A Novel Rough Fuzzy Clustering Algorithm with A New Similarity Measurement

Yang Li[1], Jian-cong Fan[1,2,3], Jeng-Shyang Pan[1,2], Gui-han Mao[1], Geng-kun Wu[1,2]

[1] College of Computer Science and Engineering, Shandong University of Science and Technology, China
[2] Provincial Key Lab. for Information Technology of Wisdom Mining of Shandong Province, Shandong University of Science and Technology, China
[3] Provincial Experimental Teaching Demonstration Center of Computer, Shandong University of Science and Technology, China
18363976385@163.com, fanjiancong@sdust.edu.cn, jengshyangpan@gmail.com, 297003471@qq.com, 728469151@qq.com

Abstract

With the emergence of exponential growth of datasets in various fields, fuzzy theory-based approaches are widely used to improve or optimize the data clustering algorithms. These improved algorithms can achieve better results than the original counterparts in practical applications. However, the fuzzy clustering algorithms including the traditional improved algorithms normally ignore the clustering boundary uncertainty, inter-class compactness and complex data problems, thereby result in the unsatisfactory clustering results. To address this issue, in this paper, a novel rough fuzzy clustering algorithm based on a new similarity measure is proposed by utilizing the upper approximation and lower approximation of rough set. We also develop the method of transforming fuzzy clustering model into rough set model. Our experiment results show that the improved algorithm can get better clustering effect.

Keywords: Similarity measure, Fuzzy clustering algorithm, Rough set, Clustering

1 Introduction

Rough set theory, as a theoretical method for studying data expression, learning and induction, has been widely studied and applied in various fields in recent years. It has become a powerful tool for data mining [1], knowledge discovery, uncertainty reasoning, concept adaptive learning, and granular computing [2-3]. The key problem in rough set research is cluster analysis [4], which groups a set of objects having high similarity [5].

For fuzzy clustering [6-8], many effective indicators have been proposed to find the optimal number of clusters and improve the quality of fuzzy clustering. Although the improved index has better performance, the measurement of compactness still appears to be decreased monotonically when the number of clusters approach the number of samples, and the problem of cluster boundaries remains unsolved.

To address this problem, we develop a novel algorithm which combines two soft data processing methods: rough set and fuzzy set. Through the rough partition domain, the fuzzy set membership function is used to deal with the boundary region. We also implemented a new similarity measure strategy to improve the robustness of clustering for different parameter selections. Experimental results show that the improved algorithm can get better clustering results.

The rest of this paper are organized as follows. In Section 2, the rough set and fuzzy clustering algorithms are briefly reviewed. In Section 3, some important preliminary knowledge used in our proposed approaches are stated. In Section 4, we present the algorithms proposed in this paper, and some theories and analysis necessary in it. In Section 5, experimental studies are conducted to verify the effectiveness of our proposed algorithm. Section 6 concludes the paper.

2 Related Work

Zadeh proposed the separation index between clustering results [9], which is the first step in the study of effectiveness indicators. However, due to the singularity of discriminant validity, the effect is not satisfactory. Bezdek proposed the concept of Partition Coefficient (PC) [10] and Partition Entropy (PE) [11]. Dave proposed an improved partition coefficient (MPC) [12] based on the defect of partition coefficient. Kim introduced an improved validity index KYI [13] using the relative sharing degree between each pair of clusters. Nevertheless, the validity index still has some limitations, for example, it only contains the relationship between the membership degree divisions of the data set samples without the prior information of

*Corresponding Author: Jian-cong Fan; E-mail: fanjiancong@sdust.edu.cn
DOI: 10.3966/160792642019072004014

the data samples, and does not take into account the true geometric structure distribution of the data set, so it limits its application to a certain extent. The researchers then proposed a series of cluster validity indicators. Instead of using the traditional Euclidean distance, Gath et al. [14] utilized the measurement of fuzzy density and fuzzy hyper volume [15], and then proposed the FHV validity index. Xie and Beni [16] proposed an effectiveness index of separation, which used the ratio of intra-class compactness and inter-class separation to quantify the information of data membership and geometric distribution of datasets. Kwon [17] developed a penalty function which effectively limited the monotonous decreasing effect of the effectiveness index when the number of clusters c is infinitely close to the number of samples n. Linkens and Chen also pointed out the shortage of partition entropy, and modified the definition of partition entropy [18].

From view of the fuzzy clustering, a lot of validity indices have been proposed to find the optimal cluster number. However, it is not difficult to find the disadvantage of above indices which only contained membership degree of data members, but not contained the distribution information of data samples and the direct linkage information of the geometric structure of data set. Therefore, those factors limited its application. And then researchers put forward a series of cluster validity indices. Zahid and Limouri considered the fuzzy partition of data set on the basis of the geometric structure of data set and proposed the cluster validity index. Pakhira made further improvement and put forward PBMF clustering validity index [19]. Arbelaitz compared 30 cluster validity indices in many different environments with different characteristics [20]. Although index and some improved validity indices had better performance, but it would still appear monotone decreasing of compactness measurement as the cluster number approaches to sample number. The separation measurement was still limited to the geometric structure of each cluster, this was because the calculation only involves the information of cluster centroid, not takes into account the shape of all clusters.

3 Preliminary

Rough set theory is a mathematical tool to deal with uncertain and incomplete information. The classical rough set theory divides the discourse domain through the equivalence relation and uses the upper and lower approximations to describe the uncertain information. In a knowledge system, some attributes might be discrete. Although it is possible to know which collection every element belongs to by discretization and other methods, it may not be clear to estimate the degree that an element belongs to the collection. For discretized data in transition, we may use rough set theory to describe it as quadruples formally. Assuming $IS = (I, C, V, f)$ is an information system, where $U : U = \{x_1, x_2, \ldots, x_n\}$ is the non-empty finite set of objects, named discourse domain; $C : C = \{\alpha \mid \alpha \in C\}$ *i*s a non-empty finite set of attributes, and each $\alpha_j \in C(1 \le j \le m)$ is a simple attribute of C; $V : V = \bigcup V_j (1 \le j \le m)$ denotes the codomain of the information function f and $V_j (1 \le j \le m)$ is the codomain of the attribute α_j; $f : f = \{f_j \mid f_j : U \to V_j (1 \le j \le m)\}$ represents the information function of the attributes in IS, and f_j is the information function of the attribute α_j. When $\forall \alpha \in C, \forall x \in U, f_\sigma(x)$ has no default value, we consider the information system is complete, otherwise it is incomplete. In the complete information system, we express it as $S = (U, A)$, and the definitions for $Q = (P \subseteq A)$ are as follows:

(1) The indistinguishable relationship $IND(P)$, $U / IND(P)$ of knowledge P constitute a division of U, abbreviated as $\frac{U}{P} = \{[u_i]_p \mid u_i \in U\}$. The equivalent class $[u_i]_p$ generated by knowledge P is called knowledge granularity.

(2) Given a knowledge base $K = (U, S)$, where U is discourse domain, S is the equivalence relations on the U, for $\forall X \subseteq U$ and the equivalent relation $R \in IND(K)$ on the U, the lower and upper approximations of our definition subsets X for knowledge R are as follows:

$$\begin{aligned} \underline{R}(X) &= \{x \mid (\forall x \in U) \wedge ([x]_R \subseteq x)\} \\ &= \bigcup \{Y \mid \forall Y \in U / R) \wedge (Y \subseteq X)\} \end{aligned} \quad \textbf{(1)}$$

$$\begin{aligned} \overline{R}(X) &= \{x \mid (\forall x \in U) \wedge ([x]_R \bigcap X \neq \phi)\} \\ &= \bigcup \{Y \mid Y \in U / R) \wedge (Y \bigcap X \neq \phi)\} \end{aligned} \quad \textbf{(2)}$$

The set $bn_R(X) = \overline{R}(X) - \underline{R}(X)$ is the boundary domain on X about the relationship R; $pso_R(X) = \underline{R}(X)$ is called the positive domain on X about the relationship R; $neg_R(X) = U - \overline{R}(X)$ is called the negative domain on X about the relationship R. Obviously, $\overline{R}(X) = pso_R(X) \bigcup bn_R(X)$.

(3) Given a discourse domain U and an equivalent relation R, $\forall X \subseteq U$ if $\overline{R}(X) = \underline{R}(X)$, then the set X is called the R- exact set or R- definable set on U regarding to knowledge R.

The fuzzy clustering [21-23] analysis methods can be roughly divided into two categories. The methods in the first category do not require the knowledge of the number of clusters underlying the data [24], and dynamically assign the data into different clusters. Fuzzy matrix is the core in many Fuzzy clustering

algorithms such as the direct clustering method [25], fuzzy transfer closure method, the maximum fuzzy tree method and so forth. For the methods in the second category, and the data is clustered based on objective function optimization. Among the various methods in the second category, the most widely used and effective method is the Fuzzy C-means (FCM) algorithm.

For dataset $X = \{x_1, x_2, \ldots, x_n\} \subset R^s$, where n is the number of data, and s is the dimensionality or the number of attributes, FCM algorithm divides the data set X into c clusters, where $2 \le c \le n$, with clustering centers $V = \{v_1, v_2, \ldots, v_c\}$. The FCM algorithm can be expressed as the following mathematical optimizing problem:

$$Minimiz\ J(X,U,V) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m \| x_j - v_i \|^2 \quad (3)$$

Subject to:

$$\sum_{i=1}^{c} u_{ij} = 1 \quad (4)$$

Where u_{ij} denotes the membership of sample x_j to cluster center v_i, $U = (u_{ij})_{c\times n}$ is the fuzzy partition matrix; m is the fuzzy weight index, also known as the fuzzy factor, which is mainly used to adjust the fuzziness degree of fuzzy partition matrix; $\| x_j - v_i \|^2$ is the Euclidean distance between sample x_j and the i-th cluster center and is used as the similarity measure.

The steps of FCM algorithm

Step 1: Given a presumed number of clusters: c, the fuzzy factor m (usually from 1.5 to 2.5), initialize the membership degree matrix $U^{(\gamma)}(\gamma = 0)$, and satisfy the equation (4);

Step 2: Update the clustering center $V^{(\gamma+1)} = \{v_1, v_2, \ldots, v_n\}$ according to (9);

$$v^{(\gamma+1)} = \frac{\sum_{j=1}^{n} u_{ij}^{(\gamma)^m} g x_j}{\sum_{j=1}^{n} u_{ij}^{(\gamma)^m}}, i = 1, 2, \ldots, c \quad (5)$$

Step 3: Update the membership matrix $U^{(\gamma+1)} = (u_{ij})$ according to (6);

$$v^{(\gamma+1)} = [\sum_{k=1}^{n} (\frac{\| x_j - v_i \|^{2(\gamma)}}{\| x_j - v_k \|^{2(\gamma)}})^{\frac{2}{(m-1)^{-1}}}], i = 1, 2, \ldots, c,\ j = 1, 2, \ldots, n \quad (6)$$

Step 4: Calculate $e = \| U^{(\gamma+1)} - U^{(\gamma)} \|$, if $e \le \eta$ (η is the threshold, usually 0.001 to 0.01), then the algorithm stops; otherwise $\gamma = \gamma + 1$, then go to step 2.

The FCM algorithm has been extensive explored. On one hand, the traditional C-means has been quite mature. On the other hand, because individual knowledge is relative and not immutable, the fuzzy C-means can objectively reflect the real-world categorization compared with hard clustering. Fuzzy clustering algorithm and many improved algorithms have been widely used in many fields. For example, Li Wenjuan and others proposed a scheduling algorithm based on two-level scheduling mode under fuzzy clustering [26-31]. They improved the traditional fuzzy clustering algorithm systematically, integrated the particle swarm optimization method and support vector in the clustering algorithm, and achieved good results finally.

Generally, this paper proposes a new similarity measure Sim_α based on rough sets to improve the fuzzy clustering algorithm from the following aspects: (1) New fuzzy centroid is defined to improve the accuracy of centroid; (2) By using the similarity measure Sim_α, the upper approximation and the lower approximation, the objective function of fuzzy clustering is improved.

4 Our Proposed Approaches

In traditional fuzzy clustering algorithms, all samples are treated equally, and noise samples can easily affect the clustering results. To minimize the influence of outlier samples, we propose a novel similarity measure Sim_α that combines with rough set. The main idea is to define a new fuzzy centroid to improve the precision of centroid, and then combine the new similarity measure Sim_α with the upper approximation and the lower approximation of rough set theory to improve the performance of objective function of fuzzy clustering.

4.1 Rough Fuzzy Clustering Algorithm for Similarity Measurement

Definition 1: Membership Function

Fuzzy sets are groups of data objects, each of which has a continuous membership degree assigned between 0 and 1 and calculated by the membership function. Fuzzy clustering is the combination of the idea of fuzzy set and clustering. In the traditional clustering analysis, an object can only belong to one cluster, while in fuzzy clustering, an object could fall in more than one clusters according to its membership. Suppose the number of objects (samples) for clustering analysis is *N*, *X* represents the set of elements of the object, and the objects are to be allocated into *C* clusters. The membership function is defined as *U*. The membership $u_{i,j}$ between the *i*-th object and the *j*-th cluster, satisfies the following constraints:

$$\begin{cases} \mu_{i,j}=1 \ \ k\neq 1, if \parallel x_j-c_i \parallel^2 \leq \parallel x_j-c_k \parallel^2 \\ \mu_{i,j}=0 \qquad\qquad\qquad other \end{cases} \tag{7}$$

Definition 2: Approximate Accuracy

Given a domain *U* and an equivalence relation *R* on it, $\forall X \subseteq U$, the approximate accuracy and the roughness degree of the set *X* defined by the equivalence relation *R* are as follows:

$$\alpha_R(X)=\frac{\underline{R}(X)}{\overline{R}(X)} \tag{8}$$

$$\rho_R=1-\alpha_R(X) \tag{9}$$

The inaccuracy of the set is caused by the existence of the boundary region, and the larger its bounding region is, the lower its accuracy would be. For each *R* and $X \subseteq U$, there is $0 \leq \alpha_R(X) \leq 1$. When $\alpha_R(X)=1$, the *R*- boundary filed of *X* is an empty set, so the set *X* is the exact of *R*-; when $\alpha_R(X)>1$, the set *X* have a non-empty *R*- boundary filed, so the set *X* is a rough set of *R*-; when *X* is an empty set, we set $\alpha_R(X)=\alpha_R(\phi)=1$. The *R*- roughness of *X* is opposite to the accuracy, which reflects the incompleteness of our knowledge of the category of aggregate *X* expression under knowledge *R*.

Given a domain *U* and an equivalence relation *R* on it, and a division $\pi(U)=\{X_1, X_2, \ldots, X_n\} \in \Pi(U)$ on domain *U*, this division is independent of knowledge *R*. Among them, subset $X_i (i=1, 2, \ldots, n)$ is an equivalent class for dividing $\pi(U)$. The *R* approximation and upper approximation of $\pi(U)$ are as follows:

$$\underline{R}(\pi(U))=\underline{R}(X_1) \bigcup \underline{R}(X_2) \bigcup \ldots \underline{R}(X_n) \tag{10}$$

$$\overline{R}(\pi(U))=\underline{R}(X_1) \bigcup \underline{R}(X_2) \bigcup \ldots \underline{R}(X_n) \tag{11}$$

Hence, the *R*- approximate classification accuracy and approximate classification quality of partition $\pi(U)$ are respectively defined as follows:

$$\begin{aligned} \alpha_R(\pi(U)) &= \frac{\sum_{i=1}^{n} |\underline{R}(X_i)|}{\sum_{i=1}^{n} |\overline{R}(X_i)|} = \frac{card(\underline{R}(\pi(U)))}{card(\overline{R}(\pi(U)))} \\ &= \frac{\underline{R}(\pi(U))}{\overline{R}(\pi(U))} \end{aligned} \tag{12}$$

$$\begin{aligned} \gamma_R(\pi(U)) &= \frac{\sum_{i=1}^{n} |\underline{R}(X_i)|}{|U|} = \frac{card(\underline{R}(\pi(U)))}{card(U)} \\ &= \frac{\underline{R}(\pi(U))}{|U|} \end{aligned} \tag{13}$$

Definition 3: Fuzzy Centroid

In fuzzy clustering, the calculation formula of centroid is:

$$C_j=\sum_{j=1}^{n} u_{ij}^m x_j / \sum_{j=1}^{n} u_{ij}^m \tag{14}$$

The $u_{i,j}$ represents the membership degree; *m* represents the fuzzy index, generally its value is 2. In this paper, we introduce the concepts, the upper approximation set, and the lower approximation set of rough sets. So, two centers of mass are computed: centroids of the upper and lower approximation sets.

According to the formula 3 and the formula 4, the calculation formula of the two centroids are:

$$C_u=w_l\left(\sum_{x_i \in \underline{R}(C_j)} u_{ij}^m / x_i / |\underline{R}(C_j)|\right) \tag{15}$$

$$C_p=w_u\left(\sum_{x_i \in (\overline{R}(C_j)-\underline{R}(C_j)))} u_{ij}^m / x_i / |\overline{R}(C_j)-\underline{R}(C_j)|\right) \tag{16}$$

$$w_l=V_l \times \gamma_R \tag{17}$$

$$w_u=V_u \times \gamma_R \tag{18}$$

Where C_n is the centroid of lower approximation sets; C_p is the centroid of upper approximation sets; V_i represents the weight of the sample under the approximate concentration; V_n represents the weight of the sample upper the approximate concentration $V_i+V_n=1$, and $V_i \geq V_n$.

Definition 4: Similarity Measurement Sim_α

Given a knowledge base $K=(U, S), R \in IND(K)$ represents single or a set of system parameter describing system characteristics. For $\forall X \subseteq U$ and the partition $\pi(U)=\{X_1, X_2, \ldots, X_n\}$ of *U* that is

independent of the system parameter *R*, the similarity measurement of the set *X* with respect to the system parameter *R* is defined as

$$Sim_{\alpha} = \frac{|U - bn_R(X)|}{|U|} \tag{19}$$

$$Sim_{\alpha}(\pi(u)) = \frac{\sum_{i=1}^{n} |U - bn_R(X_i)|}{n|U|} \tag{20}$$

In the fuzzy clustering algorithm, when the normalization condition of membership constraint is relaxed, the membership degree of the sample may be greater than 1. Then the membership degrees of samples of each cluster are quite different. In other words, some samples may have a high degree of membership for each cluster, while others may have a very low membership degree of each cluster. If a sample has a high degree of membership in a cluster during the clustering process, the final cluster may only contain this one sample, that is, the noise points are grouped into a single cluster. This is not the desired outcome. In addition, if the membership degree of some samples is very low, it will be very difficult to select the iterative termination threshold in the actual clustering process. Therefore, we improve the membership degree as follows:

$$H_{i,j} = u_{i,j}^2 + \sum_{j \in U, i \neq j} \frac{u_{i,j}^2}{(1 + Sim_{\alpha}(\pi(U)))} \tag{21}$$

Where $u_{i,j}$ is the membership value of the *i*-th object x_i relative to the *k*-th cluster center v_k, and the objective function is improved by using the concept of approximate set in the rough set theory. In the improved algorithm, the objective function is defined as

$$E = \sum_{i=1}^{C} \sum_{j=1}^{n} H_{i,j} d_{i,j} \tag{22}$$

Since the concept of rough set theory is introduced, the centroid is divided into two categories according to formula (15) ~ (16): the centroid C_p of the upper approximation set and the center C_n of the lower approximation set, and the target distance *d* between the object x_j and the two centroids is defined as

$$d_{i,j}(C_{u'}, C_{p'}, x_j)^2 = V_l \frac{\| C_u - x_j \|^2}{N} + V_u \frac{\| C_p - x_j \|^2}{N} \tag{23}$$

The improved algorithm

Step 1: Initialize the FCM algorithm related parameters and effectiveness indicators:

c=2, $c_{\max} = \sqrt{n}$, m=2.0,

$\zeta = 0.001$, iterations l=0, Fuzzy partition matrix U

Step 2: Initialize the membership matrix of the data set and use (15), (16) to calculate the cluster center, including the centroid C_u of the upper approximation set and the centroid C_p of the lower approximation set.

Step 3: Update the cluster center and membership matrix $H_{i,j}$ according to (19) (20) (21).

Step 4: If $\| U^{(\gamma+1)} - U^{(\gamma)} \| \leq \zeta$ (ζ is the threshold, generally take 0.001 to 0.01), then go to Step 5; otherwise go to Step 3;

Step 5: According to the membership degree matrix *U* and cluster center matrix *V* obtained in Step 4, each index of the algorithm is calculated.

Step 6: If $c_p \neq c_{\max}$, then $c = c + 1$, then return to Step 2; otherwise continue to Step 7.

Step 7: Calculate each index according to the formula.

Step 8: Find the minimum value of the validity index and select the corresponding *C* as the best number of clustering.

4.2 Evaluating Indicator

In the above algorithm, validity index is a critical factor. Next, we list a few commonly used validity indices.

Definition 5: SC Metric

SC measurement contains two parts: in-cluster compactness and inter-cluster separation. Defined as follows:

$$SC(U, V : c) = \frac{\frac{1}{c}\left(\sum_{i=1}^{c} \| v_i - \bar{v} \|^2\right)}{\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} (u_{ij})^m \| x_j - v_i \|^2 + \frac{1}{n(n-1)}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n} \| x_k - x_j \|^2} \tag{24}$$

Where $\frac{1}{v} = \frac{1}{c}\sum_{i=1}^{c} v_i$, $\frac{1}{c} = \left(\sum_{i=1}^{c} \| v_i - \bar{v} \|^2\right)$ is the degree of separation between clusters, $\frac{1}{n}\sum_{j=1}^{n} (u_{ij})^m \| x_j - v_i \|^2$

represents the compactness within the cluster. $\frac{1}{n(n-1)}\sum_{k=1}^{n-1}\sum_{j=k+1}^{n}\| x_k - x_j \|^2$ is equivalent to the penalty factor, which means the average distance between any two data samples in the data set. The larger value of the $SC(U,V:c)$ index means the better result of the clustering.

Definition 6: PC Metric

Bezdek proposed two clustering indicators of effectiveness, including the partition coefficient V_{PC} and the partition entropy V_{CE}, defined as follows:

$$V_{PC} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}^2}{n} \tag{25}$$

$$V_{CE} = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c}[u_{ij}\log_{\alpha}(u_{ij})] \tag{26}$$

where c is the number of clusters, n is the number of data samples, and the u_{ij} is the membership degree.

Definition 7: S Metric

Kwon proposed a new effectiveness index:

$$V_s = -\frac{\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}\| x_j - v_i \|^2 + \frac{1}{c}\sum_{i=1}^{c}\| v_i - \bar{v} \|^2}{\min_{i\neq k}\| v_i - v_s \|^2} \tag{27}$$

where $\bar{v} = \sum_{j=1}^{n} x_j / n,$, u_{ij} is the membership degree, v_i is the cluster center.

Definition 8: SC Metric

Kim tries to find out the best number of clusters using the two indexes, under-divided and over-divided indicators, defined as follows:

$$V_{sc} = \frac{1}{c}\sum_{i=1}^{c}\frac{\sum_{j=1}^{n}\| x_j - v_i \|}{n_i} + \frac{c}{\min\limits_{i\neq j}\| x_i - v_j \|} \tag{28}$$

where V is the matrix of cluster centers; $\min_{i\neq j}\| x_i - v_j \|$ denotes the minimum distance from the object to each cluster center.

Definition 9: DBI Metric

Pakhira and others put forward the DBI validity index, which involves two aspects: hard clustering and fuzzy clustering. Here we only discuss the -DBI index related to fuzzy clustering, which is defined as follows:

$$V_{DBI} = (\frac{1}{c}\times\frac{E_1}{J_m}\times D_c)^2 \tag{29}$$

where

$$E_1 = \sum_{j=1}^{n} u_{ij}\| x_j - v \|, D_c = \max_{i,j=1}^{c}\| v_i - v_j \| \tag{30}$$

J_m is defined as:

$$J_m(U,Z) = \sum_{j=1}^{n}\sum_{i=1}^{c}(u_{ij})^m \| v_i - v_j \| \tag{31}$$

5 Experimental Results and Analysis

In this paper, an improved fuzzy clustering function is proposed by introducing rough set theory and a new similarity measure. To test the performance of rough fuzzy clustering algorithm, the classical artificial datasets and real datasets from UCI data repository are used. The proposed rough fuzzy clustering algorithm is compared with the fuzzy clustering algorithm. Finally, the stability of each index with respect to fuzzy factor m is analyzed.

Section 4.1 briefly describes the experimental dataset information and evaluation criteria, Section 4.2 analyzes the clustering results of artificial datasets, and Section 4.3 gives the specific experimental results.

5.1 Methodology

The experimental datasets are listed in Table 1. These datasets have different numbers of attributes and clusters. All datasets are normalized. The parameter settings of FCM algorithm are as follows: termination threshold $\zeta = 0.001$, fuzzy factor $m = 2.0$, $\|*\|^2$ is Euclidean distance square, the effectiveness of the indicator α is generally set to 0.6, γ is generally set to 0.1.

The experimental results were evaluated using the indices of definition 5 to definition 9, denoted by V_{PC}, V_{SC}, V_S, V_{XB}, V_{DBI}, V_{CE}, V_{DI}, respectively. The smaller the value of V_{PC}, V_{SC}, V_S, V_{XB}, V_{DBI}, the better the clustering results, and the larger the value of the V_{CE}, V_{DI}, the better the clustering results. In addition, the algorithm is evaluated by Accuracy, Recall, and Mutual Information (NMI). Accuracy and recall are the two metrics widely used in the field of information retrieval and statistics to evaluate the quality of clustering results. Mutual information is a measure of the interdependence of variables. Specifically, the three metrics are defined as follows:

$$Acc = \frac{\sum_{i=1}^{k} a_i}{|U|} \tag{32}$$

$$\mathrm{Re} = \frac{\sum_{i=1}^{k}\frac{a_i}{a_i + c_i}}{k} \tag{33}$$

Table 1. Experimental artificial data set and real data set in this paper

DataSet	Size	#Attribute
Dp1	200	5
Iris	150	4
gesture_phase_a1_raw	1747	19
heart	270	13
zoo	101	16
seeds	210	7
optdigits	3823	64
machine	209	7
hepatitis	150	18
Haberman	306	5
Art	300	4
wine	178	13
ecoli	336	7
VertebralColumn_3C	310	6

In which, k is the number of clusters, and a_i represents the number of samples that are correctly classified to the cluster of c_i. U is the discourse domain that contains all the samples, and c_i indicates the number of samples that belong to the class cluster c_i but are wrongly classified to other clusters.

$$NMI(X,Y)=\frac{I(X,Y)}{\sqrt{H(X)H(Y)}} \tag{34}$$

where X and Y are random variables, $I(X,Y)$ is the mutual information of the two random variables, and $H(X)$ is the entropy of X. The definitions are as follows.

$$I(X,Y)=\sum_{h=1}^{k^{(a)}}\sum_{l=1}^{k^{(b)}} n_{h,l}\log(\frac{n\cdot n_{h,l}}{n_h^{(a)}n_l^{(b)}}) \tag{35}$$

$$H(Y)=\sum_{h=1}^{k^{(a)}} n_n{}^{(a)}\log\frac{n_h{}^{(a)}}{n} \tag{36}$$

$$H(Y)=\sum_{i=1}^{k^{(b)}} n_i{}^{(b)}\log\frac{n_l{}^{(b)}}{n} \tag{37}$$

All the experimental environments are Win10 64bit operating system, Matlab software, 4G memory, Intel (R) Core (TM) i5-3210M CPU@2.50GHz.

5.2 Experimental Results

The experimental results are shown in Table 2 and Table 3, where Table 2 shows the indexes of fuzzy clustering and Table 3 shows the indexes of rough fuzzy clustering.

Table 2. Fuzzy clustering results on various indicators on different data sets

DataSet	V_{PC}	V_{CE}	V_{SC}	V_S	V_{XB}	V_{DI}	V_{DBI}
Dp1	0.726505100232829	0.436413277552198	1.0157967246279821	0.203159344925597	0.991154549254301	0.707106781186547	1.01631773666832
Iris	0.783193859067484	0.395928229970596	0.1167046798838132	0.10011967666952033	4.44773915326333	0.104972776216303	0.688234976780471
gesture_phase_a1_raw	0.783193859067484	0.402357004008814	2.90318605759776E-06	2.3145158022235E-09	0.664522512482295	0.25207431867780099	0.414860317807038
heart	0.712624821086085	0.450049569454743	0.0157830941202204	0.0800584559041489647	3.0923146558636	0.0196383770594906	1.04120610066015
zoo	0.467512072905545	1.165657122980985	0.1959376359172796	0.20258386676421365	1.03898210677921	0.288675134594812	1.75549117440317
seeds	0.725687865332747	0.499712627068563	0.1031871269065616	0.700763295139412894	3.88381948794912	0.0857788341290416	0.755905454592989
optdigits	0.111111111111155	2.19722457733602	2.51671486231358	1.05934426276709	0.157372747405708	0.099236115527	1.64587115464036
machine	0.800253260353247	0.398541985649825	5.13569873291395E-06	0.4289845217551612473	6.77394452320352	0.041662833855083	0.478326866470177
hepatitis	0.823382598442989	0.296278740010827	0.6971304589501935	0.4000464753639300129	3.49465738831668	0.0430739664489499	0.883166924869982
Haberman	0.739323223616515	0.41433163657272	0.0619855058218884	0.204202560.7012489831	2.90718849953904	0.026157762372378	0.956277534479891
Art	0.804508192395293	0.366711772492897	0.109987846369538	0.512570284118788448	4.42464925472097	0.0241385777292517	0.630676391137862
wine	0.790939816597689	0.380407694407004	0.742532907522503197	0.683639347840709	5.69478618836047	0.0116954957036635	0.52794379936199
ecoli	0.340691447335117	1.41476183290842	1.55709467246263	0.426813610151368	1.27968647230718	0.0482626949965598	2.39779318624323
VertebralColumn_3C	0.631656072967181	0.649052792690546	0.2144558891154133	0.7123046349108855216	3.22268091496792	0.0291518664392343	1.17504881052556

Table 3. Results of each index of rough fuzzy clustering on different data sets

DataSet	V_{PC}	V_{CE}	V_{SC}	V_S	V_{XB}	V_{DI}	V_{DBI}
Dp1	0.197406895522595	0.470769405228366	0.0531849070957982	0.203159344925597	0.40907872655245	1	0.583333333333333
Iris	0.037037037037037	0.732408192445406	0.0352733680606701	0.10011967666952033	0.185185185185185	141139.359234409	0.575535398472538
gesture_phase_a1_raw	0.008	0.643775164973639	0.0171117296232016	2.3145158022235E-09	0.00800472801066742	9.47768479082	0.0109076727937957
heart	0.037037037037037	0.732408192445406	0.0143755633017475	0.0800584559041489647	0.0375572328700073	2274.20810456063	0.324816796158979
zoo	0.00291545189504373	0.555974328301518	0.000632619614132707	0.20258386676421365	0.00301963333475345	95782.6285221151	1.46196398562502
seeds	0.015625	0.693147180559945	0.0000405049955109754	0.700763295139412894	0.0371683641685104	83552.3943261416	0.328276226511074
optdigits	0.00137174211248285	0.488272128296937	2.83995765762572E-08	1.05934426276709	0.0387802238016506	12663.3014896726	1.33146854761584
machine	0.0081000019	0.64377516497364	1.83195284905798E-10	0.4289845217551612473	0.00801608240317052	13.9899910350536	0.2003246397437
hepatitis	0.121881333640064	0.674628037707514	6.46462177829262E-07	0.4000464753639300129	0.128860056687371	40.0440387489235	0.536271342263581
Haberman	0.125	0.693147180559945	9.46513190468311E-08	0.204202560.7012489831	0.133619157383862	15617.3761888606	0.28487897868633
Art	0.012564823813336	0.693147180559945	0.0000842930488979543	0.512570284118788448	0.0372807973278505	131319.791931854	0.600391626975418
wine	0.037037037037037	0.835608365453214	7.09978208749165E-08	0.683639347840709	0.0372365382758532	713.168982503712	0.779944043677399
ecoli	0.00387848050168277	1.61417020358402	0.00172713057181804	0.426813610151368	0.00388657031070512	4.02677427167527	1.83645754351612
VertebralColumn_3C	0.0156249999999999	0.693147180559945	6.58832709551406E-07	0.7123046349108855216	0.0370952528166905	2196.80573030795	0.571932729021605

From Table 2 and Table 3, we can see that rough fuzzy clustering algorithm has smaller values of V_{PC}, V_{SC}, V_S, V_{XB}, V_{DBI}, than the fuzzy clustering algorithm, while V_{CE}, V_{DI}, of the rough fuzzy clustering algorithm are larger than that of the fuzzy clustering algorithm. Therefore, the rough fuzzy clustering algorithm outperforms the fuzzy clustering algorithm.

Table 4 shows the accuracy of clustering results of each algorithm, Table 5 compares the recall rate of clustering results of each algorithm, and Table 6 is the mutual information (NMI) of clustering results of each algorithm. The comparison of these three algorithms shows that the proposed algorithm performs better than K-MEANS and FCM in accuracy, recall and mutual information. In addition, the optimal cluster number obtained by rough fuzzy clustering algorithm is consistent with the real cluster number of datasets, which means our proposed algorithm works better than other existing algorithms At the same time, our algorithm can find the correct number of clusters by choosing a suit weighted exponent *m*, which shows that the algorithm has good stability.

Table 7 compares the proposed algorithm with the original FCM algorithm, as well as the classic K-MEANS, AP, FCM clustering algorithms, and the time performance of the improved algorithm AFCM and KFCM. From the experimental results, we can see that the algorithm in this paper inherits the time advantage of FCM algorithm, and running time takes less time than other algorithms, so it has some advantages in both running time and memory consumption.

5.3 Analysis of Experimental Results of Olivetti Data Set

To compressively test our proposed algorithms, we also used Olivetti dataset. The Olivetti faces dataset contains forty face images of individuals, and each person's face image contains ten images from different angles. It is really challenging for cluster algorithms since the number of people (the number of clusters) is more than the number of face images of each person (the number of samples in the cluster).

Table 4. Comparison of algorithm accuracy

Dataset	ACC(%)					
	K-MEANS	AP	FCM	AFCM	KFCM	Algorithm in this paper
Dp1	92	91.5	98.7	98.7	98.7	98.7
Iris	89.3	88.6	89.8	89.9	90.0	90.0
gesture_phase_a1_raw	75.4	70.2	80.2	82.3	85.4	87.5
heart	62.7	59.4	75.2	84.2	80.6	85.3
zoo	87.8	85.3	92.0	91.3	90.0	92.5
seeds	80.2	78.9	87.1	89.5	90.0	91.6
optdigits	78.3	80.4	82.4	83.6	84.8	87.1
machine	83.6	85.2	89.7	91.1	93.6	96.2
hepatitis	71.4	69.8	75.2	75.9	76.4	78.5
Haberman	62.7	70.3	79.7	82.3	81.9	85.6
Art	84.7	80.67	86.5	87.6	88.7	89.6
wine	95.5	87.8	88.2	89.7	90.1	91.5
ecoli	30.1	73.21	69.8	71.9	73.8	75.6
VertebralColumn_3C	82.3	78.5	80.3	81.6	82.5	84.7

Table 5. Comparison of the algorithm recall rate

Dataset	RE(%)					
	K-MEANS	AP	FCM	AFCM	KFCM	Algorithm in this paper
Dp1	90.6	91.4	98.7	98.7	98.7	98.7
Iris	89.3	88.6	89.8	89.9	90.0	90
gesture_phase_a1_raw	60.1	70.2	75.1	77.5	79.3	82.3
heart	58.4	62.3	73.2	75.2	77.6	82.1
zoo	80.3	79.1	85.1	86.3	86.5	89.7
seeds	80.2	81.6	87.1	88.6	89.1	91.6
optdigits	70.1	77.8	79.5	80.3	81.2	85.2
machine	80.2	84.9	85.3	87.4	88.3	92.5
hepatitis	65.4	67.8	70.1	70.9	71.5	73.2
Haberman	62.7	69.5	79.7	80.2	81.9	85.6
Art	84.7	85.8	86.5	87.3	84.8	89.6
wine	96.2	88.6	90.1	90.5	91.2	93
ecoli	59.3	60.0	60.3	59.4	61.3	65.7
VertebralColumn_3C	77.1	73.2	75.4	70.9	76.3	80.2

Table 6. Comparison of algorithmic mutual information

Dataset	NMI(%)					
	K-MEANS	AP	FCM	AFCM	KFCM	Algorithm in this paper
Dp1	90.1	91.2	97.6	97.9	98.0	98.3
Iris	75.8	70.5	77.8	59.8	78.0	78.0
gesture_phase_a1_raw	66.1	69.4	80.2	81.4	83.2	85.7
heart	55.3	60.3	69.7	70.6	73.5	79.2
zoo	75.1	77.5	80.2	79.9	82.7	85.3
seeds	67.4	59.4	61.1	62.3	63.6	66.1
optdigits	73.1	74.3	78.4	80.1	84.8	87.6
machine	78.5	80.6	82.4	83.2	84.5	90.7
hepatitis	69.1	70.9	75.3	76.8	77.9	80.6
Haberman	60.6	69.1	75.4	76.0	74.9	82.3
Art	72.7	70.8	75.6	76.3	77.0	77.2
wine	85.3	69.7	71.0	72.6	71.9	76.5
ecoli	60.7	59.2	61.1	60.9	62.8	66.1
VertebralColumn_3C	75.7	76.3	74.2	76.0	77.3	78.6

Table 7. Comparison of running time of various clustering algorithms on UCI datasets (seconds /s)

Dataset	Iris	Seeds	Dp1	gesture_phase_a1_raw	heart	zoo	Wine
K-MEANS	0.059	0.122	0.884	0.309	0.349	8.727	0.098
AP	0.565	0.973	6.115	3.016	2.018	66.79	0.832
FCM	0.148	0.164	0.464	2.602	0.309	0.313	0.168
AFCM	0.425	0.932	4.251	4.445	3.104	54.265	1.203
KFCM	0.049	0.05	0.092	0.068	0.064	0.806	0.048
Algorithm in this paper	0.032	0.074	0.090	0.059	0.071	0.912	0.032

We applied our proposed algorithm to cluster the Olivetti dataset. The clustering results are showed in Figure 1 and Figure 2. Figure 1 is the original images of face data set and includes a total of ten rows and each row has two clusters. Figure 2 is the result of cluster recognition. The wrong face images in cluster recognition were added to horizontal shadows and marked an "×" for each one in the left white part.

Table 8 shows the results of ACC, RE, and NMI using various algorithms. The commonly used algorithms such as FCM and K-MEANS can only find twenty Peak density points, which means twenty groups of faces can be fully clustered and identified. On the other hand, our proposed algorithm could successfully identify thirty-five groups of faces, with more than half of the correct labels in one group. For the unrecognized faces, it is found that the images within the same cluster are quite different. In such a scenario, it is difficult to accurately identify based on distance metrics only. The accuracy may be further improved by feature extraction.

Table 8. Cluster evaluation result table

Algorithm	ACC	RE	NMI
K-MEANS	0.64	0.65	0.56
FCM	0.76	0.78	0.67
Algorithm in this paper	0.82	0.83	0.73

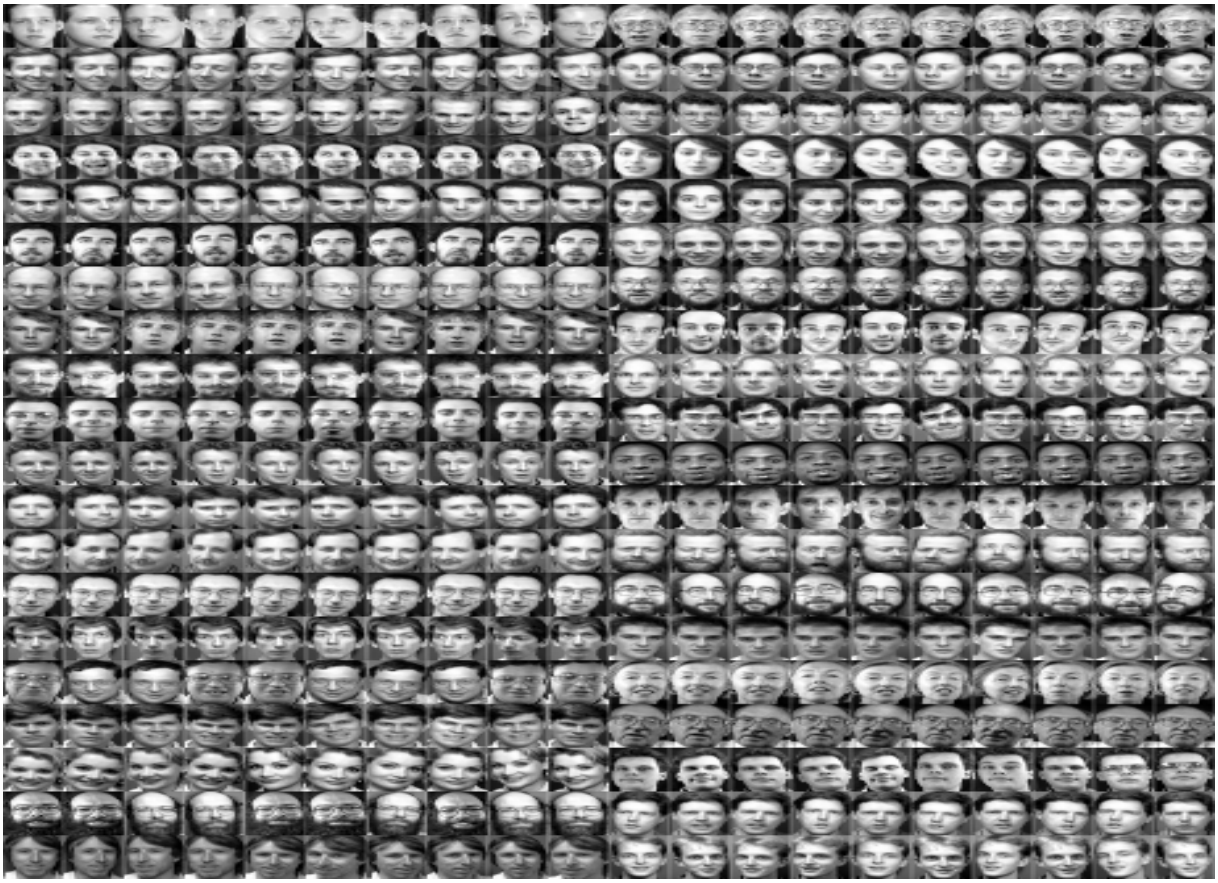


Figure 1. Face data set original graph

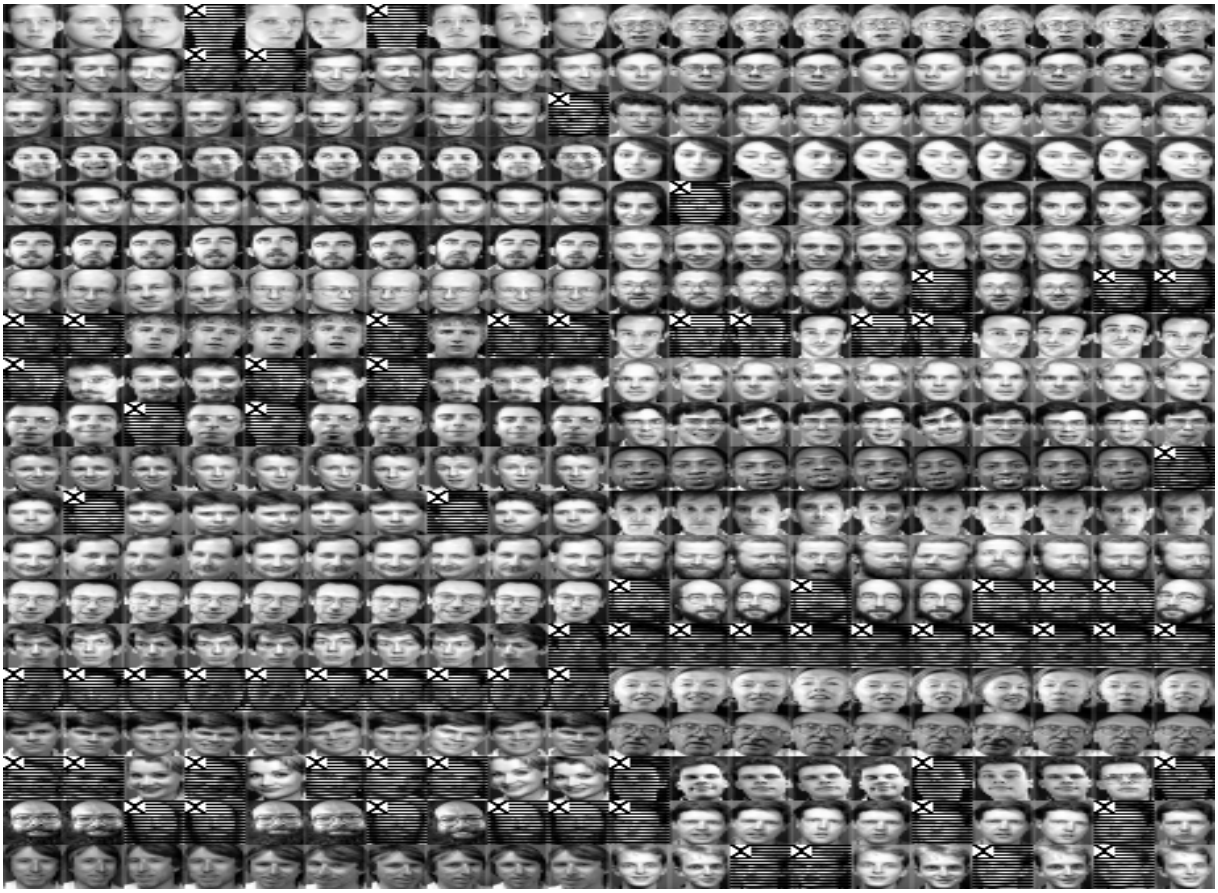


Figure 2. The result of cluster recognition

6 Conclusion

In this paper, a new measurement of similarity based on concepts of upper approximation and lower approximation in rough set theory is proposed. This measure addresses issues such as uncertain boundaries and complex data. Based on the new similarity measure, a rough fuzzy clustering algorithm is developed. Extensive experiments are conducted and the results have proved that the rough fuzzy clustering algorithm not only has good stability but also produces improved clustering results.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. This work is supported by The State Key Research Development Program of China under Grant 2017YFC0804406, Shandong Provincial Natural Science Foundation of China under Grant ZR2018MF009, National Natural Science Foundation of China under Grant 91746104, the Special Funds of Taishan Scholars Construction Project, and Leading Talent Project of Shandong University of Science and Technology.

References

[1] T. F. Chan, J. Shen, Nontexture Inpainting by Curvature-Driven Diffusions, *Journal of Visual Communication and Image Representation*, Vol. 12, No. 4, pp. 436-449, December, 2001.

[2] M. Erisoglu, N. Calis, S. Sakallioglu, A New Algorithm for Initial Cluster Centers in K-means Algorithm, *Pattern Recognition Letters*, Vol. 32, No. 14, pp. 1701-1705, October, 2011.

[3] M. M. Eusuff, K. E. Lansey, Optimization of Water Distribution Network Design Using the Shuffled Frog Leaping Algorithm, *Journal of Water Resources Planning and Management*, Vol. 129, No. 3, pp. 210-225, April, 2003.

[4] J. Gllavata, E. Qeli, B. Freisleben, Detecting Text in Videos Using Fuzzy Clustering Ensembles, *IEEE Eighth IEEE International Symposium on Multimedia*, San Diego, CA, 2006, pp. 283-290.

[5] M. Hassan, A. Chaudhry, A. Khan, J. Y. Kim, Carotid Artery Image Segmentation Using Modified Spatial Fuzzy c-means and Ensemble Clustering, *Computer Methods and Programs in Biomedicine*, Vol. 108, No. 3, pp. 1261-1276, August, 2012.

[6] J. Han, J. Pei, M. Kamber, *Data Mining: Concepts and Techniques*, Elsevier, 2011.

[7] K. Kang, H. X. Zhang, Y. Fan, A Novel Clusterer Ensemble Algorithm Based on Dynamic Cooperation, *International Conference on Fuzzy Systems & Knowledge Discovery IEEE*, Hong Kong, China, 2008, pp. 32-35.

[8] J. Y. Liang, Y. H. Qian, D. Y. Li, Theory and Method of Granular Computing for Big Data Mining, *Science China: Information Sciences*, Vol. 45, No. 11, pp. 1355-1369, January, 2015.

[9] L. A. Zadeh, Fuzzy Sets, *Information and Control*, Vol. 8, No. 3, pp. 338-353, June, 1965.

[10] M. P. Windham, Cluster Validity for Fuzzy Clustering Algorithms, *Fuzzy Sets and Systems*, Vol. 5, No. 2, pp. 177-185, March, 1981.

[11] J. C. Bezdek, Numerical Taxonomy with Fuzzy Sets, *Journal of Mathematical Biology*, Vol. 1, No. 1, pp. 57-71, May, 1974.

[12] R. N. Dave, Validating Fuzzy Partitions Obtained through C-Shells Clustering, *Pattern Recognition Letters*, Vol. 17, No. 6, pp. 613-623, May, 1996.

[13] Y. I. Kim, D. W. Kim, D. Lee, K. H. Lee, A Cluster Validation Index for GK Cluster Analysis Based on Relative Degree of Sharing, *Information Sciences*, Vol. 168, No. 1, pp. 225-242, December, 2004.

[14] I. Gath, A. B. Geva, Fuzzy Clustering for the Estimation of the Parameters of the Components of Mixtures of Normal Distributions, *Pattern Recognition Letters*, Vol. 9, No. 2, pp. 77-86, February, 1989.

[15] J. J. Niu, C. C. Huang, J. H. Li, M. Fan, Parallel Computing Techniques for Concept-Cognitive Learning Based on Granular Computing, *International Journal of Machine Learning & Cybernetics*, Vol. 9, No. 3, pp. 1-21, February, 2018.

[16] X. L. Xie, G. Beni, A New Fuzzy Clustering Validity Criterion and Its Application to Color Image Segmentation, *Proceedings of the 1991 IEEE International Symposium on Intelligent Control*, Arlington, VA, 1991.

[17] S. H. Kwon, Cluster Validity Index for Fuzzy Clustering, *Electronics Letters*, Vol. 34, No. 22, pp. 2176-2177, November, 1998.

[18] D. A. Linkens, M. Y. Chen. Input Selection and Partition Validation for Fuzzy Modelling Using Neural Network, *Fuzzy Sets and Systems*, Vol. 107, No. 3, pp. 299-308, November, 1999.

[19] M. K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity Index for Crisp and Fuzzy Clusters, *Pattern Recognition*, Vol. 37, No. 3, pp. 487-501, March, 2004.

[20] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Perez, I. Perona, An Extensive Comparative Study of Cluster Validity Indices, *Pattern Recognition*, Vol. 46, No. 1, pp. 243-256, January, 2013.

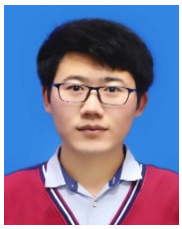
[21] J. Zhang, T. Li, H. Chen, Composite Rough Sets for Dynamic Data Mining, *Information Sciences*, Vol. 257, No. 4, pp. 81-100, February, 2014.

[22] J. Zhou, W. Pedrycz, D. Miao, Shadowed Sets in the Characterization of Rough-Fuzzy Clustering, *Pattern Recognition*, Vol. 44, No. 8, pp. 1738-1749, January, 2011.

[23] H. Zhang, H. Y. Tan, Y. H. Qian, R. Li, Q. Chen, Chinese Text Deception Detection Based on Ensemble Learning, *Journal of Computer Research and Development*, Vol. 52, No. 5, pp. 1005-1013, May, 2015.

[24] J. Fan, Z. Niu, Y. Liang, Z. Zhao, Probability Model Selection and Parameter Evolutionary Estimation for Clustering Imbalanced Data Without Sampling, *Neurocomputing*, Vol. 211, No. 10, pp. 172-181, October, 2016.
[25] Z. Feng, J. Fan, A Novel Validity Index in Fuzzy Clustering Algorithm, *International Journal of Wireless and Mobile Computing*, Vol. 10, No. 2, pp. 183-190, January, 2016.
[26] W. J. Li, Q. F. Zhang, L. D. Ping, X. Z. Pan, Cloud Scheduling Algorithm Based on Fuzzy Clustering, *Journal of Communications*, Vol. 33, No. 3, pp. 146-154, March, 2012.
[27] N. Juneam, S. Kantabutra, NC Algorithms for Minimum Sum of Diameters Clustering. *Journal of Internet Technology*, Vol. 18, No. 4, pp. 899-905, July, 2017.
[28] T. Chang, H. Wang, S. Yu, A Novel Approach for Complex Datasets Clustering/Classification, *Journal of Internet Technology*, Vol. 17, No. 3, pp. 523-530, May, 2016.
[29] J. Ye, X. Chen, J. Ma, Improved Algorithm for Secure Outsourcing of Modular Exponentiations with High Checkability, *Ubiquitous Computing*, Vol. 23, No. 3, pp. 182-191, Januar, 2016.
[30] H. Cui, M. Xie, Y. Cai, Cluster Validity Index for Adaptive Clustering Algorithms. *IET Communications*, Vol. 8, No. 13, pp. 2256-2263, September, 2014.
[31] L. Cong, S. Ding, L. Wang, Image Segmentation Algorithm Based on Superpixel Clustering, *IET Image Processing*, Vol. 12, No. 11, pp. 2030-2035, July, 2018.

Biographies

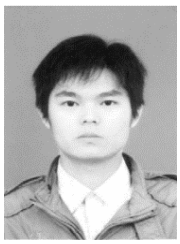


Yang Li is currently a graduate student of Shandong University of Science and Technology. His research interests include data mining and machine learning.

Jian-Cong Fan received the B.S., M.S., and Ph.D. degrees from the College of Computer Science and Engineering, Shandong University of Science and Technology, China, in 2000, 2003, and 2010, respectively. He is currently a Professor with the Shandong University of Science and Technology, Qingdao, China. His research interests include data mining and machine learning.

Jeng-Shyang Pan received the Ph.D. degree in Electrical Engineering from the University of Edinburgh, U.K. in 1996. Currently, he is a professor in College of Computer Science and Engineering, Shandong University of Science and Technology, China. He is the IET Fellow, UK and was offered Thousand Talent Program in China in 2010. His research interests include artificial intelligence and information security.

Gui-Han Mao graduate student of Shandong University of Science and Technology, majoring in machine learning, natural language processing.



Geng-Kun Wu received the Ph.D. degree from the School of Computer Science and Technology, Ocean University of China, in 2015. He worked as a postdoctoral researcher at Zhejiang University from 2015 to 2017. He is currently a Lecturer with the College of Computer Science and Engineering, Shandong University of Science and Technology. His research interests include modeling and optimization, ocean wave modeling and rendering.