Spectral Clustering with Neighborhood Attribute Reduction Based on Information Entropy

Hongjie Jia[1,2], Shifei Ding[1,2], Heng Ma[1], Wanqiu Xing[1]
1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China
Email: jiahongjie@cumt.edu.cn
2. Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190 China
Email: dingsf@cumt.edu.cn

***Abstract*—Traditional rough set theory is only suitable for dealing with discrete variables and need data preprocessing. Neighborhood rough sets overcome these shortcomings with the ability to directly process numeric data. This paper modifies the attribute reduction method based on neighborhood rough sets, in which the attribute importance is combined with information entropy to select the appropriate attributes. When multiple attributes have the same importance degree, compare the information entropy of these attributes. Put the attribute having the minimal entropy into the reduction set, so that the reduced attribute set is better. Then we introduce this attribute reduction method to improve spectral clustering and propose NRSR-SC algorithm. It can highlight the differences between samples while maintaining the characteristics of data points to make the final clustering results closer to the real data classes. Experiments show that, NRSR-SC algorithm is superior to traditional spectral clustering algorithm and FCM algorithm. Its clustering accuracy is higher, and has strong robustness to the noise in high-dimensional data.**

***Index Terms*—neighborhood rough sets, information entropy, attribute reduction, spectral clustering**

I. Introduction

Clustering analysis is an important method for data mining and information statistics. It can effectively find the intrinsic link between things and is able to describe the internal structure of the data set. The purpose of clustering is to divide the data set into several categories according to certain similarity measure, so that data points belonging to the same class have high similarity, while the data points belonging to different classes share low similarity [1]. k-means algorithm and FCM algorithm [2] are typical clustering methods. They are suitable for handling data sets of spherical structures, but for non-convex data set, the algorithm usually fall into local optimal. Spectral clustering algorithm is a new clustering method, which treats clustering problem as a graph partitioning problem. It can find a global optimal solution of the graph cut objective function by constructing the graph Laplacian matrix and its eigen-decomposition [3]. Compared with the conventional clustering algorithms, spectral clustering has obvious advantages that it can converge to the global optimal for the clustering on any shape of sample space, especially suitable for dealing with non-convex data set.

Spectral clustering is based on algebraic graph theory to clustering. As it has a solid theoretical foundation and good clustering results, the research in this area is very active at present. Blekas and Lagaris [4] use Newton's second law to analyze the interaction between data points, in order to obtain valuable similarity information, which can help reduce the overlap between classes and improve cluster perspicuity. Li and Guo [5] develop a new method of constructing similarity matrix according to the spread of neighbor relations. In this way, the similarities of data pairs within the same cluster will be enhanced, so the derived similarity matrix can well describe the real structure of data. Frederix and Van Barel [6] propose a novel sparse spectral clustering method. This method builds an approximation by analyzing the structure of Laplacian matrix, so that the complexity of eigen-decomposition can be reduced. Wang and Dong [7] introduce multi-level low-rank matrix approximation into spectral clustering and present an effective sampling method. This algorithm can effectively deal with the segmentation problem of high resolution images. Jiao et al. [8] exploit enhanced spectral embedding to improve the performance of semi-supervised clustering. Both the geometry information of data set and the given pairwise constraints are taken into account in this algorithm and it is robust to noise and outliers. Michoel and Nachtergaele [9] use hypergraph representations to detect communities in complex networks. It can simultaneously analyze multiple types of interactions and dynamically describe the functional part of the system. Adefioye et al. [10] develop a multi-view spectral clustering algorithm used for the classification of chemical compounds. They find that the compounds gathered together usually have very different chemical properties, which can help reveal the intrinsic relationship of these compounds. Ding et al. [11] apply spectral clustering to controlled islanding problem,

Manuscript received July 11, 2013; revised August 1, 2013; accepted September 1, 2013.
This work is supported by the National Natural Science Foundation of China (Nos.61379101), the National Key Basic Research Program of China (No. 2013CB329502), and the Natural Science Foundation of Jiangsu Province (No.BK20130209).

doi:10.4304/jcp.9.6.1316-1324

aiming to reduce the complexity of the problem and find an appropriate islanding solution to prevent the accidents of large area blackouts.

Nowadays, science and technology is growing by leaps and bounds and massive data result in "data explosion" [12]. These data are often accompanied by high dimensions. Traditional clustering algorithms are unable to meet the requirements of today's data analysis. When dealing with high-dimensional data, some clustering algorithms that perform well in low-dimensional data space are often unable to get good clustering results, and even invalid [13]. Therefore, designing and developing novel clustering algorithms for mining massive high-dimensional data, has become one of the researching hotspots at home and abroad [14]. Attribute reduction is an effective way to decrease the size of data, and it is often used as a preprocessing step for data mining. The essence of attribute reduction is to remove irrelevant or unnecessary attributes while maintaining the classification ability of knowledge base, in order to decrease the data size, reduce the computational complexity, and improve the efficiency of algorithms [15]. From an economic perspective, efficient attribute reduction not only can improve the knowledge clarity in intelligent information systems, but also reduce the cost of information systems to some extent. This reflects the business idea that "cost minimization, benefit maximization", which is of great significance for business intelligence.

In order to effectively deal with high-dimensional data, we combine attribute reduction with clustering analysis, and present a novel spectral clustering algorithm based on neighborhood rough sets reduction (NRSR-SC). This paper is organized as follows: section 2 describes some related concepts of neighborhood rough sets, and then uses information entropy to improve the attribute reduction algorithm based on neighborhood rough sets; section 3 analyzes the deficiencies of traditional spectral clustering algorithm processing high-dimensional data, and proposes NRSR-SC algorithm; section 4 verifies the effectiveness of NRSR-SC algorithm using the data sets in UCI machine learning database; finally, we summarize the main contents of this paper and discuss the priorities for future research.

II. Attribute Reduction with Neighborhood Rough Sets

Rough set theory is proposed by professor Pawlak [16] in 1982. It defines the concepts of domain, lower approximation and upper approximation etc. to describe the process of human learning and reasoning. In rough set theory, the acquired knowledge is represented by production rules, which are easy for users to understand, accept and use. As a mathematical tool for processing fuzzy and uncertain knowledge, rough set is widely used in selecting attribute subsets, finding decision rules, discovering knowledge dependency and other fields. Attribute reduction is one of the core contents of rough set knowledge discovery. It describes whether each attribute in the attribute set of information system is necessary, and how to remove unnecessary knowledge.

However, Pawlak rough set is based on the classic theory of equivalence relations and equivalence classes, so it is only suitable for handling discrete data. As for continuous data, which are widespread in the real world, should be discretized in advance. We can select appropriate division points to divide the range of continuous attribute values into a number of discrete intervals, and then use different integers to represent the attribute values in each subinterval. Such conversion will inevitably result in the loss of information and the processing results depend largely on the effectiveness of discrete. To solve this problem, Hu Qinghua et al. introduce neighborhood relations into rough set and propose neighborhood rough set model [17]. This model can directly analyze the attributes with continuous values, eliminating the process of discretization. Therefore, it has great advantages in feature selection and classification accuracy.

A. The Measurement of Neighborhood

Rough sets describe the problem to be addressed as an information system. Information system $IS = \langle U, A, V, f \rangle$, where U is a non-empty finite set of objects, called domain; A is a set of attributes, including condition attribute and decision attribute; V is the range of all attribute values; $f : U \times A \to V$ is an information function, indicating the mapping relationship between sample and its corresponding attribute values.

Definition 1. Domain $U = \{x_1, x_2, \cdots, x_n\}$ is a non-empty finite set in real space, for $x_i \in U$, the δ-neighborhood of x_i is defined as:

$$\delta(x_i) = \{x \mid x \in U, \Delta(x, x_i) \le \delta\} \tag{1}$$

where $\delta \ge 0$, $\delta(x_i)$ is called the neighborhood particle of x_i, Δ is a distance function. For $x_1, x_2, x_3 \in U$, Δ satisfies the following relations:

① $\Delta(x_1, x_2) \ge 0$, $\Delta(x_1, x_2) = 0$, if and only if $x_1 = x_2$;

② $\Delta(x_1, x_2) = \Delta(x_2, x_1)$;

③ $\Delta(x_1, x_3) \le \Delta(x_1, x_2) + \Delta(x_2, x_3)$.

For a sample set of N attributes, distance is usually calculated by P-norm:

$$\Delta_p(x_1, x_2) = \left(\sum_{i=1}^{N} \left| f(x_1, a_i) - f(x_2, a_i) \right|^p\right)^{1/p} \tag{2}$$

where $f(x,a_i)$ is the value of attributes a_i of sample x. If a_i is symbolic attribute, define:

① $\left| f(x_1, a_i) - f(x_2, a_i) \right| = 0$, if x_1, x_2 have the same value on attribute a_i;

② $\left| f(x_1, a_i) - f(x_2, a_i) \right| = 1$, if x_1, x_2 have different values on attribute a_i.

For example, domain $U = \{x_1, x_2, x_3, x_4, x_5\}$, a is an attribute of U, $f(x,a)$ represents the attribute value of sample x on attribute a. $f(x_1, a) = 1.1$, $f(x_2, a) = 1.2$,

$f(x_3,a)=1.6$, $f(x_4,a)=1.8$, $f(x_5,a)=1.9$. Set the neighborhood size $\delta=0.2$. As $|f(x_1,a)-f(x_2,a)|\le 0.2$, so $x_2\in\delta(x_1)$, $x_1\in\delta(x_2)$. Then we can derive each sample's δ-neighborhood: $\delta(x_1)=\{x_1,x_2\}$, $\delta(x_2)=\{x_1,x_2\}$, $\delta(x_3)=\{x_3,x_4\}$, $\delta(x_4)=\{x_3,x_4,x_5\}$, $\delta(x_5)=\{x_4,x_5\}$. If domain U contains multiple attributes, the δ-neighborhood of samples can be calculated in a similar way.

B. Measure Knowledge Entropy

Definition 2. Given a domain $U=\{x_1,x_2,\cdots,x_n\}$ located in real space. A represents the attribute set of U; D represents the decision attribute. If A is able to generate a family of neighborhood relationship of domain U, then $NDT=\langle U,A,D\rangle$ is called a neighborhood decision system.

For a neighborhood decision system $NDT=\langle U,A,D\rangle$, domain U is divided into N equivalence classes by decision attribute D: $X_1,X_2,\cdots,X_N$. $\forall B\subseteq A$, the lower approximation, upper approximation and decision boundary of decision attribute D about B are respectively defined as:

$$\underline{N_B}D=\bigcup_{i=1}^{N}\underline{N_B}X_i \tag{3}$$

$$\overline{N_B}D=\bigcup_{i=1}^{N}\overline{N_B}X_i \tag{4}$$

$$BN(D)=\overline{N_B}D-\underline{N_B}D \tag{5}$$

where $\underline{N_B}X_i=\{x_i\mid\delta_B(x_i)\subseteq X_i,x_i\in U\}$, $\overline{N_B}X_i=\{x_i\mid\delta_B(x_i)\cap X_i\neq\varnothing,x_i\in U\}$.

To illustrate the concept of upper and lower approximation, here we give an example of the classification of two classes. Suppose the domain U contains two equivalence classes, as shown in Figure 1. One class of samples marked with "*", and the other samples marked with "+". It can be seen from the figure that the samples in the circular neighborhood of sample x_1 are all from class "*", so x_1 belongs to the lower approximation of class "*"; the samples in the neighborhood of x_3 are all from class "+", so x_3 belong to the lower approximation of class "+"; the neighborhood of sample x_2 covers both class "*" samples and class "+" samples, so x_2 is a boundary sample. Thus, this definition is consistent with our intuitive understanding to classification problems.

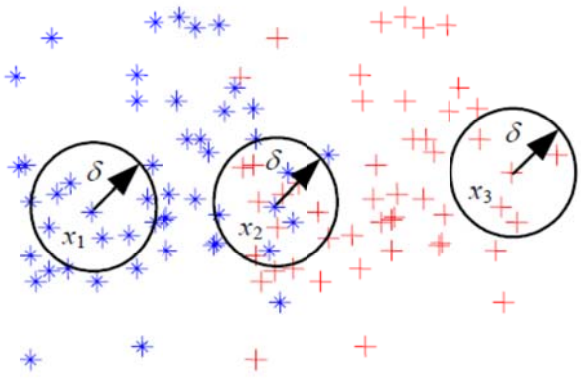


Figure 1. Neighborhood rough set model.

The lower approximation $\underline{N_B}D$ of decision attribute D is also called positive decision region denoted as $POS_B(D)$. The size of $POS_B(D)$ reflects the separable degree of domain U in a given attribute space. The greater the positive region is, means the sharper the boundaries of each class and the less overlap.

According to the nature of positive region, we can define the dependence of decision attribute D on condition attribute B:

$$\gamma_B(D)=\frac{Card(\underline{N_B}D)}{Card(U)} \tag{6}$$

where $0\le\gamma_B(D)\le 1$. $\gamma_B(D)$ represents the ratio that the samples fully contained in certain type of decision accounted for all the samples, according to the description of condition attribute B, in the sample set. Obviously, the greater the positive region $\underline{N_B}D$, the stronger the dependence of decision D on condition B.

C. Neighborhood Attribute Reduction based on Information Entropy

In a neighborhood decision system $NDT=\langle U,A,D\rangle$, $\forall B\subseteq A$, $\forall a\in B$, if $\gamma_{B-a}(D)<\gamma_B(D)$, then a is called essential to B; if $\gamma_{B-a}(D)=\gamma_B(D)$, then a is called redundant. When $\forall a\in B$ are all essential to B, then B is called independent.

Definition 3. Given a neighborhood decision system $NDT=\langle U,A,D\rangle$, If B satisfied the following two conditions, then $B\subseteq A$ is called a reduction of A.

① $\forall a\in B$, $\gamma_{B-a}(D)<\gamma_B(D)$;

② $\gamma_A(D)=\gamma_B(D)$.

The defined condition ① requires that a reduction can not contain redundant attributes, that is, the reduction must be independent; condition ② requires that of the distinguish ability of the system should remain unchanged after reduction. If $B_1,B_2,\cdots,B_k$ is the total reduction of system NDT, then $Core=\bigcap_{i=1}^{k}B_i$ is called the core of decision system.

Definition 4. Given a neighborhood decision system $NDT = \langle U, A, D \rangle$, $B \subseteq A$, $\forall a \in A - B$, then the significant degree of *a* relative to *B* is defined as:

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \tag{7}$$

Using the attribute significance index we can design the attribute reduction algorithm based on neighborhood rough sets: first calculate the importance degree of all of the remaining attributes, and then add the attribute having the greatest importance to the reduction set; repeat the process until the importance degree of all the remaining attributes is 0, which means that adding any new attribute, the dependent function values of system are no longer changed [18]. However, sometimes several attributes may have the same greatest importance degree. Traditional reduction algorithms take the approach of randomly choosing one of the attributes, which is obviously arbitrary does not taking into account the impact of other factors on attribute selection and may lead to poor reduction results.

From the viewpoint of information theory to analyze attribute reduction can improve the reduction accuracy, and many scholars have conducted research in this area. Wang Guoyin et al. [19] propose a decision table reduction algorithm based on conditional information entropy. Ding Shouzhen et al [20] combine information entropy theory with the deterministic relationship based on positive region. It can express the probability of causation between attributes, and has strong robustness to noise data. Wu Shangzhi et al. [21] introduce mutual information into decision table to describe attribute importance, and thus obtain the relative reduction of attributes. The father of information theory Shannon noted that redundancy exists in any information [22]. The size of redundancy is related with the occurrence probability or say uncertainty of each symbol (numbers, letters or words) in information. The basic role of information is to eliminate the uncertainty of things. The greater the uncertainty, the greater the entropy value, and the larger the amount of information required to make it clear [23]. The definition of entropy is given below.

Definition 5. Given knowledge *P* and its partition $U / P = \{X_1, X_2, \cdots, X_n\}$ exported on domain *U*. The information entropy of knowledge *P* is defined as:

$$H(P) = -\sum_{i=1}^{n} p(X_i) \log p(X_i) \tag{8}$$

where $p(X_i) = |X_i| / |U|$ represents the probability of equivalence class X_i on domain *U* [24].

In this paper, we use information entropy as another criterion to evaluate attributes. When there are multiple attributes having the greatest importance degree, then compare the information entropy of these attributes and select the attribute with the minimum entropy (carrying the least uncertain information) to be incorporated into the reduction set, so that we can get better attribute reduction results. This improved attribute reduction algorithm is shown as Algorithm 1.

Algorithm 1: Neighborhood attribute reduction algorithm based on information entropy

Input: Neighborhood decision system $NDT = \langle U, A, D \rangle$.

Output: The reduced attribute set *red*.

Step 1: $\forall a \in A$, calculate the neighborhood relation N_a.

Step 2: Initialize $red = \varnothing$.

Step 3: $\forall a_i \in A - red$, calculate the importance degree of each attribute $SIG(a_i, red, D) = \gamma_{red \cup a}(D) - \gamma_{red}(D)$.

Step 4: If $\max_i(SIG(a_i, red, D))$ contains only one attribute, then select a_k so as to satisfy $SIG(a_k, red, D) = \max_i(SIG(a_i, red, D))$; otherwise, calculate the information entropy of these attributes and select a_k so as to satisfy $H(a_k) = \min_i(H(a_i))$.

Step 5: If $SIG(a_k, red, D) > 0$, put a_k into the reduction set, $red = red \cup a_k$, then go to step 3; otherwise, output *red* and the algorithm ends.

III. Spectral Clustering Algorithm based on Neighborhood Rough Sets Reduction

Spectral clustering uses the idea of graph partitioning to solve data clustering problem. Given a data set $X = \{x_1, x_2, \cdots, x_n\}$ containing *n* data points, $x_i \in R^l$. With the points of set *X*, we can construct an undirected weighted graph *G*(*V*,*E*), in which every point $x_i \in X$ being a vertex of *G* and the affinity value w_{ij} between each pair of points (x_i,x_j) being the edge weight. All the weight values constitute the affinity matrix $W \in R^{n \times n}$. The similarity between data points can be measured by angle cosine, correlation coefficient or Gaussian kernel function. Gaussian kernel function is more common in spectral clustering and its expression is:

$$w_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{2\sigma^2}\right) \tag{9}$$

where $d(x_i,x_j)$ is a distance function between two points usually described by Euclidean distance; σ is the scale parameter controlling the speed that weight value falls off with $d(x_i,x_j)$.

Since a single parameter σ is not easy to deal with multi-scale clustering problem, Manor et al. [25] proposed a self-tuning spectral clustering algorithm. Instead of using a uniform σ, this algorithm calculates an adaptive parameter σ_i for each point x_i according to their neighborhood information, where σ_i is the Euclidean distance from point x_i to its *p*-th nearest neighbor. This similarity measure is called self-tuning Gaussian kernel function, which is described by the following formula:

$$w_{ij} = \exp\left(-\frac{d^2(x_i, x_j)}{\sigma_i \sigma_j}\right) \tag{10}$$

where $\sigma_i = \frac{1}{p}\sum_{k=1}^{p} d(x_i, x_k)$. As the neighborhood information of each point is considered, self-adjusting spectral clustering can effectively separate clusters from sparse background (as shown in Figure 2). So in this paper we use self-tuning similarity measure to calculate the affinities of data points.

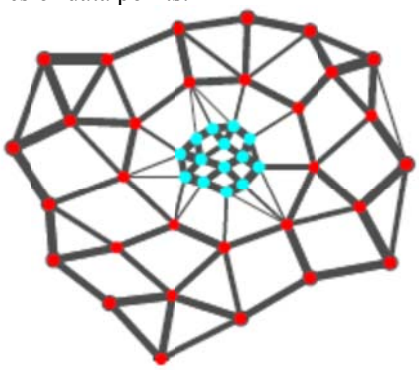


Figure 2. Local adaptive similarity measure.

Through the edge weights we can compute the degree of each vertex to compose the degree matrix $D \in R^{n\times n}$ of graph G. The degree of vertex i is denoted by d_i, which is the total weights of the edges connected to i, shown as formula (11). Thus D is a diagonal matrix with d_i being the diagonal elements and 0 being other elements.

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{11}$$

According to the similarities between data points, we can create the similarity graph to describe samples points. Then the goal of clustering becomes finding an optimal partition of the graph, so that the internal nodes of the same sub-graph have greater connecting weight, while the connecting weights between different sub-graphs are much smaller. Many different graph cut methods have been proposed, for example Minimum cut, Ratio cut, Normalized cut and MinMax cut, etc. These methods typically define a number of indicate vectors to describe the attribution of data points. Given a subset of the graph, a data point is either in this subset or not, which is a discrete optimization problem. Theoretically speaking, the desired clustering results can be obtained by minimizing or maximizing the objective function of graph cut method. However, it has been proved that finding the best possible partition of a graph is an NP-hard problem [26]. With the help of spectral method, the original problem can be solved in polynomial time, and get an approximate optimal solution. Spectral method relaxes the original discrete attribution degree constraint to the real number field, so that the indicate vector may get the values in a continuous range. Then according to Rayleigh-Ritz theory, make eigen-decomposition of the Laplacian matrix, and use the resulting eigenvector to segment sub-graphs [27]. Alternatively, spectral clustering can be also interpreted from the view of random walk [28]. In random walk theory, each point on the graph may jump to the points adjacent to it in a certain probability. A probability transition matrix is composed of these jump probabilities. Dividing a graph into several sub-graphs should try to keep the random walk process among the vertices within the same sub-graph in a large probability and rarely jumping from one sub-graph to another.

Massive high-dimensional data processing has been a problem in data mining. High-dimensional data is often accompanied by the "curse of dimensionality", so traditional spectral clustering algorithms can not play to their strengths very well. Moreover, real data sets often contain noise and irrelevant features, likely to cause “dimension trap”. It would interfere with the clustering process of algorithms, affecting the accuracy of clustering results [29]. To solve this problem, we propose a spectral clustering algorithm based on neighborhood rough sets reduction (NRSR-SC). Its basic idea is: first, according to the concept of neighborhood rough set compute the importance degree of each attribute; then make reduction to the attributes based on attribute importance, removing redundant attributes and retaining the most important attribute, under the premise of ensuring the distinguish ability of the system; next measure the similarities between data points to construct the similarity matrix and Laplacian matrix; map the points of the original data set to R^k space with the largest k eigenvectors of Laplacian matrix; at last, clustering the data points in the R^k space by k-means or other conventional algorithm. The process of NRSR-SC algorithm is described in Algorithm 2.

Algorithm 2: Spectral clustering algorithm based on neighborhood rough sets reduction

Input: Data set $X = \{x_1, x_2, \cdots, x_n\}$, the number of clusters k.

Output: k divided clusters.

Step 1: Reduce the attributes of sample points according to Algorithm 1 and obtain the reduced attribute set *red*.

Step 2: On the basis of the reduction attribute set *red*, form the affinity matrix $W \in R^{n\times n}$ by Equation (10) and establish the degree matrix $D \in R^{n\times n}$ of the graph by Equation (11).

Step 3: Using the affinity matrix W and the degree matrix D to construct the Laplacian matrix $L_{sym} = D^{-1/2}(D-W)D^{-1/2}$.

Step 4: Make eigen-decomposition for Laplacian matrix L_{sym}, choosing its largest k eigenvalues and their corresponding eigenvectors $u_1, \cdots, u_k$ to form matrix $U = [u_1 \vdots \cdots \vdots u_k] \in R^{n\times k}$ by arranging these eigenvectors vertically.

Step 5: Convert each row vector of matrix U into a unit vector through normalization and get matrix Y: $y_{ij} = u_{ij} / \left[\sum_{j=1}^{n} u_{ij}^2\right]^{1/2}$.

Step 6: Every row of matrix Y can be seen as a new point in R^k space. Then we can use a certain clustering algorithm such as k-means to group these points.

Step 7: Establish a one to one relationship between the original data points and each row vector of matrix Y, and

determine the attribution of each data point in accordance with the clustering results.

IV. Experimental Analysis

A. Clustering Evaluation Method

To test the effectiveness of NRSR-SC algorithm, we select five high dimensional data sets from the UCI machine learning repository for experiments. The description of these data sets is shown in Table I.

TABLE I.
DATA SETS DESCRIPTION

Data set	Data Characteristic			
	Sample number	Condition attribute number	Decision attribute number	Class number
Ionosphere	351	34	1	2
Sonar	208	60	1	2
Small soybean	47	35	1	4
WDBC	569	30	1	2
Zoo	101	16	1	7

The merits of clustering results can be measured by many indicators. F-measure is an evaluation index based on artificial annotation, derived from information retrieval field. F-measure contains precision rate and recall rate. These two indicators describe the difference between the clustering results and the real classes from different angles. F-score calculated by the precision and recall rate, is a comprehensive index to evaluate a cluster. It gives an objective evaluation to the generated clusters [30]. Suppose there are k classes in the data set, and class i is associated with cluster i^* in clustering results. We can compute the F-score of class i using the following three formulas:

$$P(i) = N_{ii^*} / N_{i^*} \tag{12}$$

$$R(i) = N_{ii^*} / N_i \tag{13}$$

$$F(i) = \frac{2 \times P(i) \times R(i)}{P(i) + R(i)} \tag{14}$$

where $P(i)$ and $R(i)$ are respectively the precision rate and the recall rate; N_{ii^*} is the size of the intersection of class i and cluster i^*; N_i is the size of class i; N_{i^*} is the size of cluster i^*.

The total F index of the clustering results is the weighted average of each class's F-score:

$$F = \frac{1}{n} \sum_{i=1}^{k} [N_i \times F(i)] \tag{15}$$

where n is the number of sample points; k is the class number of data set; N_i is the size of class i. $F \in [0,1]$, the greater the F index is, means the clustering results of the algorithm is closer to the real data category.

B. Experimental Results and Discussion

Using the real world data sets, NRSR-SC algorithm is compared with the traditional spectral clustering (SC) and FCM algorithm in the experiment. The clustering results of these three algorithms on each data set are shown in Figure 3 to Figure 7. The horizontal axis of the figure is the class label of obtained clusters, and the vertical axis is their F-scores.

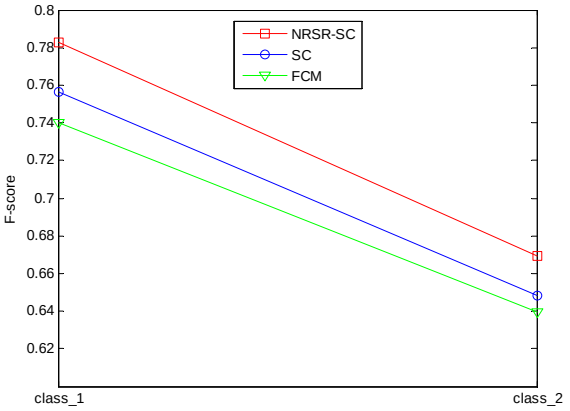


Figure 3. Clustering results on Ionosphere dataset.

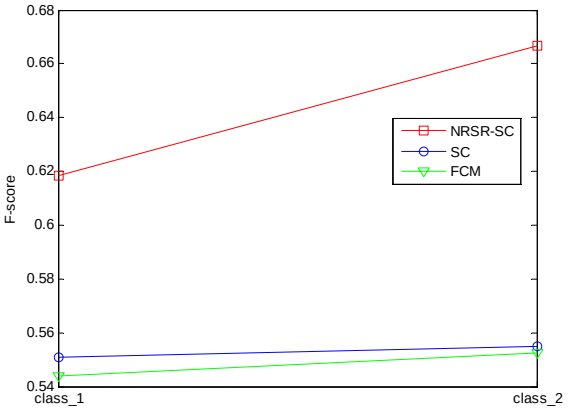


Figure 4. Clustering results on Sonar dataset.

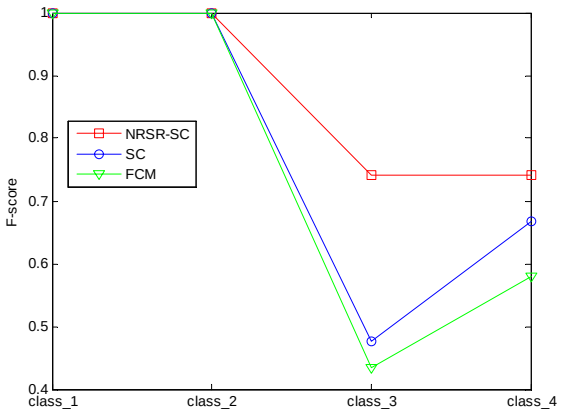


Figure 5. Clustering results on Small soybean dataset.

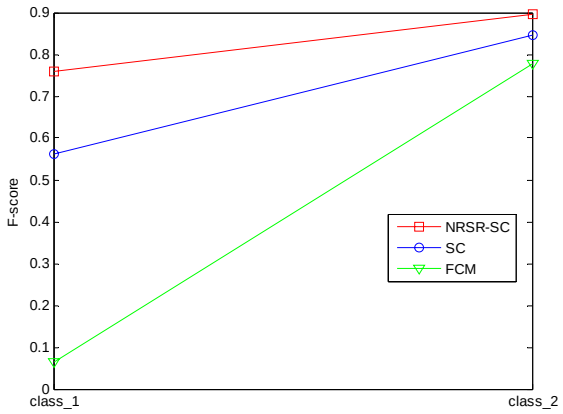


Figure 6. Clustering results on WDBC dataset.

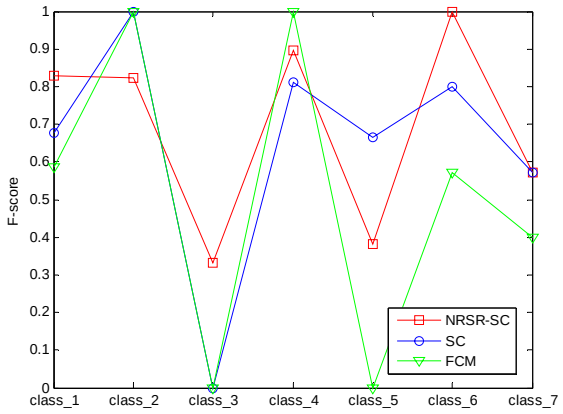


Figure 7. Clustering results on Zoo dataset.

In the above figures, the F-score of FCM algorithm is relatively low. This is mainly due to that in FCM algorithm, the initialization of cluster centers is a random process. Improper cluster centers would cause a greater impact on the clustering results. Moreover, FCM algorithm does work well with multi-scale clustering problem, easily dropping into local optimum. As spectral clustering algorithm, the indexes of NRSR-SC and SC algorithm are much higher. They turn clustering problem into graph partitioning problem. Using the spectral method and Laplace transform, the algorithm may find the global optimum solution, resulting in a more balanced class division. But the curve of SC algorithm is slightly below the proposed NRSR-SC algorithm. Because the information in each attribute of the samples is different, and they also make different contributions to clustering. Traditional spectral clustering algorithm does not take this into account, susceptible to the interference of noise and irrelevant attributes, so the clustering performance is not very well. For further comparison, the number of condition attributes as well as the overall F index of each data set processed by NRSR-SC algorithm, SC algorithm and FCM algorithm are given below, as shown in Table II and Table III.

TABLE II.
CONDITION ATTRIBUTE NUMBER OF DATA SETS

Data set	Number of condition attributes	
	NRSR-SC	SC and FCM
Ionosphere	6	34
Sonar	8	60
Small soybean	2	35
WDBC	7	30
Zoo	5	16

TABLE III.
OVERALL F INDEX OF ALGORITHMS

Data set	Algorithm		
	NRSR-SC	SC	FCM
Ionosphere	0.7421	0.7177	0.7040
Sonar	0.6442	0.5530	0.5483
Small soybean	0.8511	0.7680	0.7281
WDBC	0.8443	0.7411	0.5113
Zoo	0.7822	0.7239	0.6496

From the experimental data in Table 2 and Table 3 we can see that compared with FCM algorithm and conventional spectral clustering algorithm, NRSR-SC algorithm can better handle high-dimensional data, and generate more accurate clustering results. NRSR-SC algorithm uses neighborhood rough sets to optimize data samples, removes non-related properties and retains the attributes contributing the most to clustering, so that the sample points within the same cluster are more compact, while the sample points of different clusters are more separate, thus improving the accuracy of clustering. The neighborhood attribute reduction based on information entropy diminishes the negative impact of noise data and redundant attributes on the clustering. Not only reduce the complexity of problem solving, but also can better describe the approximate relationship between data points, and play the advantages of spectral clustering algorithm to enhance its robustness and generalization ability.

V. CONCLUSION

Spectral clustering is based on algebraic graph theory and can effectively solve many practical problems. However, suffering from the interference of noise and irrelevant attributes, traditional spectral clustering algorithm does not work well on high-dimensional data. In order to reduce the computational complexity and weaken the negative impact of noise data and redundant attributes on clustering, this paper propose a spectral clustering algorithm based on neighborhood rough sets reduction (NRSR-SC). Information entropy is introduced into the neighborhood rough sets in this algorithm, so that redundant attributes can be removed and the attributes making the greatest contribution to clustering can be reserved, under the premise of maintaining the ability to

distinguish different kind of samples. Then, based on the reduced attribute collection, the similarities between sample points are calculated to construct the affinity matrix and Laplacian matrix. At last, use spectral method to get the final clustering results. Experiments show that, when dealing with high-dimensional data, NRSR-SC algorithm has a strong anti-jamming ability and the efficiency and accuracy has improved significantly. In the future, we will study how to use NRSR-SC algorithm to process large data problems, and consider its application to web data mining, image retrieval and other fields.

REFERENCES

[1] Sun JG, Liu J, Zhao LY. Clustering Algorithms Research. Journal of Software, 2008, 19(1): 48-61.

[2] Jiang HF, Gu JF, Liu YJ, et al. Study of Clustering Algorithm based on Fuzzy C-Means and Immunological Partheno Genetic. Journal of Software, 2013, 8(1): 134-141.

[3] Nascimento MCV, de Carvalho ACPLF. Spectral methods for graph clustering - A survey. European Journal of Operational Research, 2011, 211(2): 221-231.

[4] Blekas K, Lagaris IE. A Spectral Clustering Approach Based on Newton's Equations of Motion. International Journal of Intelligent Systems, 2013, 28(4): 394-410.

[5] Li XY, Guo LJ. Constructing affinity matrix in spectral clustering based on neighbor propagation. Neurocomputing, 2012, 97: 125-130.

[6] Frederix K, Van Barel M. Sparse spectral clustering method based on the incomplete Cholesky decomposition. Journal of Computational and Applied Mathematics, 2013, 237(1): 145-161.

[7] Wang LJ, Dong M. Multi-level Low-rank Approximation-based Spectral Clustering for image segmentation. Pattern Recognition Letters, 2012, 33(16): 2206-2215.

[8] Jiao LC, Shang FH, Wang F, Liu YY. Fast semi-supervised clustering with enhanced spectral embedding. Pattern Recognition, 2012, 45(12): 4358-4369.

[9] Michoel T, Nachtergaele B. Alignment and integration of complex networks by hypergraph-based spectral clustering. Physical Review E, 2012, 86(5): 056111.

[10] Adefioye AA, Liu XH, Moor BD. Multi-view spectral clustering and its chemical application. International Journal of Computational Biology and Drug Design, 2013, 6(1-2): 32-49.

[11] Ding L, Gonzalez-Longatt FM, Wall P, Terzija V. Two-Step Spectral Clustering Controlled Islanding Algorithm. IEEE Transactions on Power Systems, 2013, 28(1): 75-84.

[12] Li H, Ding SF. Research and Development of Granular Neural Networks. Applied Mathematics & Information Sciences, 2013, 7(3): 1251-1261.

[13] Xu L, Ding SF, Guo FF. A Rough Kernel Clustering Algorithm Based on Improved Attribute Reduction. Journal of Guangxi Normal University (Natural Science Edition), 2011, 29(3): 105-109.

[14] Li XW. A New Text Clustering Algorithm Based on Improved K-means. Journal of Software, 2012, 7(1): 95-101.

[15] Zhu H, Ding SF, Zhao H, et al. Attribute Granulation Based on Attribute Discernibility and AP Algorithm. Journal of Software, 2013, 8(4): 834-841.

[16] Pawlak Z. Rough sets [J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[17] Hu QH, Yu DR, Xie ZX. Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation. Journal of Software, 2008, 19(3): 640-649.

[18] Hu QH, Zhao H, Yu DR. Efficient Symbolic and Numerical Attribute Reduction with Neighborhood Rough Sets. Pattern Recognition and Artificial Intelligence, 2008, 21(6): 732-738.

[19] Wang GY, Yu H, Yang DC. Decision Table Reduction based on Conditional Information Entropy. Chinese Journal of Computers, 2002, 25(7): 759-766.

[20] Ding SZ, Sang L, Zhu QY, et al. Attribute reduction in rough set based on information entropy and it's application. Computer Engineering and Applications, 2007, 43(35): 245-248.

[21] Wu SZ, Gou PZ. Attribute Reduction Algorithm on Rough Set and Information Entropy and Its Application. Computer Engineering, 2011, 37(7): 56-61.

[22] Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal, 1948, 27: 379-423, 623-656.

[23] Ding SF, Zhu H, Xu XZ, Shi ZZ. Entropy-Based Fuzzy Information Measures. Chinese Journal of Computers, 2012, 35(4): 796-801.

[24] Jia WK, Ding SF, Xu XZ, et al. Factor Analysis Feature Extraction Algorithm Based on Shannon Entropy. Pattern Recognition and Artificial Intelligence, 2011, 24(3): 327-331.

[25] Zelnik-Manor L, Perona P. Self-tuning spectral clustering. In Proceeding of NIPS, 2005: 1601-1608.

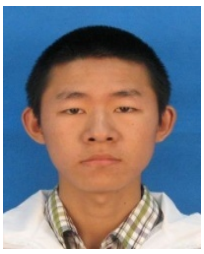
[26] Li JY, Zhou JG, Guan JH, et al. A survey of clustering algorithms based on spectra of graphs. CAAI Transactions on Intelligent Systems, 2011, 6(5): 405-414.

[27] Jia HJ, Ding SF, Zhu H, et al. A feature weighted spectral clustering algorithm based on knowledge entropy. Journal of Software, 2013, 8(5): 1101-1108.

[28] Zhang XC, You QZ. An improved spectral clustering algorithm based on random walk. Frontiers of Computer Science in China, 2011, 5(3): 268-278.

[29] Wang XZ, Wang TT, Zhai JH. An Attribute Reduction Algorithm Based on Instance Selection. Journal of Computer Research and Development, 2012, 49(11): 2305-2310.

[30] Tao XM, Song SY, Cao PD, et al. A Spectral Clustering Algorithm Based on Manifold Distance Kernel. Information and Control, 2012, 41(3): 307-313.



Hongjie Jia is currently a graduate student now studying at School of Computer Science and Technology, China University of Mining and Technology, and his supervisor is Prof. Shifei Ding. He received his B.Sc. degree in computer science from China University of Mining and Technology in 2011. His research interests include pattern recognition, machine learning, and spectral clustering et al.

Shifei Ding is a professor and Ph.D. supervisor at China University of Mining and Technology. His research interests include intelligent information processing, pattern recognition, machine learning, data mining, and granular computing et al. He has published 3 books, and more than 150 research papers in journals and international conferences. He received his B.Sc., M.Sc. degree in

mathematics and computer science in 1987, 1998 respectively from Qufu Normal University, received Ph.D. degree in computer science from Shandong University of Science and Technology in 2004, and received post Ph.D. degree in computer science from Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, and Chinese Academy of Sciences in 2006.

Heng Ma is currently a graduate student now studying in School of Computer Science and Technology, China University of Mining and Technology, and his supervisor is Prof. Shifei Ding. He received her B.Sc. degree in computer science from China University of Mining and Technology in 2013. His research interests include cloud computing, feature selection, pattern recognition, machine learning et al.

Wanqiu Xing is currently a graduate student now studying in School of Computer Science and Technology, China University of Mining and Technology, and her supervisor is Prof. Shifei Ding. She received her B.Sc. degree in computer science from China University of Mining and Technology in 2012. His research interests include cloud computing, feature selection, pattern recognition, machine learning et al.