Laplacian Matrix Based Spectral Graph Clustering

Ambika P. R., Bharathi Malakreddy A.

Abstract: Recent attention in the research field of clustering is focused on grouping of clusters based on structure of a graph. At present, there are plentiful literature work has been proposed towards the clustering techniques but it is still an open challenge to find the best technique for clustering. This paper present a comprehensive review of our insights towards emerging clustering methods on graph based spectral clustering. Graph Laplacians have become a core technology for the spectral clustering which works based on the properties of the Laplacian matrix. In our study, we discuss correlation between similarity and Laplacian matrices within a graph and spectral graph theory concepts. Current studies on graph-based clustering methods requires a well defined good quality graph to achieve high clustering accuracy. This paper describes how spectral graph theory has been used in the literature of clustering concepts and how it helps to predict relationships that have not yet been identified in the existing literature. Some application areas on the graph clustering algorithms are discussed. This survey outlines the problems addressed by the existing research works on spectral clustering with its problems, methodologies, data sets and advantages. This paper identifies fundamental issues of graph clustering which provides a better direction for more applications in social network analysis, image segmentation, computer vision and other domains.

Keywords: Clustering, Laplacian, spectral graph.

I. INTRODUCTION

Clustering is one of the important process in Data Mining technology which groups similar data points into same cluster and dissimilar data points into different clusters based on similarity measures or dissimilarity measure. Most used similarity measures are Euclidean distance, cosine similarity, Manhattan distance measure, Jaccard coefficient etc. Clustering techniques identifies underlying structure and discovers meaningful data patterns from the clusters. Clustering has always been an active topic of research in today's Big Data world. In today's real world problems, data are collected in high dimensional space. Detecting clusters in high dimensional structures are a challenging task in the data mining problem . The difficulty that traditional clustering algorithms encounter in dealing with high dimensional data sets motivated the invention of subspace clustering, which

has been studied extensively in recent years. The abundance of Dimensional Reduction (DR) methods invented to solve curse of dimensionality problem of High Dimensional Data (HDD)[5]. The present subspace clustering is more focused on improving clustering performance and not much research has been reported to assist in solving DR problems [6]. Recently, a new approach in clustering called spectral methods has started to get a lot of attention in data mining domain. The spectral methods for clustering works by computing the Eigen vectors of matrix based on the distance between points and then map them to clusters. Spectral Clustering has become quite popular over the last few years and several new algorithms have been published. Spectral graph theory deals with the connectivity structures in a graph by casting a graph to an algebraic structure and by analyzing the spectra of the same. The connectivity and Spectral Graph has been discussed by Mohar[21]. A standard clustering method using spectra of graph has been proposed by Shi &Malik[22]. Graph Laplacians have become a core technology for the computer vision and data analytics. Kannan et al., and Luxburg et al.[20], they have done numerous research contributions in clustering. Belkin 2003; Nadler et al., 2006, proposed works focused on dimensionality reduction using graph based clustering. These extensive literature works has provided several useful insights towards new emerging clustering methods and also discusses several gaps between theoretical concepts and Applications. In recent years, the spectral clustering method has gained attentions because of its better-quality clustering performance.

1.1 Background and Motivation

Graph based clustering draws much of the attentions of the research communities in recent days. This is because, graph is a natural way of representation of today's real life problems; especially in network based problems. Graph based clustering algorithms are useful for producing clusters where the problem is modeled by using graphs. Graphs are very useful to represent high dimensional unstructured data. So, graph based clustering for large, high dimensional datasets gaining popularity gradually. In Spectral clustering, using spectral graph theory data points are grouped together based on the similarity and Laplacian matrix concepts . Graph based clustering methods identifies the set of nodes of a known graph and assign them into clusters. The Laplacian matrix based clustering approaches are dominant at present. The spectral/eigenvector domain is a low-dimensional space where the data points are easily separable into different clusters .

Laplacian Matrix Based Spectral Graph Clustering

This survey discusses explanation of Laplacian based spectral clustering techniques generally applied in the data mining applications. This paper also outlines some of the important research contributions towards spectral clustering.

1.2 Outline of the survey

This survey begin by defining fundamentals of a Spectral Graph theory, laplacian matrix, Eigen value and Eigenvectors in Section 2. Section 3 graph based spectral clustering is discussed. In section 4, Existing research review on Laplacian based spectral clustering is discussed.

Applications are reviewed in section 5. Lastly, we give the conclusion in Section 6.

II. A STUDY ON SPECTRAL GRAPH THEORY

Spectral Graph Theory is a fundamental mathematical concept which explores Eigen properties of graph by integrating both linear algebra method and graph theory [10]. This algebraic graph theory studies the relation between graph properties and the spectrum of the adjacency matrix or Laplacian matrix. In this paper, we study the properties of Laplacian matrices and its properties. The eigen values of the Laplacian matrix of a graph are closely related to the connectivity of the graph. The main advantage of spectral graph theory provides both local and global geometrical properties of graph structure. A spectral graph theory have been used in several prominent applications in many domains such as graph partition problems as in fig 2.1, data mining, pattern recognition, machine learning, social networks, quantum computing, bioinformatics and computer sciences[9].
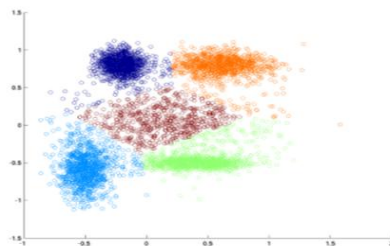

Figure 2.1 Application of graphs in Clustering technique[16].

This section explores the study of Laplacian matrix, its eigenpair properties such as, eigenvalues and eigenvectors. Graph is problem representation model which is used to represent pair wise relationship between objects. There are several algebraic representation of a graph available. Out of which, Adjacency Matrix, Laplacian Matrix and Incidence Matrix are frequently used for solving problems related to clustering. This theory provides many useful algorithms, as well as some that can be rigorously analyzed. There are two categories of graph Laplacians; Un-normalized (L_{UNorm}) and normalized Laplacian Matrix(L_{Norm}).

Adjacency Matrix A: Consider an undirected graph G = (V, E) with number of vertices V and number of edges E. If there is a edge exists between vertices i and j then the element a_{ij} = edge(i,j) of A has a real positive weight assigned to that edge and otherwise it is zero.

Weighted Adjacency Matrix W: Weights on the edges of a graph represents similarity between the vertices.. That is, w_{ij} = w_{ji} ≥0 if i and j are connected, w_{ij} = w_{ji} = 0 otherwise.

Un-normalized Laplacian Matrix L: It is real symmetric and positive semi-definite values. its smallest eigenvalue is 0. The Laplacian matrix is defined as ,$L = D_k - A$.

Normalized Laplacian Matrix: There are two variants of matrices for normalized graph Laplacians which are closely related to each other. *Symmetric matrix* of L i.e Lsym= I - $D^{-1/2}$ W $D^{-1/2}$ and *random walk* Lrw= D^{-1} L = I – D^{-1} W. Among these |V| Eigenvalues, different eigenvalues hold some specific spectral properties and the corresponding eigenvectors carry different information that can be used to cluster the graph [12].
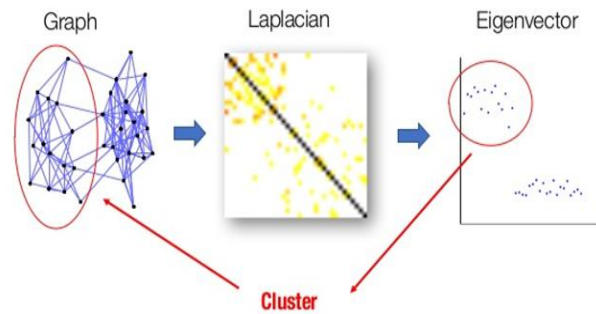

Figure 2.2 Graph partitioning by looking at eigen vectors using Laplacian Matrix[Fiedler73]

Eigenvalues and Eigenvectors

In spectral clustering algorithm it is very important one has to calculate the initial k eigenvectors of Laplacian matrix. There exists some methods to calculate the first eigenvectors like power method or Krylov subspace methods such as the Lanczos method proposed by Golub et.al, in 1996. There are two important eigenvectors ; *Principal Eigen vector* is the Eigen vector with a eigen value having largest rank value. *The Fiedler Eigen vector* refers to the second smallest Laplacian Eigen-value or a Fielder value of a connected graph and it minimizes λ_2.

III. GRAPH BASED SPECTRAL CLUSTERING

This section is to briefly outlines the spectral clustering algorithms. Von Luxburg [20] gives a extensive introduction to spectral clustering and its properties. Graph clustering is a technique used in cluster analysis which groups related vertices in a graph. There are many graph clustering methods proposed because of its extensive applicability and popularity. Spectral clustering is one powerful approach of graph clustering algorithms and it works by using spectral characteristics of the Laplacian of a graph which divide the nodes of the graph into natural clusters. The fundamental concepts of spectral clustering lies in spectral graph theory. Gaussian kernel similarity function is used in standard conventional algorithm [2] to find the similarities between the data points. Recent attention in Spectral clustering has increased because of it unchallenging execution and superior performance in numerous graph-based clustering problems. Spectral clustering techniques reduces High-Dimensional data into low-dimensional by the spectrum properties of the Laplacian matrix (L). To perform spectral clustering [14], we need 3 main steps:

1. *Pre-processing:* Formation of the associated Laplacian matrix(L). Consider an input matrix defined as the similarity matrix and then construct Degree matrix (D) and Laplacian matrix(L) from the similarity matrix of the graph i.e., between our *N* objects. In Un-normalized Spectral Clustering, Laplacian matrix can be defined as: $L_G = D_G - W_G$. In Normalized Spectral Clustering, Laplacian matrix is defined as:
$L_{rw} = I - D^{-1} W$. Where L_{sym} is a symmetric matrix and L_{rw} is related to random walk.

2. *Decomposition:* Obtain Eigen values and Eigen vectors of Laplacian matrix L. Find x1,x2.....xk, first k largest Eigen vectors of L. These form the new columns in a new matrix X.

3. *Clustering:* Assign and label each data point to a low-dimensional representation based on one or more eigenvectors [15].

Figure 3.1 Laplacian based spectral clustering.

Table 3.1 Existing Research Contributions on Graph Laplacian based spectral clustering

IV. RESULTS AND DISCUSSIONS

Graph based clustering draws much of the attentions of the research communities in recent days. In the present literature reviews, there is no distinct principle of "Laplacian graph matrix" and how its variant matrices are defined. As a result, a special attention is necessary for understanding literature on graph Laplacians. Recent emerging graph clustering classifies the clustering with graph representation. There exists several graph representation methods; k-nearest neighbor graph, similarity graph and graph subspace clustering. According to these methods, their performance extremely relies on the pre-defined graph.

At present, numerous amount of research work proposed towards different graph based clustering techniques. The following table 3.1 gives outlines of the existing research contributions being carried out on laplacian based spectral graph clustering techniques.
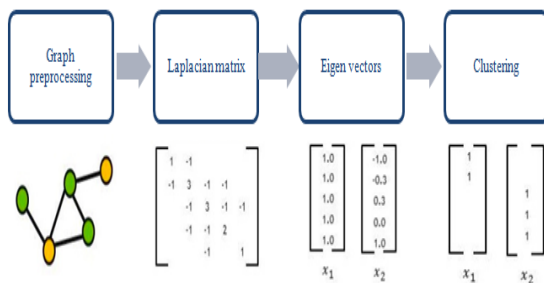
Author	Problem	Methodology	Dataset	Advantage
Kun Zhan[1]	To improve the graph quality	Multi View Graph Learning (MVGL)	UCI Digits Caltech COIL-20	Gives better performance compared to robust multi-view k-means clustering, similarity network fusion and MVSC.
Feiping Nie et.al [2]	Constrained Laplacian Rank method which learns k-components within a graph	L1-norm and L2-norm based on the Laplacian graph	7 benchmark datasets: Yeast, Abalone, OIL20,COIL100,AR, XM2VTSand UMIST.	Compared with different clustering methods and outperforms; K-means, Ratio Cut Normalized Cut and NMF methods.
Kavitha K R et.al[3]	k-Means is unsuccessful to find similarity in the structure of data	spectral clustering based on PCA for the Laplacian graph	jain dataset two circle dataset	**Laplacian graph matrices** with accurate result of PCA measure
Abhishek, Eti Goel [6]	scalability issue	Map Reduce approach for semi-supervised learning	two moons and Amazon Kindle Review data set	**Laplacian graph** performs better compared to SVM
Arthur Guillon et al [7]	Subspace clustering	Weighted Laplacian Fuzzy Clustering	Data set on real data artificial data	the effect of the subspace clustering term gives better performance than regularization term.
Xi Peng et.al [8]	robust subspace learning and clustering	subspace clustering and learning algorithms applied on L2-graph	Hopkins155 database	L2-graph outperforms :L1-graph, low rank representation, least square regression.
Armando di Nardo et al [9]	urban water supply	water management tasks in networks	c-town parete network	Good network performance results. Identifying water loss becomes simple

Hein et al[19]	multiple graphs clustering using unsupervised and semi-supervised technique	Linked Matrix Factorization (LMF)	SIAM	LMF shows better clustering performance in an unsupervised setting
T. Xiang et.al [18]	This paper addresses two issues; Automatically determine the number of clusters and Perform effective clustering on noisy and sparse data	Measures the relevance of an eigenvector for effective clustering	synthetic data and real-world data	approximate the number of clusters properly ; expose natural clustering for sparse data and noisy data.
Yi Jiang et al [17]	Irrelevant or noisy features in data, misleads existing spectral feature selection methods to choose 'wrong' features.	Eigen value Sensitive Criteria (EVSC) for feature selection on graph Laplacian Eigen values	PIX10P(PIX) ORL10P(ORL) GLA-BRA(GLA) CLL-SUB(CLL) TOX-171(TOX)	In comparison with two methods, Laplacian Score and SPEC, EVSC demonstrates excellent performance and high stability.
Huazhong Ninga,Wei Xu et.al [14]	To handle dynamic data	The eigenvalue system updates and it generates instant cluster labels	web–blogs data	Insertion, deletion and similarity change in the data points, are integrated in the proposed framework
Parthajit Roy et.al [16]	Computational efficiency	All-pair shortest path distances of graph and its Laplacian spectra.	Flame dataset	It can handle non-separability as well as non-convexity of the clusters.

V. APPLICATIONS OF GRAPH CLUSTERING

Spectral clustering technique recently has many applications in data clustering, image segmentation, and dimension reduction. It is widely used in many diverse domains because of its simple implementation approach by solving eigen-gap problem. Spectral clustering empirically very successful in several applications.

Communication networks: In any communication networks, graph clustering helps for data analysis, modeling, prediction evolution of the network and anomalous use. Clustering is used to identify substructures, analyze the connectivity in computer networks. Clustering is also used in the AD-Hoc network and sensor networks.

Pattern recognition:. The spectral clustering in pattern recognition used only for informative eigenvectors. The advantage of spectral clustering in pattern recognition is it gives more efficient and accurate number of clusters and better clustering results. The disadvantage of spectral clustering is not to perform effective clustering for noisy and sparse data.

Drug Discovery: The process designed to discover drugs which is useful for modern pharmaceutical industry. Spectral clustering which gives promising results in the area of protein library analysis. The eigen-pairs of a matrix used as the basis for partitioning a dataset.

Image segmentation (Shi & Malik 2000) : In digital image processing, segmentation is important for image description and classification approach. Many research works have been proposed to use spectral clustering in image segmentation because it improves the quality of image segmentation and reduces the computational complexity.

Speech separation: Spectral clustering technique has been widely proposed in speech separation domain. However, spectral clustering is able to find extended clusters and is more robust to noise.

VI. CONCLUSION

Recently Graph Laplacians have become a core technology for the domain of machine learning. Spectral clustering technique works based on the properties of the Laplacian matrix. This survey discusses explanation of fundamental techniques generally applied and review some of the important existing research publications on Laplacian matrix based graph clustering. The usage of laplacian based clustering approaches are dominant at present. Recently, Spectral clustering technique used in many applications in data clustering, image segmentation, graph partition problem, social network analysis and dimensionality reduction. This manuscript discuss some important applications of graph clustering. And also, we discussed many positive insights, but still there are some gaps exist between theory and practice. A plentiful of spectral clustering algorithms have been proposed , but still there exists many challenges to be solved to meet the demands of advanced applications. The fundamental issues of graph clustering identified in this survey provides a better direction for our research work towards Laplacian based clustering for High dimensional Data(HDD).

REFERENCES

1. Kun Zhan, Changqing Zhang et.al, "Graph learning for multi-view clustering", IEEE transactions on cybernetics 2017.
2. Feiping Nie et.al, " The Constrained Laplacian Rank Algorithm for Graph-Based Clustering" AAAI- 2016.
3. [Kavitha K R et.al, " Improved spectral clustering using pca based similarity measure on different laplacian graphs" IEEE International Conference on ' computational intelligence and computing research ', 2016.
4. Naumov, Maxim (NVIDIA) and Moon,Timothy "Parallel Spectral Graph Partitioning" March 2016.

5. Ambika P.R, Dr.Bharathi Malakreddy A, " Review of Existing Research Contribution towards Dimensional Reduction Methods in High Dimensional Data" Springer Lecture Notes on Data Engineering and Communications Technologies Volume 15 (Page No-409-421).
6. Abhishek, Eti Goel et.al, " Parallel Graph Laplacian For Large Datasets", CICT'17, IEEE 2017.
7. Arthur Guillon, Marie-Jeanne Leso "Laplacian Regularization for Fuzzy Subspace Clustering" IEEE 2017.
8. Xi Peng, Zhiding Y ,"Constructing the L2-Graph for Robust Subspace Learning and Subspace clustering" IEEE, 2017.
9. Armando di Nardo "Applications of graph spectral techniques to water distribution network management" 2018,
10. Arsic B et. al, " Graph spectral techniques in computer sciences" 2012.
11. Ulrike von Luxburg " A Tutorial on Spectral Clustering " 2007.
12. Shi, J, Malik, J " Normalized cuts and image segmentation" IEEE Transactions 2000
13. Fiedler M , "Algebraic connectivity of graphs" Czechoslovak journal, pp. 298–305, 1973.
14. Huazhong Ninga et.al " Incremental spectral clustering by efficiently updating the Eigen-system" Pattern Recognition PP-113-127 Elsevier 2009.
15. Jianbo Shi, Jitendra Malik "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence vol.22, pp.888-905, Aug 2000.
16. Parthajit Roy et.al "A Shortest Path Similarity Matrix based Spectral Clustering" International Journal of Computer Sciences and Engineering Volume-4, Special Issue 2016.
17. Jiang Yi , Ren Jiangtao " Eigenvalue sensitive feature selection" International Conference on Machine Learning USA, 2011.
18. Xiang T , Gong S, "Spectral clustering with eigenvector selection" Pattern-Recognition volume 41 pp. 1012-1029, 2008.
19. Hein, Von Luxburg et.al, "From Graphs to Manifolds :weak and strong point wise consistency of graph laplacians" COLT 2005.
20. Mohar, "Some Applications of Laplace eigenvalues of graphs", Graph Symmetry: Algebraic Methods and Applications, NATO ASI, Series.C-497, pp.225-275, pub.Kluwer, Editor. G. Hahn and Sabidussi, 1997.
21. Jianbo Shi and Jitendra Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, no.8, pp.888-905, August, 2000.

AUTHORS PROFILE

Ambika P. R. received her B.E. degree and M.Tech. degree in Computer Science and Engineering from Visvesvaraya Technological University (VTU), Belgaum, Karnataka. She is a Research Scholar in the Department of CSE at BMS Institute of Technology and Management, Bengaluru under Visvesvaraya Technological University. Currently, she is working as an Assistant Professor in the Department of CSE, City Engineering College, Bengaluru, Karnataka. Her areas of interest includes data mining, data science, big data analytics and Machine learning.

Dr. Bharathi Malakreddy A has B.E. degree in Computer Science and Engineering, M.Tech. in CSE and Ph.D. in Computer Science and Engineering, University of JNTU Hyderabad. Presently working as Professor and HOD, Department of Artificial Intelligence and Machine Learning at BMS Institute of Technology and Management, Bengaluru. Worked as visiting faculty for IGNOU and Infosys campus connect program. Areas of interest are Wireless Sensor Network, Medical Imaging, IoT, Big Data and Cloud Computing. Published papers in IEEE, Springer, Elsevier and LNSE.