processed dataset and compare our fusion results with true clusters by computing clustering accuracy [11], defined as:

$$AC = \frac{\sum_{i=1}^{k} \max_{j=1,2,\cdots,k} n_{ij}}{|U|}$$

where if the true clustering partition is $CR = \{C_1, C_2, \cdots, C_k\}$ and the clustering fusion result is $CF = \{F_1, F_2, \cdots, F_k\}$, then $n_{ij} = |C_i \cap F_j|$, $i, j \leq k$.

Four weak partitions are generated each time for clustering fusion. We compare the average clustering accuracy and variance over 100 trials with CSPA, HGPA, MCLA (all graph-based clustering fusion algorithms), IWCE [13] (a clustering fusion weighted iteration model based on rough set theory), and DSCE [5] (a multigranulation information fusion clustering ensemble method based on evidence theory). Using clustering accuracy as the evaluation metric, we obtain the results shown in Table 4 and Figure 1 [Figure 1: see original paper].

The comparison shows that MGIDA achieves the best clustering accuracy on datasets 3, 6, and 10; suboptimal accuracy or very close results on datasets 2, 4, 5, 7, 8, and 9 (differences: -0.0113 on dataset 4, -0.0323 on dataset 5, -0.0023 on dataset 7); and clearly outperforms HGPA. MGIDA is inferior to MCLA only on dataset 3, superior to MCLA on datasets 2, 4, 7, 8, and 9, and inferior on other datasets with small accuracy gaps. MGIDA is inferior to IWCE only on datasets 4 and 5, and inferior to DSCE only on dataset 9. Overall, the proposed algorithm is never the worst-performing algorithm on any dataset. Algorithm 2 generates different weak partitions from the same data by essentially adding varying degrees of noise through data distortion. Therefore, algorithms performing well on such noisy datasets exhibit good robustness. Since our algorithm achieves optimal or near-suboptimal accuracy on all datasets, it demonstrates good performance and robustness across datasets.

Algorithm 2 randomly maps multi-dimensional data to one-dimensional space, causing some algorithms to perform averagely on certain datasets and poorly on datasets with complex distributions. This occurs because data distortion prevents generated clustering members from properly reflecting the true data distribution.

Time Efficiency. The time efficiency is shown in Table 5 and Figure 2 [Figure 2: see original paper]. Let each run generate h clustering members, each with K clusters containing p elements. After computing the boundary using multi-granulation rough sets, $|BN| = |U| - n$. The complexity of computing the compatibility degree matrix in Step 1 is $(hKp)^2$. The time complexity of Otsu' s algorithm in Step 2 is $|U| \log |U|$. The comparison count for solving lower approximation and boundary in Step 3 is $(Kp)^2$. The time complexity for reclassifying boundary elements in Step 4 is no greater than $|U|(|BN|!)$. Therefore, the total time complexity of MGIDA is $O((hKp)^2) + |U| \log |U| + (Kp)^2 + |U|(|BN|!)$. This shows that universe size and number of clusters are the most important

factors affecting runtime, and the algorithm has lower time complexity when the boundary region is small. Table 5 and Figure 2 show that MGIDA achieves the best time efficiency on 4 datasets and suboptimal efficiency on all datasets except dataset 1. The algorithm has relatively small time complexity. Notably, MGIDA has high time efficiency on small datasets, with average performance on larger datasets.

4 Conclusion

This paper first proposes a multi-granulation decision-inconsistent rough set model and subsequently presents a clustering fusion algorithm based on multi-granulation rough sets. The research investigates clustering fusion algorithms from a new perspective. Data experiments were conducted to compare the algorithm with other clustering fusion methods, using clustering accuracy as the metric to verify its effectiveness. Experimental results show that the new clustering fusion algorithm can achieve optimal accuracy on some datasets and suboptimal or near-suboptimal accuracy when not optimal. The algorithm demonstrates good robustness and time efficiency advantages on small datasets.

The K-means algorithm performs poorly on non-convex distributed data. However, based on Theorem 1, the proposed clustering fusion algorithm is less affected by data distribution for non-convex data. Improving the algorithm for application to non-convex distributed data represents a worthwhile research direction.

References

[1] Qian Yuhua, Liang Jiye, Yao Yiyu, et al. MGRS: a multi-granulation rough set [J]. Information Sciences, 2010, 180 (6): 949-970.
[2] Lin Guoping, Liang Jiye, Qian Yuhua. An information fusion approach by combining multigranulation rough sets and evidence model [J]. Information sciences, 2015, 314: 184-199.
[3] Antoine C, Cédric W, Pierre G, et al. Collaborative clustering: why, when, what and how [J]. Information Fusion, 2018, 39: 81-95.
[4] 阳琳赟, 王文渊. 聚类融合方法综述 [J]. 计算机应用研究, 2005, 22 (12): 8-10. (Yang Linbin, Wang Wenyuan. A survey of clustering fusion methods [J]. Application Research of Computers, 2005, 22 (12): 8-10.)
[5] Li Feijiang, Qian Yuhua, Wang Jieting, et al. Multigranulation information fusion: a Dempster-Shafer evidence theory-based clustering ensemble method [J]. Information Sciences, 2017, 378: 389-409.
[6] 谢岳山, 樊晓平, 廖志芳, 等. 一种基于图论的加权聚类融合算法 [J]. 计算机应用研究, 2013, 30 (4): 1015-1016. (Xie Yueshan, Fan Xiaoping, Liao Zhifang, et al. A graph-based weighted clustering ensemble algorithm [J]. Application Research

of Computers, 2013, 30 (4): 1015-1016.)
[7] Fred A. Finding consistent clusters in data partitions [C]// Proc of International Workshop on Multiple Classifier Systems. Berlin: Springer, 2001: 309-318.
[8] Ayad H, Kamel M. Finding natural clusters using multi-cluster combiner based on shared nearest neighbors [C]// Proc of International Workshop on Multiple Classifier Systems. Berlin: Springer, 2003: 166-175.
[9] Faceli K, De Souto M C P, De Araújo D S A, et al. Multi-objective clustering ensemble for gene expression data analysis [J]. Neurocomputing, 2009, 72 (13): 2763-2774.
[10] Yi Hong, Kwong S, Wang Hanli, et al. Resampling-based selective clustering ensembles [J]. Pattern Recognition Letters, 2009, 30 (3): 298-305.
[11] Nguyen N, Caruana R. Consensus clusterings [C]// Proc of the 7th IEEE International Conference on Data Mining. 2007: 607-612.
[12] Otsu N. A threshold selection method from gray-level histograms [J]. IEEE Trans on Systems, Man, and Cybernetics, 1979, 9 (1): 62-66.
[13] Jain A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31 (8): 651-666.
[14] 阳琳赟, 王路, 卓晴, 等. 基于粗糙集理论的聚类融合加权迭代模型 [J]. 清华大学学报: 自然科学版, 2009 (8): 1106-1108. (Yang Linbin, Wang Lu, Zhuo Qing, et al. Weighted iteration model of clustering fusion based on rough set theory [J]. Journal of Tsinghua University: Natural Science Edition, 2009 (8): 1106-1108.)
[15] Hu Jie, Li Tianrui, Wang Hongjun, et al. Hierarchical cluster ensemble model based on knowledge granulation [J]. Knowledge-Based Systems, 2016, 91 (C): 179-188.
[16] Huang Dong, Wang Changdong, Lai Jianhuang. Locally Weighted Ensemble Clustering [J]. IEEE Trans on Cybernetics, 2018, 48 (5): 1460-1473.
[17] Kausar N, Abdullah A, Samir B B, et al. Ensemble clustering algorithm with supervised classification of clinical data for early diagnosis of coronary artery disease [J]. Journal of Medical Imaging & Health Informatics, 2016, In-Press.
[18] Teng Geer, He Changheng, Xiao Jin, et al. Cluster ensemble framework based on the group method of data handling [J]. Applied Soft Computing, 2016, 43 (C): 35-46.

Note: Figure translations are in progress. See original paper for figures.