Research on Spectral Clustering Algorithms and Prospects

Shifei Ding
School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, China;
Key Laboratory of Intelligent Information Processing
Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
e-mail: dingsf@cumt.edu.cn

Liwen Zhang, Yu Zhang
School of Computer Science and Technology
China University of Mining and Technology
Xuzhou, China
e-mail: ji.wenji@163.com
e-mail: zhang_yu1015@hotmail.com

Abstract—Along with the expansion and in-depth of the application domain of cluster analysis, one kind of new cluster algorithm called Spectral Clustering algorithm has been aroused great concern by scholars, Spectral Clustering algorithm is newly developing technique in the field of machine learning in recent years. Unlike the traditional clustering algorithms, this can solve the clustering of non-convex sphere of sample spaces and has globally optimal solution. This paper introduces the principle, the induction summary to the current research situation of Spectral Clustering algorithm as well as in various application domains. Firstly, the analysis and induction of some Spectral Clustering algorithms have been made from several aspects, such as the ideas of algorithm, key technology, advantage and disadvantage. On the other hand, some typical Spectral Clustering algorithms have been selected to analyze and compare. Finally, it points out the key problems and future directions.

Keywords—cluster analysis; spectral clustering; Laplacian Matrix; graph partition; eigenvalue

I. INTRODUCTION

So far, clustering analysis has not yet accepted definition in academic community. Spectral clustering refers to a class of techniques which rely on the eigenstructure of a similarity matrix to partition points into disjoint clusters with points in the same cluster having high similarity and points in different clusters having low similarity [1]. Clustering can be a stand-alone tools and also as a pretreatment process of the model algorithms. Clustering plays an important role in the field of pattern recognition and image processing [2].

Cluster analysis as a data pretreatment process is the basis of further analysis and data processing [3]. It is also known as unsupervised learning process, since there is no priori knowledge about the data set. It also acts as an important data processing analysis tools and methods which are aimed at completing the exploratory function. Cluster analysis is a multivariate statistical method which to study “birds of a feather flock together” and also is one method of three practical methods [4], another two are regression analysis and discriminate analysis. It is mainly to study the individual or similarity distance or similarity measure, according to some clustering rules to divide the individual or data set into different clusters, making the same cluster within high similarity and different clusters have a smaller similarity [5]. In recent years, cluster analysis has been a hotspot to analyze data and extract information in the field of pattern recognition and data analysis [6].

There are many cluster algorithms, but each algorithm is optimized the certain aspects of data features, such as: minimize the within-class distance, maximize inter-class distance, etc. So far, there is no signal cluster algorithm can be used generally to reveal the structure of a variety of multi-dimensional data sets [7]. Each algorithm imposes some structure on the data set explicitly or implicitly, which result in the difficult clustering validity assessment. The traditional cluster algorithms, such as K-means, Fuzzy C means (FCM) algorithm, etc. most of these algorithms need to assume that cluster objects have certain characteristics and it can form a number of different regions, and based on the convex spherical sample space. But when the sample space is not convex, the algorithm will be trapped in local optimum. In order to solve this problem, a new cluster algorithm has been proposed, known as Spectral Clustering algorithm [8].

Spectral Clustering algorithms are based on the spectra graph theory. They treat the data clustering as a graph partitioning problem without make any assumption on the form of the data clusters, namely, the clustering of data sets mapped to the Laplacian matrix’s row vector which composed of the fist k feature vectors. Not only can the n-dimensional data sets convert into k-dimensional data sets (many times, k<<n), to achieve the purpose of dimensionality reduction, and have a good clustering results [9]. Spectral Clustering algorithm is a point-to-point cluster algorithm, and has a good application prospects. In recent years, Spectral Clustering algorithm is more studied and more increasingly widespread as a cluster analysis algorithm. It is a new branch in the cluster analysis, it was originally used for load balancing and parallel computing, VLSI design [10] and other areas, and it is beginning to be used in machine learning recently, and quickly becomes an international hot spot in the field of machine learning. Currently, Spectral Clustering is attracted more attention in the field of text mining [11, 12], information retrieval [13] and image segmentation [14], and has achieved research results.
This paper introduces the basic principles of Spectral Clustering algorithm, some typical algorithm, study situation and the applications in the field of machine learning, meanwhile points out the key issues and the development trends in the future.

II. FUNDAMENTAL PRINCIPLES

A. Basic Knowledge of Graph

Spectral Clustering based on the spectral graph partition theory, the spectral graph algorithm converts the solution space from discrete to continuous, and thus transforms the graph segmentation problem into the matrix eigenvalue problem. At present it has been a popular high-performance algorithm.

Spectral Clustering algorithm treats the clustering data sets as an undirected complete graph \( G=(V,E) \), where the vertex sets should be correspond with the data sets, that is, a vertex \( v_i \) corresponds a data \( x_i \), and the weight \( w_{ij} \) means the similarity between \( v_i \) and \( v_j \) which have connected edge [15]. So the clustering issue becomes a graph partition issue. Its mission is to cut \( V \) into \( k \) disjoint subsets of vertices, the value \( V=(V_1,V_2,...,V_k) \), so that to make the same vertices in one subset with high similarity, different subsets have lower similarity.

Different graphs can be expressed in different matrices; the same graph can also be described with different matrices, for the same problem different matrixes achieve different results. The following will give the commonly used matrix diagram [16, 17, and 18].

1) 2-value adjacency matrix: Graph \( G_k=(V_k,E_k) \), where \( V_k \) is the point sets, \( E_k \in V_k \times V_k \) is edge sets, \( i,j \) is the row and column, the corresponding 2-value adjacency matrix can be expressed as:

\[
A_k(i, j) = \begin{cases} 1, (i, j) \in E_k \\ 0, \text{ others} \end{cases}
\]  

2) The weighted adjacency matrix: for the graph \( G_k \), using \( d(i,j) \) represent the distance between two points, then the weighted adjacency matrix can be expressed as:

\[
A_k(i, j) = \begin{cases} \exp(-\frac{d(i, j)^2}{\sigma^2}), (i, j) \in E \\ 0, \text{ others} \end{cases}
\]  

3) Laplacian matrix: for the graph \( G_k \), \( A_k(i, j) \) is 2-value adjacency matrix, \( D_k(i,i) \) is the degree matrix, then the matrix \( L_k(i,j) = D_k(i,i) - A_k(i,j) \) is called Laplacian matrix. The most used Laplacian matrixes are summarized in the following table 1.

<table>
<thead>
<tr>
<th>Type</th>
<th>Laplacian Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unnormalized</td>
<td>( L=D-W )</td>
</tr>
<tr>
<td>Symmetric</td>
<td>( L_{Sy} = D^{-1/2}LD^{-1/2} )</td>
</tr>
<tr>
<td>Asymmetric</td>
<td>( L_{Az} = D^{-1}L = -D^{-1}W )</td>
</tr>
</tbody>
</table>

B. Spectral Theory

In 1973, Donath and Hoffman [19] related the graph partition problem with the eigenvectors of the similar matrix at the earliest, after then Fiedler [20] presented the relationship between the dichotomy of the graph and the second eigenvector of the Laplacian matrix. In 1992, Hagen and Ratio-Cut proposed Ratio-Cut which based on the spectral analysis [21]. Spectral Clustering algorithms derive the new features of clustering objects through the theory of matrix analysis, using the new features to cluster the original data.

An arbitrary undirected graph \( G=(V,E) \) can be expressed using a symmetric matrix, and also become a graph adjacency matrix. Then the graph is divided into two sub-graphs \( A \) and \( B \) which are not connected (where \( A \cup B=V \), \( A \cap B=\Phi \)). The similarity of internal graph is the biggest, and the similarity of the sub-graphs is the smallest [22]. The idea is to minimize the sum of weights, that is to minimize the cost function \( \text{cut}(A,B) = \sum_{i \in A, j \in B} w_{ij} \) (where, \( A \) and \( B \) are the sub-graphs, \( w_{ij} \) is the weight). Division criteria will have a direct impact on the merits of clustering results. Because the weight of the graph can combined with the various features of the clustering object, so Spectral Clustering algorithm is simple and can handle complex data types.

C. SPECTRAL CLUSTERING Algorithm Model

According to the proposal of Spectral Clustering, many researchers proposed the different concrete realization methods, but these methods follow the four main issues to explore:

Step1: to build the similarity matrix \( A \); Step2: by calculating the first \( k \) eigenvalues and eigenvectors of the \( k \), to construct feature vector space; Step3: to determine the number of clusters; Step4: using the cluster algorithm to cluster the eigenvectors of the feature vector space.

III. TYPICAL SPECTRAL CLUSTERING ALGORITHMS

A. Based on the Graph Partition of the Spectral Clustering Algorithms

Spectral Clustering can be interpreted from several angles, such as figure cut set theory, random migration point and the perturbation theory [23]. But no matter what kind of theory, Spectral Clustering has been converted to the eigenvector problem of Laplacian matrix, and then the eigenvectors are clustered. Assumed that the data \( X=[x_1,x_2,...,x_n] \in \mathbb{R}^d \) is divided into c clusters, all of which the
Normalized-Cut (N-Cut) is recognized as the best Spectral Clustering algorithm [24].

Based on N-Cut Algorithm for the Spectral Clustering:

Step 1: Construct the similarity matrix \( w_{ij} \in R^{N \times N} \), where \( w_{ii} = 0 \), when \( i \neq j \), then \( w_{ij} = \exp(-1 \times x_i - x_j^2 / \sigma^2) \);

Step 2: Calculate the diagonal matrix \( D = \text{Diag} (w \cdot 1_N) \), in which \( D_{ii} = \sum_j w_{ij} \);

Step 3: Make the similarity matrix L standardization, and \( L = (D^{-1/2})^T W (D^{-1/2}) \);

Step 4: Calculate k largest eigenvectors of the L and build characteristic matrix \( T = (t_1, t_2, \ldots, t_k) \in R^{N \times K} \);

Step 5: the row vectors of the T are normalized to the T-unit length, and take them as k samples of the RK space;

Step 6: using the K-means algorithm, to cluster the K numbers of normalized row vectors:

Step 7: the original data \( x_i \) is divided into the \( f^th \) cluster, if and only if when the \( f^th \) row of the T is divided into the \( f^th \) cluster.

From this algorithm, we can see that there are two parameters which affect the final clustering effect: the scale parameter \( \sigma \) of the intimacy function and the number of clusters \( c \). However, \( \sigma \) is a pre-determined parameter, is an artificial selected data, so that the similar function will have some limitations. We should seek a certain algorithm to automatically determine the size of \( \sigma \), thus to eliminate the artificial factors. But there are some defects when use the Laplacian matrix, in the literature [25], it referred that the clustering results by the Laplacian matrix are basically the same when the diagonal elements \( d_{ij} \) of the diagonal matrix \( D \) are the same; on the other hand, the clustering results are vary widely. For that reason, the specific environment which Lapacian matrix applies should be further explored and studied. At present, the Spectral Clustering has been proposed a number of algorithms, their main differences are: how to select the eigenvectors and how to use the eigenvectors to cluster. However, it involves the solving of the eigenvalue problem; the computational complexity of the algorithm, the similarity matrix is equivalent to weight matrix of the graph, so the changed similarity matrix will not affect the algorithm’s structure.

From this algorithm, we can see that there are two parameters which affect the final clustering effect: the scale parameter \( \sigma \) of the intimacy function and the number of clusters \( c \). However, \( \sigma \) is a pre-determined parameter, is an artificial selected data, so that the similar function will have some limitations. We should seek a certain algorithm to automatically determine the size of \( \sigma \), thus to eliminate the artificial factors. But there are some defects when use the Laplacian matrix, in the literature [25], it referred that the clustering results by the Laplacian matrix are basically the same when the diagonal elements \( d_{ij} \) of the diagonal matrix \( D \) are the same; on the other hand, the clustering results are vary widely. For that reason, the specific environment which Lapacian matrix applies should be further explored and studied. At present, the Spectral Clustering has been proposed a number of algorithms, their main differences are: how to select the eigenvectors and how to use the eigenvectors to cluster. However, it involves the solving of the eigenvalue problem; the computational complexity of the algorithm, the similarity matrix is equivalent to weight matrix of the graph, so the changed similarity matrix will not affect the algorithm’s structure.

### B. Based on the Density Consistent Criteria of the SPECTRAL CLUSTERING Algorithm

Spectral Clustering algorithm is based on the spectra graph theory, is a new clustering algorithm which can divide arbitrarily shaped data. It has been successfully applied to text mining, image retrieval [27] and other fields. However, Spectral Clustering is difficult to correctly find some clusters in which the density is much larger, and the parameter selection depends on many experiments and personal experience. Spectral Clustering algorithm Based on density consistent criteria (DCSC) sends the consistent-density similarity matrix into the classical compactness criteria of the Spectral Clustering algorithm, to replace the original similarity matrix to improve the clustering performance. As the Spectral Clustering algorithm is a typical graph partition algorithm, the similarity matrix is equivalent to weight matrix of the graph, so the changed similarity matrix will not affect the algorithm’s structure.

The traditional Euclidean distance can only reflect the local consistency characteristics of the clustering structure, but does not reflect the characteristics of global consistency. In order to overcome this shortcoming, the consistent-density is introduced in the Spectral Clustering algorithm. Of the algorithm the data points will be seen as the vertex \( V \) of a weighted undirected graph \( G=(V,E) \), edge sets \( E=\{W_{ij}\} \) mean the similarity between the data points. The algorithm designs a density-sensitive similarity measure:

\[
W_{ij} = \min_{\rho \in R} \left( \rho \cdot \text{diag} (P_{ii}) - 1 \right) + 1 \tag{4}
\]

The new similarity measure does not require the kernel function, and can directly compute the similarity distance. The consistent-density similarity measure only concerns one free parameter \( \rho \), thus reducing difficulty of the parameters.

Literature [28] combines the artificial data sets and proposed the arbitrary expansion factor \( \rho \) in the interval \( [1, e^{29}] \) to solve the practical problems, thereby confirming with the original algorithm, this algorithm has a stronger recognition of data sets’ pop structures, as well as is more

### TABLE II. SPECTRAL ALGORITHMS BASED ON THE GRAPH CUT

<table>
<thead>
<tr>
<th>Classification Criteria</th>
<th>Cost Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum Cut</td>
<td>( \sum_{a \in A, b \in B} w(u,v) )</td>
</tr>
<tr>
<td>Normalized Cut</td>
<td>( \text{Ncut}(A,B) = \sum_{a \in A, b \in B} \text{cut}(A,B) \cdot \frac{\text{assoc}(A,V)}{\text{assoc}(B,V)} )</td>
</tr>
<tr>
<td>Ratio Cut</td>
<td>( \text{Rcut}(A,B) = \frac{\text{cut}(A,B)}{\min(</td>
</tr>
<tr>
<td>Average Cut</td>
<td>( \text{Avcut}(A,B) = \sum_{a \in A, b \in B} \text{cut}(A,B) \cdot \frac{</td>
</tr>
<tr>
<td>Min-max Cut</td>
<td>( \text{Mcut}(A,B) = \sum_{a \in A, b \in B} \text{cut}(A,B) \cdot \frac{\text{assoc}(A,V)}{\text{assoc}(B,V)} )</td>
</tr>
<tr>
<td>Multiway Normalized Cut</td>
<td>( \text{MNcut}(A,B) = \sum_{a \in A, b \in B} \text{cut}(A,B) \cdot \frac{\text{assoc}(A,V)}{\text{assoc}(B,V)} )</td>
</tr>
</tbody>
</table>
robust to the random noise. Since the algorithm’s time complexity is \(O(n^3)\), has the same time complexity with the original clustering algorithm, but there are some differences in the factors. The algorithm does not inherit the advantage to reduce the data dimension; it is difficult to deal with out-of-core data problem.

C. Based on Sub-space of the Spectral Clustering algorithm

Under normal circumstances, we hope that through the nature of data sets’ local neighborhood to cluster the information of the data sets, in order to initially realize the so-called “popular convergence criteria [29]”, namely it will focus the data sets which on the same pop structure on the same class, so that even if the different pop structures overlap or cross it also can be successfully identified. Obviously, in the traditional clustering algorithm, a simple similarity criterion is difficult to identify a number of cross-clustering data. Sub-space Spectral Clustering algorithm has a strong ability to identify the data sets, can effectively solve the above problems. Spectral Clustering algorithm uses the eigenvector to construct simplified data space, can make the distributed architecture of the data in the sub space more apparent at the time of reducing the data dimension. The algorithm introduce the similarity matrix of the sub-space neighborhood \(w_{\text{sub}}\) [30], greatly enhanced the ability to identify the cross-shape data. However, the similarity matrix of similarity matrix are studied not yet ripe, so the theoretical research and applications should be considered more in more data sets.

\[
\begin{align*}
    w_{\text{sub}}(i,j) &= \frac{w_{\text{sub}}(i,j) + w_{\text{sub}}(j,i)}{2} \\
    w_{\text{sub}}^+(i,j) &= \begin{cases} 
        1, & x_i \in N_{\text{sub}}(x_j) \\
        0, & x_i \notin N_{\text{sub}}(x_j)
    \end{cases}
\end{align*}
\]

D. Based on the Multi-layer Determine the Types of the Spectral Clustering Algorithm [31]

The key problems of Spectral Clustering are to automatically determine multi-layer cluster number and deal with the out-of-core data. Multi-SC algorithm which can determines the type automatically (ACNAFLDS) is a multi-layer algorithm which can handle large scale data sets. The core idea of this algorithm is to combine the large-scale data sets according to a certain correlation to become a small group of data sets, then use the new algorithm to cluster the small data sets, finally step by step split and fine-tuning to complete all the clustering data.

This algorithm was proposed based on the ACNA [32], compared with the traditional clustering algorithms, ACNA can automatically determine the number of cluster, but for the large images, ACNA computational cost takes so high that easily exceed the computer’s memory. ACNAFLDS overcome this shortcoming of the ACNA, not only be able to automatically determine the number of simple clusters, but also solve the problems of the computer’s memory which are caused by the eigenvector and similarity matrix handling the massive data. ACNAFLDS is suitable for large images and fast, simple and easy to implement. This algorithm is effectively applied in image segmentation.

IV. PROSPECTS

With the Spectral Clustering algorithm generated, many scholars make it better development, but they found that the reason hinder its development is lack of theoretical basis. Although there are a variety of algorithms, the difference only lies in dealing with the difference of matrix, the relationship between the matrix spectrum and eigenvectors is not clear, and most of the existing Spectral Clustering algorithms need a given number of clusters in advance. Spectral Clustering algorithm should face and solve two main problems they are: 1) How to measure the similarity and dissimilarity between the data sets? 2) How can we quickly and efficiently find the optimal partition?

Spectral Clustering solely related to the number of data points, but has noting to do with the dimension, which can be avoided the singularity problem caused by the high-dimensional eigenvector. Spectral Clustering is also a distinguishing method, does not assume the global structure. Although Spectral Clustering is an extremely competitive clustering method, it is still in the early stage. There are some problems to be solved. Under normal circumstances, Spectral Clustering standards are measured from several perspectives, and not fully using the quantitative object criteria. Here are the six main criteria for Spectral Clustering algorithm:

- owning the capacity to handle large data sets;
- handling any shape, including clearance of nested data;
- the results are or aren’t related with the data input, namely the algorithm is or isn’t independent of the input sequence;
- the ability to process the data noise;
- the need to know the number of clusters, and the field of knowledge which is given by the users;
- the ability to handle the data which have much attribute, that is, whether the sensitive to the data dimension.

A variety of Spectral Clustering algorithms have their own advantages and disadvantages. Because of the actual complexity and the data diversity, each algorithm only can solve a set of problems. Therefore, the users should base on specific problems to select appropriate clustering algorithm. In recent years, along with the development of traditional methods and the new technologies in the field of data mining, machine learning and artificial intelligence, Spectral Clustering algorithm has been considerable development. Not difficult to find the new trends: (1) the convergence development of traditional clustering methods; (2) new methods are emerged; (3) according to actual needs, various fields of technology are targeted blended. In short, Spectral Clustering algorithm combines data mining, pattern recognition, mathematics, image and many other areas of research. As the theory of these areas is developed, improved and mutual penetrated, as well as new technologies emerging, cluster analysis will be developed faster.
ACKNOWLEDGMENT

This work is supported by the Basic Research Program (Natural Science Foundation) of Jiangsu Province of China (No.BK2009093), the National Natural Science Foundation of China (No.60975039), and the Opening Foundation of Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (No.IIP2006-2).

REFERENCES

[3] H. ZHAO, Study on a number of clustering issues in data mining. Xi'an: Xi'an University of Electronic Science and Technology, 2005