ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

A Review on Spectral Clustering and its Applications

Dr. S. Meenakshi[1], R. Renukadevi[2]

Associate Professor, Dept. of Computer Science, Gobi Arts & Science College, Gobichettipalayam, Tamilnadu, India[1]

Research Scholar, Dept. of Computer Science, Gobi Arts & Science College, Gobichettipalayam, Tamilnadu, India[2]

ABSTRACT: Clustering is one of the most important approaches used for exploratory data analysis. Clustering aims at grouping the similar patterns into the same cluster and finding the meaningful structure of the data. Due to the rapid growth of data, recent past spectral clustering has become the well accepted method for clustering and it predicts cluster labels by exploiting the various similarity graphs of data points. Spectral clustering is a technique which relies 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 various clusters having low similarity. Spectral clustering method plays a vital role in many applications such as data mining, pattern recognition, machine learning, image segmentation and speech processing for discovering meaningful patterns of data. This research paper presents a review on spectral clustering and its applications.

KEYWORDS: Clustering; spectral clustering; spectral clustering algorithms; spectral clustering applications; spectral clustering graphs

I. INTRODUCTION

Clustering is one of the most widely used approaches for exploratory data analysis. Clustering is a fundamental technique that has been widely explored and applied in various application domains such as data mining, pattern recognition, machine learning and image segmentation. Clustering aims at grouping the similar patterns into the same cluster and discovering the meaningful structure of the data. In the past many clustering algorithms have developed which include K-means clustering, C-means clustering and Fuzzy C-Means clustering. These traditional algorithms are depending upon the knowledge or acquisition of similarity information to relate data items to each other. K-means clustering is one of the most classical data clustering algorithms and has extensively applied in practice because of its simplicity and effectiveness [5].

Due to the rapid growth of data, the various challenges have posed on clustering are: partitioning the high dimensional data into different clusters, correlations among related clustering tasks, capturing of individual tasks, and clustering out-of-sample data [25]. In order to handle the above challenges and achieve better clustering performance, recent past spectral clustering technique has been proposed. Spectral clustering performs well in partitioning the data with more complicated structures compared to traditional clustering techniques since spectral clustering puts more efforts on mining the intrinsic data geometric structures. The fundamental idea of spectral clustering is predicts cluster labels by exploiting the different similarity graphs of data poin ts. Spectral clustering uses information obtained from the eigenvalues and eigenvectors of their adjacency matrices for partitioning of graphs.

Spectral clustering is a well accepted method for clustering and it uses the top eigenvectors of a matrix derived from the distance between points [15]. Spectral clustering is simple to implement, can be solved efficiently by standard linear algebra methods, and outperforms traditional clustering algorithms such as the K-means algorithm [21]. Given a sparse similarity graph, in spectral clustering can be implemented efficiently even for large data sets [22]. Spectral clustering is a powerful technique in data analysis that has found increasing support and application in many areas. This research paper presents a review on spectral clustering and its applications.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

The paper is organized as follows. Section II describes the concept and process of spectral clustering. Section III presents the key steps, stages, and categories of spectral clustering algorithm. Section IV provides the application areas of spectral clustering and Section V concludes the paper.

II. SPECTRAL CLUSTERING

Spectral clustering is a technique which relies 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]. Spectral clustering is an important clustering technique inspired by spectral graph theory and often performs better than traditional clustering methods such as K-means and C-Means clustering. Spectral clustering is used for grouping N data points in an I-dimensional space into several clusters. Each cluster is parameterized by its similarity, which means that the points in the same group are similar and points in different groups are dissimilar to each other. Spectral clustering refers to the general technique of partitioning the rows of a matrix according to their components in the top few singular vectors of the matrix. Spectral clustering reduces the dimensions using the eigenvalues of the similarity matrix of the data.

In spectral clustering, the first step is to create a similarity graph with vertices corresponding to the data points to be clustered and edges corresponding to the affinities between data points. This graph can be represented by an adjacency matrix W, also commonly referred to as an affinity matrix, where w_{ij} denotes the edge weight or affinity between vertices i and j. The data is represented by an n×p matrix X, with rows corresponding to data points and columns to features. The affinities wij are given by a positive semi-definite similarity function s (X_i, X_j) where X_i denotes the i^{th} row of X.

The two mathematical objects used by spectral clustering are similarity graphs and graph Laplacians [21].

A. Similarity graphs:

Given a set of data points $x_1 \ldots x_n$ and some notion of the similarity is $s_{ij} >= 0$ between all pairs of data points x_i and x_j, the primary goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in various groups are dissimilar to each other. The goal of constructing similarity graph is to model the local neighborhood relationships between the data points.

Representing the data in the form of the similarity graph is G = (V, E). Each vertex v_i in the graph represents a data point x_i. Two vertices are connected if the similarity s_{ij} between the corresponding data points' x_i and x_j is positive or larger than a certain threshold, and the edge is weighted by s_{ij}. Now the problem of clustering can be reformulated by using the similarity graph: to find a partition of the graph such that the edges between various groups have very low weights (which means that points in different clusters are different from each other) and the edges within a group have high weights (which means that points within the same cluster are same to each other).
The different similarity graph constructions which are exists to transform a given set $x_1 \ldots x_n$ of data points with pairwise similarities s_{ij} or pairwise distances d_{ij} into a graph. The several popular similarity graphs used in spectral clustering are €-neighborhood graph, K-nearest neighbor graphs and fully connected graph.

B. Graph Laplacians:

Graph Laplacian matrices are the main object for spectral clustering. The two types of graph Laplacians with their important properties are defined as the following.

- The unnormalized graph Laplacian

The Unnormalized graph Laplacian matrix is defined as L= D – W and the matrix L satisfies the following properties:

Step 1: For every vector f € R^n.
Step 2: L is symmetric and positive semi-definite.
Step 3: The smallest eigenvalue of L is 0 and the corresponding eigenvector is the constant one vector.
Step 4: L has 'n' non-negative and real-valued eigenvalues.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

- The normalized graph Laplacian

There are two matrices such as symmetric matrix and random walk are called as normalized graph Laplacians which are closely related to each other and are defined as follows:

$$L_{sym:} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$
$$L_{rw:} = D^{-1} L = I - D^{-1} W.$$

Where, the first matrix L_{sym} is denoted as symmetric matrix and the second matrix L_{rw} is denoted as closely related to a random walk.

III. SPECTRAL CLUSTERNG ALGORITHM

The spectral clustering methods are common graph-based approaches to unsupervised clustering of data. Spectral clustering algorithms are typically starts from the local information encoded in a weighted graph on the data and cluster according to the global eigenvectors of the corresponding (normalized) similarity matrix. [14]. In spectral clustering for each individual task, an explicit mapping function is used simultaneously learnt for predicting cluster labels by mapping features to the cluster label matrix. Meanwhile, that the learning process can naturally incorporate discriminative information to further improve clustering performance.

Spectral clustering algorithm consists of one significant step to construct a similarity matrix and the goal of constructing the similarity matrix is to model the local neighborhood relationships between the data vertexes. A good similarity matrix is greatly responsible for the performance of spectral clustering algorithms. The spectral algorithm depends on the time it takes to find the top k singular vectors [11]. This algorithm is start by presenting the data points in the form of similarity graph, and then need to find a partition of the graph so that the points within a group are same and the points between different groups are dissimilar to each other and the partition can be done in various normalization methods [26].

The key steps of spectral clustering algorithm are:

Gives a set of points S = {S_1..., Sn} in a high dimensional space R.

- Form the affinity matrix A€R.
- Define D to be the diagonal matrix and construct the Laplacian matrix L.
- Obtain the eigenvectors and eigenvalues of L.
- Find x_1, $x_{2...}$ xk, the k largest eigenvectors of L and form the matrix X= [x_1, x2$_{...}$ xk] €$R^{n\square k}$ by stacking the eigenvectors as columns.
- Form the matrix Y from X by renormalizing each of X's rows to have unit length.
- Treating each row of Y as a point in R^k, cluster them into k clusters using any clustering algorithm.
- Finally assign the original point S_i to cluster j if only if row i of the matrix Y has assigned to cluster j.

IV. LITERATURE SURVEY

Spectral clustering algorithm efficiency is mainly based on the fact that it does not make any assumptions on the form of the clusters and this property comes from the mapping of the original space to an eigen space. Spectral clustering is a promising approach to clustering and the various notable spectral clustering algorithms have been proposed in the literature which include Shi and Malik, 2000 [17]; Meila, Shi, 2001 [13]; Ding, Zhang, Zhang et al., 2010 [4]; Ng, Jordan and Weiss, 2002 [15]; Kannan, Vempala and Vetta, 2000 [11]; Ding, He and Simon, 2001 [28]; Bach and Jordan, 2006 [1]; von Luxburg, 2007 [21], Zhang and Jordan, 2008 [27].
The various existing spectral clustering algorithms perform spectral clustering which consists of three stages as follows [22]:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

- **Preprocessing** – This stage is used to form the normalization of the similarity matrix S.
- **Spectral Mapping** – In this stage, eigenvectors of the preprocessed similarity matrix are computed. Each data point 'i' is mapped to a tuple representing the values of component 'i' in the aforementioned eigenvectors.
- **Postprocessing/Grouping** – This stage clusters the data in the original or spectral domain.

In general, the developed spectral algorithms are categorized based on the number of eigenvectors they use for partitioning [22]. They are recursive spectral, multiway spectral and non-spectral methods. Recursive spectral algorithms split the data into two partitions based on a single eigenvector and are then are recursively used to generate more partitions. Multicut spectral algorithms split the data into multiway partitions directly using information in multiple eigenvectors. Non-spectral methods use a simple grouping algorithm that clusters the data quickly.

V. APPLICATIONS OF SPECTRAL CLUSTERNG

Spectral clustering has extensive applications in many fields and it has been successfully applied to image segmentation, pattern recognition, educational data mining and speech processing.

- **Image segmentation**

In image segmentation, spectral clustering calculates the similarity between each pair of pixels in the process of segmenting images. The spectral clustering in image segmentation is used to find similar matrix feature vector which reflects the similarity between data detect the internal structure of the data and solve them with the standard linear algebra method. Spectral clustering technique can cluster any sample space in any shape. Spectral clustering algorithm is to build an undirected graph and then make a multi-channel division. The similarity matrix is provided as an input and consists of a quantitative evaluation related to similarity of each pair of points in the dataset.

Many research works have been proposed to use spectral clustering in image segmentation [23], [24]. Image segmentation based on spectral clustering improves the quality of image segmentation and reduce the computational complexity. Spectral method reduces dimensions using the Eigen values of the similarity matrix of the data and is used to group number of data points. The advantage of spectral clustering in image segmentation is generating good results and also reducing the computation lay out. However the disadvantage of spectral clustering is when the image resolution is high, the spectral clustering method can lead to overlarge adjacency matrix.

- **Pattern recognition**

In pattern recognition the spectral clustering algorithm is used to clusters data using eigenvectors of a similarity matrix derived from dataset. The spectral clustering in pattern recognition used only for informative eigenvectors is employed for determining the number of clusters and performing clustering. Spectral clustering gives more efficient and accurate estimation of the number of clusters and better clustering results.

Spectral clustering technique has been widely used in pattern recognition research work [10], [28], [16]. The advantage of spectral clustering in pattern recognition is used to derive the similarity/affinity matrix from original dataset. The disadvantage of spectral clustering is not to perform effective clustering for noisy and sparse data.

- **Speech processing**

In speech separation, the spectral clustering is aimed to provide a methodology for finding elongated clusters while being more robust to noise. Spectral clustering in speech separation relies 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. The use of spectral clustering in speech separation is time-consuming for feature selection.

Spectral clustering technique has been widely used in speech separation research work [1], [6], [9]. The advantage of spectral methods in speech separation has involved the design of numerical approximation schemes that exploit the different time scales present in speech signals. The disadvantage of applying spectral method in speech separation is to manipulate similarity matrices of dimension.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

- **Educational data mining**

In educational data mining, the spectral clustering is used to find the meaningful clusters to predict the student performance in new representation. Spectral clustering is a graph theoretic for metric modification such that it gives a much more global notion of similarity between data points. The spectral clustering in educational data mining finds groupings by analyzing the top eigenvectors of the affinity matrix and returns the better results.

Spectral clustering technique has been widely used in educational data mining research work [19], [7], [3], [18], [12], [20]. The advantage of spectral clustering in educational data mining is to find better prediction with less time. The disadvantage of spectral clustering is prediction can made only one cluster model.

VI. CONCLUSION

Spectral clustering is one of the most widely used clustering approaches for exploratory data analysis. Spectral clustering is simple to implement and can be solved efficiently by standard linear algebra methods, and performs better than traditional clustering algorithms. This research paper reviews the concept, algorithm and the application areas of spectral clustering. Spectral clustering algorithms relies 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. Several spectral clustering algorithms have been proposed in the literature which is categorized based on the number of eigenvectors they use for partitioning. Even though, a number of spectral clustering algorithms have been developed, there are still many issues to be handled to meet the demands of advanced applications.

REFERENCES

1. Bach F. R.., and Jordan M. I., "Learning spectral clustering, with application to speech separation", Journal of Machine Learning Research, Vol. 7, pp. 1963-2001, 2006.
2. Bach F. R.., and Jordan M. I., "Spectral clustering for speech separation, Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods", pp. 221-253, Jan 2009.
3. Baker R., and Yacef K., "The State of Educational Data mining in 2009: A Review and Future Visions", Journal of Educational Data Mining, Vol. 1, Issue. 1, pp. 3-17, Oct 1, 2009.
4. Ding S., Zhang L., and Zhang Y., "Research on Spectral Clustering Algorithms and Prospects", The 2nd International Conference on Computer Engineering and Technology, Vol. 6, pp. 149-153, 2010.
5. Feiping Nie, Dong Xu.,and Tsang I. W., and Changshui Zhang ., "Spectral Embedded Clustering", In Proceedings of the 21st International Conference on Artificial Intelligence, pp. 1181-1186, Jul 11, 2009.
6. Gold B., Morga N., "Speech and Audio Signal Processing: Processing and Perception of Speech and Music", John Wiley & Sons, 1999.
7. Gong Y., Rai D., and Beck J., Heffernan N., "Does Self-Discipline Impact Students Knowledge and Learning", In Proceedings of the 2nd International Conference on Educational Data Mining, pp. 61-70, Jul 2009.
8. Hebert P. A., Macaire L., "Spatial-Color pixel classification by spectral clustering for color image segmentation", In Proceedings of the 3rd IEEE International Conference on Information and Communication Technologies: From Theory to Applications, Damascus (Syria), pp. 1-5, 2008.
9. Jang G. J., Lee T. W., "A maximum likelihood approach to single-channel source separation", Journal of Machine Learning Research 4, pp. 1365–1392, 2003.
10. Jiang M. F., Tseng S. S., Su C. M., "Two-phase clustering process for Outliers Detection Pattern Recognition Letters", Vol. 22, Issue. 6, pp. 691-700, May 31, 2001.
11. Kannan R., Vempala S., and Vetta A., "on clusterings: Good, bad and spectral", Journal ACM, Vol. 51, Issue. 3, pp. 497-515, May 2004.
12. Madhyastha T., Tanimoto S., "Student Consistency and Implications for Feedback in Online Assessment Systems", In Proceedings of the 2nd International Conference on Educational Data Mining, pp. 81-90, 2009.
13. Meila M., Shi J., "A random walks view of spectral segmentation", 2001.
14. Nadler B., Galun M., "Fundamental Limitations of Spectral Clustering", In Advances in Neural Information Processing Systems, pp. 1017-1024, 2006.
15. Ng A.Y., Jordan M.I., and Weiss Y., "On Spectral Clustering: Analysis and an algorithm", In Advances in Neural Information Processing Systems, Vol.13, Issue. 7, pp. 849-856, 2001.
16. Porikli F., Haga T., "Event detection by eigenvector decomposition using and frame features", In IEEE conference on computer vision and pattern recognition workshop, pp. 114-121, 2004.
17. Shi J., and Malik J., "Normalized cuts and image segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, Issue. 8, pp. 888-905, 2000.
18. Tanimoto S. L, "Improving the Prospects for Educational Data Mining", In Proceedings of the Complete On-Line Proceedings of the Workshop on Data Mining for User Modeling, at the 11th International Conference on User Modeling (UM 2007), 106110, 2007.
19. Trivedi S., and Pardos Z.A., Sarkozy G.N., and Heffernan N.T., "Spectral clustering in educational data mining", In M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper(editors), In Proceedings of the 4th International Conference on Educational Data Mining, pp. 129-138, 2011.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

20. Thai-Nghe N., Horvath T., and Schmidt-Thieme L., "Context-Aware factorization for personalized student's task recommendation", In Proceedings of the international workshop on personalization approaches in learning environments, pp. 13-18.
21. Ulrike von Luxburg, "A tutorial on spectral clustering, Statistics and computing", Vol. 17, Issue. 4, pp. 395-416, 2007.
22. Verma D., Meila M., "A comparison of spectral clustering algorithms", University of Washington Tech Rep UWCSE030501, Issue. 1, pp. 1-18, 2003.
23. WANG Chongjun., and LI Wu Jun, DING Lin., and TIAN Juan., CHEN Shifu., "Image Segmentation Using Spectral Clustering", In Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence, IEEE Computer Society, Nov 16, 2005.
24. Xiaohong Zhao., and Yanhua Qu., and Hong Zhang., "Sports Video Segmentation using Spectral Clustering", Journal of Multimedia, Vol. 9, Issue. 7, pp. 873-678, Jul 2014.
25. Yang Y., Zhigang Ma ., and Yi Yang ., Nie F., and Shen H.T., "Multitask spectral clustering by exploring intertask correlation", IEEE transactions on cybernetics, Vol. 45, Issue. 5, pp. 1083-1094, May 2015.
26. Zelnik-Manor L., Perona P. L., "Self-tuning spectral clustering", In Advances in Neural Information Processing Systems, pp. 1601-1608, 2004.
27. Zhang Z., and Jordan M.I., "Multiway Spectral Clustering: A Margin-Based Perspective", Statistical Science, Vol. 23, Issue. 3, pp. 383-403, 2008.
28. Zha H., He X., and Ding C., Simon H., Gu M., "Bipartite graph partitioning and data clustering", ACM, In Proceedings of the tenth international conference on Information and knowledge management, pp. 25-32, Oct 5, 2001.

BIOGRAPHY

R. Renukadevi is an M.Phil research scholar in Department of Computer Science, Gobi Arts & Science College, Gobichettipalayam. She received her M.sc. degree in the year 2015. Her area of interest is Data mining.

Dr. S. Meenakshi is working as an Associate Professor in the Department of Computer Science at Gobi Arts & Science College, Gobichettipalayam and her academic qualification includes M.C.A., M.Phil and Ph.D in Computer Science. She has over 26 years of teaching experience at the under-graduate and post-graduate levels. She has published research papers in International Journals and her areas of interests include Object-Oriented Programming Systems, Advanced Database Systems and Data Mining.