p-Spectral Clustering Based on Neighborhood Attribute Granulation

Shifei Ding, Hongjie Jia, Mingjing Du, Qiankun Hu

p-Spectral Clustering Based on Neighborhood Attribute Granulation

Shifei Ding[1,2], Hongjie Jia[1,2], Mingjing Du[1,2], Qiankun Hu[1,2]

1 School of Computer Science and Technology, China University of Mining and Technology,
Xuzhou 221116, China
2 Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences
Beijing 100190, China
{dingsf, jiahongjie}@cumt.edu.cn

Abstract. Clustering analysis is an important method for data mining and information statistics. Data clustering is to find the intrinsic links between objects and describe the internal structures of data sets. *p*-Spectral clustering is based on Cheeger cut criterion. It has good performance on many challenging data sets. But the original *p*-spectral clustering algorithm is not suitable for high-dimensional data. To solve this problem, this paper improves *p*-spectral clustering using neighborhood attribute granulation and proposes NAG-pSC algorithm. Neighborhood rough sets can directly process the continuous data. We introduce information entropy into the neighborhood rough sets to weaken the negative impact of noise data and redundant attributes on clustering. In this way, the data points within the same cluster are more compact, while the data points between different clusters are more separate. The effectiveness of the proposed NAG-pSC algorithm is tested on several benchmark data sets. Experiments show that the neighborhood attribute granulation will highlight the differences between data points while maintaining their characteristics in the clustering. With the help of neighborhood attribute granulation, NAG-pSC is able to recognize more complex data structures and has strong robustness to the noise or irrelevant features in high-dimensional data.

Keywords: *p*-spectral clustering, rough set, attribute reduction, neighborhood granulation

1 Introduction

Spectral clustering treats clustering problem as a graph partitioning problem. It can solve the graph cut objective function using the eigenvectors of graph Laplacian matrix [1]. Compared with the conventional clustering algorithms, spectral clustering is able to recognize more complex data structures, especially suitable for non-convex data sets. Recently, an improved version of normalized cut named Cheeger cut has aroused much attention [2]. Research shows that Cheeger cut is able to produce more balanced clusters through graph *p*-Laplacian matrix [3]. *p*-Laplacian matrix is a nonlinear generalization form of graph Laplacian.

 p-spectral clustering is based on Cheeger cut to group data points. As it has solid theoretical foundation and good clustering results, the research in this area is very active at present. Dhanjal et al. present an incremental spectral clustering which

updates the eigenvectors of the Laplacian in a computationally efficient way [4]. Gao et al. construct the sparse affinity graph on a small representative dataset and use local interpolation to improve the extension of the clustering results [5]. Semertzidis et al. inject the pairwise constraints to a small affinity sub-matrix and use a sparse coding strategy of a landmark spectral clustering to preserve low complexity [6].

Nowadays, science and technology is growing by leaps and bounds and massive data result in "data explosion". These data are often accompanied by high dimensions. 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 [7]. 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. 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. In order to effectively deal with high-dimensional data, we design a novel attribute reduction method based on neighborhood granulation and combine it with p-spectral clustering. The proposed algorithm inherits the advantages of neighborhood rough set and graph p-Laplacian. Its effectiveness is demonstrated by comprehensive experiments on benchmark data sets.

This paper is organized as follows: section 2 introduces p-spectral clustering; section 3 uses information entropy to improve the attribute reduction based on neighborhood rough sets; section 4 improves p-spectral clustering with the neighborhood attribute granulation; section 5 verifies the effectiveness of the proposed algorithm using benchmark data sets; finally, we summarize the main contribution of this paper.

2 p-Spectral Clustering

The idea of spectral clustering comes from spectral graph partition theory. Given a data set, we can construct an undirected weighted graph $G=(V,E)$, where V is the set of vertices represented by data points, E is the set of edges weighted by the similarities between the edge's two vertices. Suppose A is a subset of V, the complement of A is written as $\bar{A} = V \setminus A$. The cut of A and $\bar{A}$ is defined as:

$$cut(A, \bar{A}) = \sum_{i \in A, j \in \bar{A}} w_{ij} \tag{1}$$

where w_{ij} is the similarity between vertex i and vertex j.

In order to get more balanced clusters, Cheeger et al. propose Cheeger cut criterion, denoted as $Ccut$ [8]:

$$Ccut(A, \bar{A}) = \frac{cut(A, \bar{A})}{\min\{|A|, |\bar{A}|\}} \tag{2}$$

where $|A|$ is the number of data points in set A. Cheeger cut is to minimize formula (2) to get a graph partition. The optimal graph partition means that the similarities within a cluster are as large as possible, while the similarities between clusters are as small as possible. But according to the Rayleigh quotient principle, calculating the optimal Cheeger cut is an NP-hard problem. Next we will try to get an approximate solution of Cheeger cut by introducing p-Laplacian into spectral clustering.

Hein et al. define the inner product form of graph p-Laplacian Δ_p as follows [9]:

$$\langle f, \Delta_p f \rangle = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^p \tag{3}$$

where $p \in (1,2]$, f is the eigenvector of p-Laplacian matrix.

Theorem 1. For $p > 1$ and every partition of V into $A, \bar{A}$ there exists a function (f,A) such that the functional F_p associated to the p-Laplacian satisfies

$$F_p(f,A) = \frac{\langle f, \Delta_p f \rangle}{\|f\|^p} = cut(A, \bar{A}) \left| \frac{1}{|A|^{\frac{1}{p-1}}} + \frac{1}{|\bar{A}|^{\frac{1}{p-1}}} \right|^{p-1} \tag{4}$$

where $\|f\|^p = \sum_{i=1}^{n} |f_i|^p$. The expression (4) can be interpreted as a balanced graph cut criterion, and we have the special cases

$$\lim_{p \to 1} F_p(f,A) = Ccut(A, \bar{A}) \tag{5}$$

Theorem 1 shows that Cheeger cut can be solved in polynomial time using p-Laplacian operator. So the solution of $F_p(f)$ is a relaxed approximate solution of Cheeger cut and the optimal solution can be obtained by the eigen-decomposition of p-Laplacian:

$$\lambda_p = \arg \min_{p \to 1} F_p(f) \tag{6}$$

where λ_p is the eigenvalue corresponding to eigenvector f.

Specifically, the second eigenvector $v_p^{(2)}$ of p-Laplacian matrix will lead to a bipartition of the graph by setting an appropriate threshold [3]. The optimal threshold is determined by minimizing the corresponding Cheeger cut. For the second eigenvector $v_p^{(2)}$ of graph p-Laplacian Δ_p, the threshold should satisfy:

$$\arg \min_{A_t = \{i \in V | v_p^{(2)}(i) > t\}} Ccut(A_t, \bar{A}_t) \tag{7}$$

3 Neighborhood Attribute Granulation

Rough set theory is proposed by professor Pawlak in 1982 [10]. Attribute reduction is one of the core contents of rough set knowledge discovery. However, Pawlak rough set is only suitable for discrete data. To solve this problem, Hu et al. propose neighborhood rough set model [11]. This model can directly analyze the attributes with continuous values. Therefore, it has great advantages in feature selection and classification accuracy.

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:

$$\delta(x_i) = \{x \mid x \in U, \Delta(x, x_i) \leq \delta\} \tag{8}$$

where $\delta \geq 0$, $\delta(x_i)$ is called the neighborhood particle of x_i, Δ is a distance function.

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 is $\underline{N_B}D = \bigcup_{i=1}^{N} \underline{N_B}X_i$, where $\underline{N_B}X_i = \{x_i \mid \delta_B(x_i) \subseteq X_i, x_i \in U\}$.

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

$$\gamma_B(D) = \frac{Card(\underline{N_B}D)}{Card(U)} \tag{9}$$

where $0 \le \gamma_B(D) \le 1$. Obviously, the greater the positive region $\underline{N_B}D$, the stronger the dependence of decision D on condition B.

Definition 3. 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:

$$SIG(a, B, D) = \gamma_{B \cup a}(D) - \gamma_B(D) \tag{10}$$

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 [12]. Here, we use information entropy as another criterion to evaluate attributes. The definition of entropy is given below.

Definition 4. 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:

$$H(P) = -\sum_{i=1}^{n} p(X_i) \log p(X_i) \tag{11}$$

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

If multiple attributes have the same greatest importance degree, then we may compare their information entropy and select the attribute with the minimum entropy (because it carries the least uncertain information). Incorporate the selected attribute into the reduction set, and repeat this process for each attribute until the reduction set no longer changes. This improved attribute reduction algorithm is shown as Algorithm 1.

Algorithm 1. Neighborhood attribute granulation with 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.

4　*p*-Spectral Clustering Based on Neighborhood Attribute Granulation

Massive high-dimensional data processing has been a challenge problem in data mining. High-dimensional data is often accompanied by the "curse of dimensionality", so traditional *p*-spectral clustering algorithms cannot 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 [13]. To solve this problem, we propose a novel *p*-spectral clustering algorithm based on neighborhood attribute granulation (NAG-pSC). The detailed steps of NAG-pSC algorithm is given in Algorithm 2.

Algorithm 2. *p*-Spectral clustering algorithm based on neighborhood attribute granulation.

Input: dataset $X = \{x_1, x_2, \cdots, x_n\}$, the cluster number *k*.

Output: *k* divided clusters.

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

Step 2. Affter attribute granulation, calculate the similarities between data points based on the new data set *red*, and form the affinity matrix $W \in \mathbb{R}^{n \times n}$ using Gaussian kernel.

Step 3. Initialize the first cluster $A_1 = V$ and set the cluster number $s = 1$.

Step 4. Repeat from Step 4 to Step 8.

Step 5. Construct *p*-Laplacian matrix according to formula [3] with the affinity matrix *W*.

Step 6. Calculate the second eigenvector $v_p^{(2)}$ of graph *p*-Laplacian Δ_p, and search an appropriate threshold value that satisfies formula [7].

Step 7. Use $v_p^{(2)}$ to split each cluster $A_i (i = 1, 2, \cdots, s)$ and minimize the overall Cheeger cut objective function.

Step 8. $s \Leftarrow s + 1$.

Step 9. When the number of clusters $s == k$, stop the loop and output the clustering results.

5 Experimental Analysis

To test the effectiveness of the proposed NAG-pSC algorithm, we use six benchmark data sets to do the experiments. The characteristics of these data sets are shown in Table 1.

Table 1. Data sets used in the experiments

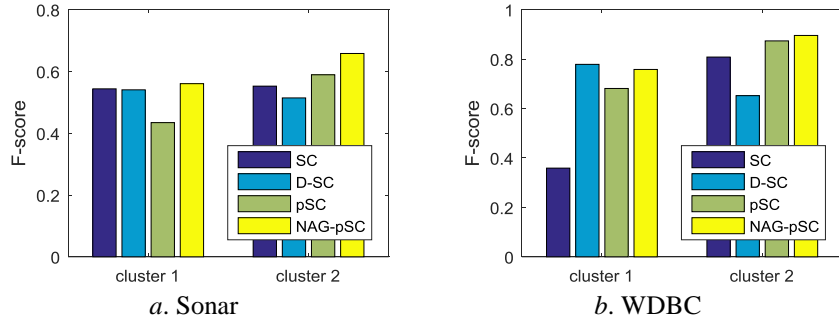| Data set	Instance number	Condition attribute number	Decision attribute number	Class number
Sonar	208	60	1	2
WDBC	569	30	1	2
Colon Cancer	62	2000	1	2
Duke Breast Cancer	44	7129	1	2

In this paper, we use F-measure to evaluate the merits of clustering results [14]. The F-score of each class i and the total F index of the clustering results are defined as:

$$F(i) = \frac{2 \times P(i) \times R(i)}{P(i) + R(i)} \tag{12}$$

$$F = \frac{1}{n} \sum_{i=1}^{k} [N_i \times F(i)] \tag{13}$$

where $P(i) = N_{ii*} / N_{i*}$ is the precision rate and $R(i) = N_{ii*} / N_i$ is 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 $i*$; n is the number of data points; k is the class number; 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.

In the experiment, NAG-pSC algorithm is compared with the traditional spectral clustering (SC), density sensitive spectral clustering (D-SC) [1] and p-spectral clustering (pSC) [3]. The threshold δ is important in neighborhood rough set. Hu et al. recommend a value range [0.2, 0.4] of δ based on experimental analysis [11]. So we set the neighborhood size δ via a cross-validatory search in the range [0.2, 0.4] (with step size 0.05) for each data set. The clustering results of these four algorithms are shown in Figure 1. The horizontal axis of the figure is the cluster label, and the vertical axis is the F-score of each cluster.
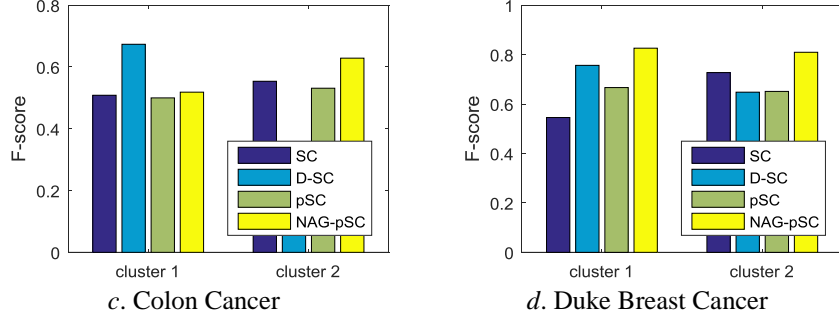

a. Sonar

b. WDBC

c. Colon Cancer   *d*. Duke Breast Cancer

Figure 1. Clustering results on different datasets

From Figure 1 we can see that, the performance of SC algorithm is close to D-SC algorithm. This is mainly because that they all based on graph theory and turn the clustering problem into a graph partitioning problem. Using the *p*-Laplacian transform, pSC may find the global optimum solution. SC works well on Sonar data set. D-SC deals well with Colon Cancer data set. pSC can generate balanced clusters on WDBC data set. But for high dimensional clustering problems, their F-scores are lower than the proposed NAG-pSC algorithm. Because the information in each attribute of the instances is different, and they also make different contributions to the clustering. Improper feature selection would cause a greate impact on the clustering results. Traditional spectral clustering algorithm does not take this into account, susceptible to the interference of noise and irrelevant attributes. For further comparison, Table 2 lists the overall *F* index for each algorithm and the number of condition attributes of different data sets.

Table 2. Total *F* index of different algorithms

Data set		Algorithm			
		SC	D-SC	pSC	NAG-pSC
Sonar	*F* index	0.5483	0.5314	0.5170	0.6126
	Condition attributes	60	60	60	17
WDBC	*F* index	0.6408	0.7331	0.8019	0.8443
	Condition attributes	30	30	30	13
Colon Cancer	*F* index	0.5377	0.5453	0.5202	0.5895
	Condition attributes	2000	2000	2000	124
Duke Breast Cancer	*F* index	0.6322	0.7177	0.6593	0.8182
	Condition attributes	7129	7129	7129	167

Table 2 shows that NAG-pSC algorithm can well deal with high-dimensional data. NAG-pSC algorithm uses neighborhood rough sets to optimize data instances. The neighborhood attribute reduction based on information entropy diminishes the negative impact of noise data and redundant attributes on the clustering. So in most cases, NAG-

pSC algorithm has higher clustering accuracy. NAG-pSC algorithm combines the advantages of *p*-spectral clustering and neighborhood attribute granulation. It has good robustness and strong generalization ability.

6   Conclusions

To improve the performance of *p*-spectral clustering on high-dimensional data, we modify the attribute reduction method based on neighborhood rough sets. In the new method, the attribute importance is combined with information entropy to select the appropriate attributes. Then we propose NAG-pSC algorithm based on the optimized attribute reduction set. Experiments show that NAG-pSC algorithm is superior to traditional spectral clustering, density sensitive spectral clustering and *p*-spectral clustering. In the future, we will study how to apply NAG-pSC algorithm to web data mining, image retrieval and other realistic scenes.

References

1. Yang, P., Zhu, Q., Huang, B.: Spectral clustering with density sensitive similarity function. Knowledge-Based Systems, 24(5): 621–628 (2011)
2. Bresson, X., Szlam, A.D.: Total Variation and Cheeger Cuts. Proceedings of the 27th International Conference on Machine Learning, pp.1039–1046 (2010)
3. Bühler, T., Hein, M.: Spectral Clustering Based on the Graph p-Laplacian. Proceedings of the 26th International Conference on Machine Learning, pp.81–88 (2009)
4. Dhanjal, C., Gaudel, R., Clémençon, S.: Efficient Eigen-Updating for Spectral Graph Clustering. Neurocomputing, 131: 440–452 (2014)
5. Cao, J., Chen, P., Dai, Q., et al.: Local Information-Based Fast Approximate Spectral Clustering. Pattern Recognition Letters, 38: 63–69 (2014)
6. Semertzidis, T., Rafailidis, D., Strintzis, M.G., et al.: Large-Scale Spectral Clustering Based on Pairwise Constraints. Information Processing & Management, 51: 616–624 (2015)
7. Jia, H., Ding, S., Xu, X., et al.: The Latest Research Progress on Spectral Clustering. Neural Computing and Applications, 24(7–8): 1477–1486 (2014)
8. Jia, H., Ding, S., Du, M.: Self-Tuning p-Spectral Clustering Based on Shared Nearest Neighbors. Cognitive Computation, 7(5): 622–632 (2015)
9. Hein, M., Audibert, J.Y., Von Luxburg, U.: Graph Laplacians and Their Convergence on Random Neighborhood Graphs. Journal of Machine Learning Research, 8(12): 1325–1368 (2007)
10. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences, 11(5): 341–356 (1982)
11. Hu, Q.H., Yu, D.R., Xie, Z.X.: Numerical Attribute Reduction Based on Neighborhood Granulation and Rough Approximation. Journal of Software,

19(3): 640–649 (2008)
12. Ding, S.F., Zhu, H., Xu, X.Z., et al.: Entropy-Based Fuzzy Information Measures. Chinese Journal of Computers, 35(4): 796–801 (2012)
13. Du, M., Ding, S., Jia, H.: Study on Density Peaks Clustering Based on K-Nearest Neighbors and Principal Component Analysis. Knowledge-Based Systems, 99: 135–145 (2016)
14. Jia, H., Ding, S., Zhu, H., et al.: A Feature Weighted Spectral Clustering Algorithm Based on Knowledge Entropy. Journal of Software, 8(5): 1101–1108 (2013)