ISSN:0975-9646

A Survey : Spectral Clustering Applications and its Enhancements

S.V.Suryanarayana[#1], Dr.G.Venkateswara Rao[*2] Dr. G.Veereswara Swamy [*3]

[#] *Department of Computer Science and Engineering ,GITAM University Visakhapatnam,India*

[*] *Department of Information Technology,GITAM University Visakhapatnam,India*

Abstract - The spectral clustering algorithm is an algorithm for placing N data points in an I-dimensional space into different clusters. Each cluster is described by its similarity, which means that the points in the same cluster are similar and points in different clusters are dissimilar to each other. Recently, spectral clustering has become an increasingly espouse tool and has been applied in many areas such as statistics, machine learning, pattern recognition, data mining, and image processing. This survey paper discusses these methods in detail and later on introduces the relationship between spectral clustering and k-means clustering and spectral clustering's applications in image segmentation, educational data mining, entity resolution and speech separation. It also mention the improvements in this algorithm using Nystrom methods.

1. INTRODUCTION

Clustering is a task of grouping a set of objects into classes with similar characteristics. There are many data clustering algorithms that do a good job. However, recently spectral techniques for data clustering have emerged as a powerful tool for clustering data. To solve the clustering problem we calculate the eigenvectors and eigen values of the graph laplacian which is a similarity measure between two data points. The clustering is obtained from the eigenvectors. Many algorithms have been proposed for spectral clustering which are small disparity of the above technique. In this survey report, we will discuss spectral clustering, a more powerful and specialized clustering algorithm[4](compared to K-means).

There are several motivations for clustering as following:

a) A good clustering has predictive power.

b) Clusters can be useful in communication because they allow people to put objects with similar features into one category and to name them.

c) Failures of one cluster model may draw special attention to interesting objects.

d) Clusters may serve as models of learning processes in neural systems.

Spectral clustering[4] glean its name from spectral analysis of a graph, which is how the data are represented. Spectral clustering techniques reduce dimensions using the eigen values of the similarity matrix of the data. The similarity matrix[4] is provided as an input and consists of a quantitative evaluation of the relative similarity of each pair of points in the dataset. The spectral clustering algorithm is an algorithm 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. We start the algorithm by presenting the data points in the form of similarity graph, and then we need to find a partition of the graph so that the points within a group are similar and the points between different groups are dissimilar to each other. The partition can be done in various ways such as minimum cut method, ratio cut method, and normalized and MinMaxCut Method[3]. This paper will discuss the relationship between spectral clustering and k-means clustering and spectral clustering's applications in different areas.

2. SPECTRAL CLUSTERING ALGORITHM

Spectral clustering is is appealingly simple: Given some data, you build an affinity (or kernel) matrix, analyze its spectrum, and often get a perfect clustering from the dominant eigen vectors for free. This simple algorithm[4] or its slightly more complex variants which yield so good results are widely appreciated for applications.

Here are the key steps of spectral clustering algorithm:

Given a set of points $S = \{s_1, \ldots, s_n\}$ in a high dimensional space.

1. Form a distance matrix $D \in R^2$. This distance measure is Euclidean, but other measures also make sense.
2. Transform the distance matrix to an affinity matrix by $A_{ij} = \exp(-\sigma_{ij})$ if $i \neq j$,0 if $i = j$. The free parameter σ^2 controls the rate at which affinity drops off with distance.
3. Form the diagonal matrix D whose (i,i) element is the sum of A's i^{th} row, and create the Laplacian matrix $L = D^{-1/2} AD^{-1/2}$
4. Obtain the eigenvectors and eigenvalues of L.
5. Form a new matrix from the vectors associated with the k largest eigenvalues. Choose k by using eigen gap method.
6. Each item now has a vector of k coordinates in the transformed space. Normalize these vectors to unit length.
7. Cluster in k-dimensional space. The result will be k well-a separated clusters.

Spectral clustering is a more advanced algorithm compared to k-means as it uses several mathematical concepts (i.e. degree matrices weight matrices, similarity matrices, similarity graphs, graph Laplacians, eigenvalues and eigenvectors) in order to divide similar data points in the same group and dissimilar data points in different groups. This Spectral Clustering works well for many real world data sets eventhough, it needs some modification in terms of improving its time complexity, space complexity.

3. SPECTRAL CLUSTERING APPLICATIONS IN RECENT LITERATURE

Spectral Clustering has been extensively used in many areas, including in the statistics, machine learning, pattern recognition, data mining, and image processing.

3.1 Image segmentation

In digital image processing, segmentation is important for image description and classification. Clusters can be formed for images build on pixel intensity, color, texture, location, or some combination of these. "Spectral clustering involves the eigen decomposition of a pair wise similarity matrix, which is intractable for sufficiently large images. Down-sizing the image, however, will cause a loss of finer details and can lead to inaccurate segmentation results" (Tung, Wong, and Clausi, 2010). So Tung et al. (2010) [7]proposed a method of spectral clustering to large images using a combination of block wise processing and stochastic ensemble consensus. The idea of this method is to perform an over-segmentation of the image at the pixel level using spectral clustering, and then merge the segments using a combination of stochastic ensemble consensus and a second round of spectral clustering at the segment level. And we use stochastic ensemble consensus o integrate both global and local image characteristics in determining the pixel classifications. This step also removes blockwise processing artifacts. (Tung et al., 2010) Tung et al. (2010)[7] also presented the experimental results on a set of natural scene images (from the Berkeley segmentation database) of the normalized cut, the self-tuning spectral clustering.They conclude that "the proposed method achieves segmentation results that are comparable to or better than the other two methods. In particular, detailed structures are better preserved in the segmentation, as reflected in the higher recall values" (Tung et al., 2010)[7]

3.2 Educational Data Mining

With quickly increasing data repositories from different educational areas, useful information and data in educational data mining is playing a outstanding role in student learning since it can answer important research question about student learning. K-means clustering is a simple and powerful tool to monitor students academic performance by discovering the key characteristics from students performance and using these characteristics for future prediction. Furthermore, we are able to boost the student performance prediction by using spectral clustering. Trivedi, Pardos, Sarkozy and Heffernan[6] implemented spectral clustering for analyzing data set of 628 students state test scores from the 2004-2005 school year and the features included the various dynamic features. The data was collected using the ASSISTments tutor in two schools in Massachusetts and ASSISTments is an brilliant Tutoring System developed at Worcester Polytechnic Institute, MA, USA. The prediction was the MCAS test scores for the same students in the following year. The technique for making a prediction for a test point includes the following steps and is shown in figure 1:

1. Divide the data into K clusters.
2. Apply a separate linear regression model to each cluster.
3. Each such predictor (such as linear regression) represents a model of the cluster and is called a cluster model. And the collection of cluster models is called a prediction model , where K indicates the number of clusters. (Trivedi et al.)[6]

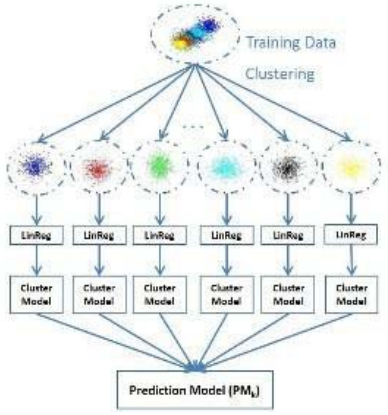

Figure 3.1 The technique for making a prediction for a test point (Trivedi et al.)

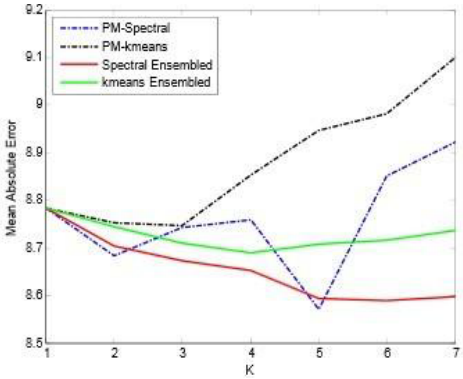

Figure 3.2 The 5 fold cross validated errors using different prediction models from K=1 to 7 for both k-means and spectral clustering (Trivedi et al.).

3.3 Entity resolution

In many telecom and web applications, the demand of entity resolution is getting bigger and bigger. Entity resolution is to recognize whether the objects in the same source represent the same entity in the real world. This problem emerge often in the area of information integration when there lacks a unique identifier across multiple data sources to represent a real world entity. Blocking is an important technique for improving the computational efficiency of the algorithms for entity resolution. To solve the entity resolution problem, Shu, Chen, Xiong and Meng proposed an efficient spectral neighborhood (SPAN) algorithm based on spectral clustering. SPAN is an unsupervised and unconstrained algorithm and it is applicable in many applications where the number of blocks is unknown beforehand. (Shu et al.) SPAN uses the vector space model in the way of representing each record by a vector of *qgrams*. A qgram is a length q substring of blocking attribute value. And the algorithm is implemented in the following steps:

1. Define the similarity matrix for the records based on the vector space model.
2. Derive SPAN based on spectral clustering.

3. Use Newman-Girvan modularity as the stopping criterion for blocking.

Shu et al. compared SPAN with three common blocking algorithms, Sorted Neighborhood, Canopy Clustering and Bigram Indexing. The experiments were performed on both published synthetic data and real data and the results indicate:

1. SPAN is fast and scalable to large scale datasets while Canopy Clustering and Bigram Indexing are not.
2. SPAN outperforms the other three when data have low or medium noise.
3. SPAN is much more robust than Canopy Clustering and Bigram Indexing in respect to the tuning parameters because the performance of Canopy Clustering and Bigram Indexing require a large number of labeled data and thus are often not possible with data in the real world applications. (Shu et al.)

3.4 Speech separation

While linkage algorithms and k-means algorithms are very popular in speech processing and robust to noise, they are only best suited for rounded linearly separable clusters. However, spectral clustering is able to find extended clusters and is more robust to noise than the above two algorithms. Bach and Jordan applied spectral clustering to data from four different male and female speakers with speech signals of duration 3 seconds based on a cost function that characterized how close the eigen structure of a similarity matrix W is to a partition E. According to Bach and Jordan[2], "minimizing this cost function with respect to the partition E leads to a new clustering algorithm that takes the form of weighted k-means algorithms. Minimizing them with respect to W yields a theoretical framework for learning the similarity matrix". The basic idea of their algorithm is to combine the knowledge of physical and psychophysical properties of speech with learning algorithms. The physical properties provide parameterized similarity matrices for spectral clustering and the psychophysical properties help generate segmented training data. There were 15 parameters to estimate using Bach and Jordan's [2] spectral learning algorithm. For testing, they used mixes from speakers which were different from those in the training set (the four different male and female speakers with speech signals of duration 3 seconds). Bach and Jordan's analyzed that the performance of the separation is good enough to obtain audible signals of reasonable quality even though some components of the "black" speaker are missing. As we can seen from the results, the proposed approach was successful in demixing the speech signals from two speakers.

3.5 spectral clustering of protein sequences

An important problem in genomics is the automatic inference of groups of homologous proteins from pair wise sequence similarities. Several approaches have been proposed for this task which is “local” in the sense that they assign a protein to a cluster based only on the distances between that protein and the other proteins in the set. It was shown recently that global methods such as spectral clustering have better performance on a wide variety of datasets.

Spectral Clustering of Protein Sequences Using Sequence-Profile Scores Rajkumar Sasidharan1, Mark Gerstein1, Alberto Paccanaro2*

An important problem in today’s genomics is that of grouping together evolutionary related proteins when only sequence information is available. Genome sequencing projects have led to a huge increase in the number of known protein sequences. Grouping together sequences with common evolutionary origin provides a high-level view of sequence space. It facilitates identification of general features which may be associated with given biological functions. If some of the sequences are of unknown biological function their placement in a particular neighbourhood may give a clue to their function. From a biological perspective it is desirable to group together as many evolutionarily related sequences as possible, while not contaminating the clusters with false positives. Clearly a very conservative cut-off for defining relatedness would exclude the latter possibility but it would most likely mean that many sequences remain singletons, because the distance to the nearest neighborhood is deemed to be too far for membership to that community. In addition to a meaningful grouping of sequences, we require a fast algorithm for computing the distances. However the measure of distance (or similarity) may not capture all functional relationships, as some sequences with common evolutionary origin can have very weak sequence similarity; recognizing these distant relationships is difficult. We have shown that our spectral clustering in combination with a distance measure obtained from a sequence-profile method like PSI-BLAST provides better clustering than using a distance measure obtained from pair wise methods like BLAST or other local methods in our experiments, the F-measure (which provides a quantitative measure on cluster quality) was consistently better.

3.6 A Text Image Segmentation Method Based on Spectral Clustering

Images generally contain rich messages from textual information, such as street name, construction identification, public transport stops and a variety of signal boards. The textual information assists the understanding the essential content of the images. If computers can automatically recognize the textual information from an image, it will be highly valuable to improve the existing technology in image and video retrieval from high-level semantics (Lienhart, 2002, pp.256-268). For instance, road signs and construction identification in a natural environment can be captured into images by cameras and the textual information will be detected, segmented, and recognized automatically by machines. These messages then can be synchronized as human voice to be used as instructions for visually impaired person. In addition to the example, textual information extraction plays a major role in images retrieval based on contents, cars auto-drive, vehicle plate recognition and automatics.In general, automatic textual extraction consists of text detection, localization, binarization and recognition etc. In a natural scene texts could have different

backgrounds and characters in the text message can also have variety of forms. And, existing OCR (Optical Character Recognition)engine can only deal with printed characters against clean backgrounds and can not handle characters embedded in shaded, textured or complex backgrounds. So that characters are separated from the text in the detected region accurately is very necessary. Currently, many researchers have done a lot of work in the text detection and a lot of methods of text detection and location have been proposed. (Mariano, 2000;D. Chen,2004; Zhong,2000; X.L Chen,2004;X. Chen,2004) Compared to the text detection in natural scenes, specialized study of the characters extraction from natural environment is not more. The purpose of this paper is to extract accurate binary characters from the localize text regions so that the traditional OCR can work directly. In our approach, the histogram of intensity is used for the object of grouping, we partition the image into two parts using the gray levels of an image rather than the image pixels. For most images, the number of gray levels is much smaller than the number of pixels. Therefore, the proposed algorithm occupies much smaller storage space and requires much lower computational costs and implementation complexity than other similar algorithms.

4. RECENT IMPROVEMENTS IN SPECTRAL CLUSTERING ALGORITHM

Over the past decade, spectral clustering methods have gained popularity as a method to perform data clustering one of the most basic tasks of machine learning. These methods enjoy some important advantages, such as the ability to cluster non-vectorial data, and often yield superior empirical performance. Moreover, they are well-studied and supported theoretically. In our literature survey we identified the major advances in the spectral clustering algorithm. Here we are notifying some of the improvements.

4.1 Improvement in time complexity Anna Choromanska1⋆, Tony Jebara2, Hyungtae Kim2, Mahesh Mohan3, and Claire Monteleoni3[8] propose and analyze a fast spectral clustering algorithm with computational complexity linear in the number of data points that is directly applicable to large-scale datasets. The algorithm combines two powerful techniques in machine learning: spectral clustering algorithms and Nystrom methods[8] commonly used to obtain good quality low rank approximations of large matrices. The proposed algorithm applies the Nystrom approximation to the graph Laplacian to perform clustering.We provide theoretical analysis of the performance of the algorithm and show the error bound it achieves and we discuss the conditions under which the algorithm performance is comparable to spectral clustering with the original graph Laplacian

4.2 Time and Space Efficient Spectral Clustering via Column SamplingMu Li1 ,Xiao-Chen Lian1,James T. Kwok2 and Bao-Liang Lu1[9] As only several eigenvectors are required in the procedure, a general approach to alleviate this problem is by using low-rank matrix approximations, among which the Nystrom method is the most popular. It samples mx n columns from the original n × n matrix, and then forms a low-rank approximation of the full matrix by using the correlations between the sampled columns and the remaining n − m columns. As only a portion of the full matrix is computed and stored, the Nystrom method can reduce the time and space complexities significantly. Fowlkeset al. successfully applied this to spectral clustering for image segmentation [6]. Besides this, the Nystrom method has also been popularly used for tasks such as Gaussian processes and manifold learning .

4.3 Spectral Clustering on a Budget Ohad Shamir, Naftali Tishby[10] focus on the problem of performing spectral clustering under a budget constraint. Namely, a situation where we can only query a limited number of entries from the similarity matrix, but still wish to cluster comparably well as if we had the entire matrix at hand. **Ohad Shamir, Naftali Tishby[10]** propose and study, theoretically and empirically, two algorithms for this task. The first algorithm is a simple and efficient randomized procedure, with formal performance guarantees. The theoretical analysis indicates that its performance improves as the data is more easily clustered. In particular, for well clustered data and an n×n similarity matrix, a budget of O˜(n) (i.e., linear up to logarithmic factors) will suffice. The second algorithm is adaptive, and has better empirical performance. On the flip side, it is much more computationally demanding, and without better theoretical guarantees.

4.4 Active Spectral Clustering via Iterative Uncertainty Reduction Fabian L. Wauthier, Nebojsa Jojic and Michael I. Jordan[11] propose an active learning algorithm for spectral clustering that incrementally measures only those similarities that are most likely to remove uncertainty in an intermediate clustering solution. In many applications, similarities are not only costly to compute, but also noisy. We extend our algorithm to maintain running estimates of the true similarities, as well as estimates of their accuracy. Using this information, the algorithm updates only those estimates which are relatively inaccurate and whose update would most likely remove clustering uncertainty. We compare our methods on several datasets, including a realistic example where similarities are expensive and noisy. The results show a significant improvement in performance compared to the alternatives

4.5 Parallel Spectral Clustering traditional spectral clustering suffers from a scalability problem in both memory use and computational time when a dataset size is large. To perform clustering on large datasets, Yangqiu Song1,4, Wen-Yen Chen2,4, Hongjie Bai4,Chih-Jen Lin3,4, and Edward Y. Chang4 propose an improved algorithm to parallelize both memory use and computation on distributed computers. Through an empirical study on a large document dataset of 193, 844 data instances and a large photo dataset of 637, 137, they demonstrate that their parallel algorithm can effectively alleviate the scalability problem.

5 CONCLUSION

Clustering is one of the most popular techniques used in many applications such as statistics, machine learning, pattern recognition, data mining, and image processing. And spectral clustering has become more and more widely used since it is a simple method in cluster analysis and often outperforms traditional clustering algorithms like k-means. In this paper, we have introduced the theoretical background and procedures of spectral clustering. In short, it makes used of eigen values of the similarity matrix of the data to perform dimensionality reduction before clustering in fewer dimensions and then we find a partition of the graph so that the points within a group are similar and the points between different groups are dissimilar to each other. This partition can be done in various ways. This paper also presented the connection between weighted kernel k-means and spectral clustering and spectral clustering's applications in different areas such as image segmentation, educational data mining, entity resolution and speech separation. This paper also presented the different improvements in terms of reducing the time complexity based on Nystrom Methods. All these vast applications in various domains motivated me to some significant improvement in this Spectral clustering algorithm.

REFERENCES

[1] Auffarth B. (2007) Spectral Graph Clustering. Retrieved February12, 2013, from http://wwwlehre.inf.uos.de/~bauffart/spectral.pdf

[2] Bach F. & Jordan M. *Spectral Clustering for Speech Separation*. etrieved March 4, 2013, from http://www.cs.berkeley.edu/~jordan/papers/bach-jordan-spectral-clustering-chapter.pdf

[3] Dhillon, Guan & Kulis. *Kernel k-means, Spectral Clustering and Normalized Cuts*. Retrieved February 20, 2013, from http://www.cs.utexas.edu/users/inderjit/public_papers/kdd_spectral_kernelkmeans.pdf

[4] Hagen, L.and Kahng,A.(1992). New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Computer Aided Design,11(9),1074-1085 Luxbug U.(2007)A Tutorial on Spectral Clustering, Retrieved February 10,2013, from http://www.kyb.mpg.de/fileadmin/user_upload/files/publications/attachments/Luxburg07_tutorial_4488%5B0%5D.pdf

[5] Shu L., Chen A., Xiong M. & Meng W. *Efficient Spectral neighborhood Blocking for Entity Resolution*. Retrieved March 1, 2013, from http://www.cs.binghamton.edu/~meng/pub.d/ICDE11_conf_full_065_update.pdf#page=1&z oom=auto,0,800

[6] Trivedi S., Pardos Z. Sarkozy G. & Heffernan N. *Spectral Clustering in Educational Data Mining*. Retrieved February 26, 2013, from http://web.cs.wpi.edu/~nth/pubs_and_grants/papers/2011/EDM%202011/Trivedi%20Spectral%20Clustering%20in%20Educational%20Data%20Mining.pdf

[7] Tung F., Wong A. & Clausi D. (2010) Enabling Scalable Spectral Clustering for Image Segmentation. Retrieved February 22, 2013, from http://vip.uwaterloo.ca/files/publications/Enabling%20scalable%20spectral%20clustering%20for%20image%20segmentation.pdf

[8] Anna Choromanska, Tony Jebara, Hyungtae Kim, Mahesh Mohan, and Claire Monteleoni. Fast Spectral clustering via the Nystrom Method The 24th International Conference on Algorithmic Learning Theory ,Singapore,Repubic of Singapore,October 6-9,2013.Retrieved from http://www.columbia.edu/~aec2163/NonFlash/Papers/ALT2013FSCVTNM.pdf

[9] Mu Li, Xiao-Chen Lian, James T. Kwok , Bao-Liang Lu. Time and space efficient spectral clustering via column sampling. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011 retrieved from http://bcmi.sjtu.edu.cn/~blu/papers/2011/cvpr2011-muli-paper.pdf

[10] Ohad Shamir, Naftali Tishby; Spectral Clustering on a Budget. JMLR W&CP 15:661-669, 2011. Retrieved from http://www.jmlr.org/proceedings/papers/v15/shamir11a/shamir11a.pdf

[11] Fabian L. Wauthier, Nebojsa Jojic and Michael I. Jordan, Active Spectral Clustering via Iterative Uncertainty Reduction. KDD '12 Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining Pages 1339-1347. Retrieved from http://www.stats.ox.ac.uk/~wauthier/wauthier-jojic-jordan-kdd12.pdf