IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 9, September 2017

A Technical Study on Spectral and Ensemble Clustering Methods and Its Applications

Ms. D. Priyadarshini[1], Vidhya V[2]

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[1]

M. Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India[2]

Abstract: Cluster analysis enclosed a number of different algorithms and methods for grouping objects into a respective cluster using the similarity among objects. The attractiveness of cluster analysis is its ability to find groups or clusters directly from the given data. Many clustering approaches and algorithms have been developed and successfully applied to many applications. Spectral clustering groups the objects with high similarity measure and eigenvalues. This paper gives an overview of the various types of clustering and the research conclusion from the recent techniques. The analysis of the various application involved with the spectral clustering is studied with its problem analysis.

Index Terms: Consensus Clustering, Ensemble Clustering, Spectral Clustering, Co-association Matrix, Weighted K-means.

1. INTRODUCTION

Clustering is the unsupervised learning process, which groups the objects into a set of groups called as clusters based on the density or distance. Clustering is the process of segmenting the objects or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups [1]. In simple words, the aim of clustering is to split groups with similar behavior and assign them into clusters. In many applications [2], the notion of a cluster is unfortunately not well defined. The definition of a cluster depends on the nature of the data, the desired results and the goal of the application. This paper gives the overview of the spectral and ensemble clustering along with the challenges and opportunities of new clustering algorithm development.

Spectral Clustering: Spectral clustering associated with three major steps, which are pre-processing, decomposition and grouping. In the pre-processing step, the spectral algorithms construct the matrix representation of the graph. The decomposition step computes the eigenvalues and eigenvectors of the matrix and maps the each point to a least dimensional representation. This is based on the eigenvector. Finally, the grouping process performed by assigning the points to two or more clusters. The spectral clustering methods care connectivity rather than the proximity. Spectral clustering treats the data clustering as a graph partitioning problem without make any assumption on the form of the data clusters [3].

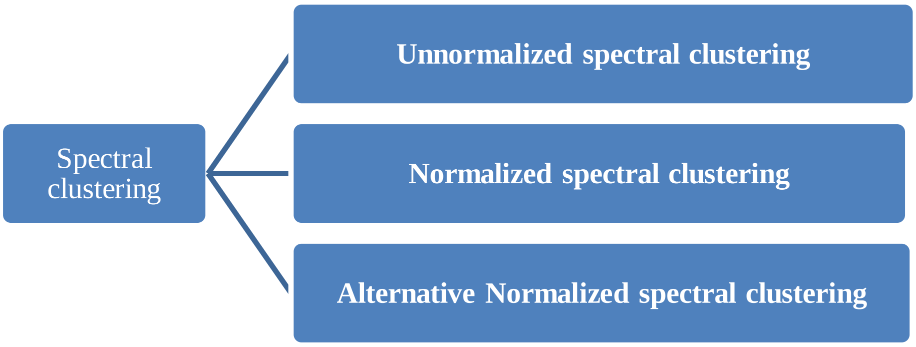
Spectral clustering
Unnormalized spectral clustering
Normalized spectral clustering
Alternative Normalized spectral clustering

Fig 1.0 types of spectral clustering algorithm

The spectral algorithm can be categorized into three types, which are normalized, un-normalized and alternative methods which is shown in Fig 1.0.

Ensemble Clustering: Ensemble clustering, also known as consensus clustering is emerging as a promising solution for multi-source and/or heterogeneous data clustering. The idea here is that by combining multiple partitions (clustering

DOI10.17148/IJARCCE.2017.6938

ensembles) of the same data, we can obtain a better data partitioning. A reliable separation of clusters are formed by using a co occurrence of matrix, a clustering ensemble is generated with the parameter values using the same clustering algorithm [4]. This allows applying different clustering algorithms on the same data which leads to different clusters or even using different data representations of the data with different clustering algorithms. Finally the results are combined together and re-clustered. Ensemble clustering is the collection of result from various basic partitions. The portioned or chunked minor clusters are collected together. This comes with the two types, with or without the explicit global objective functions, which compares the similarity between the portioned clusters. It utilizes the summarized utility function to improve the ensemble cluster performance.

In the paper [5] authors proposed a Quadratic Mutual Information based objective function for consensus clustering, and used Kmeans [6] clustering to find the solution. Further, they used the expectation-maximization [7] algorithm with a finite mixture of multinomial distributions for consensus clustering.

In this paper [8], authors proposed SEC and uncover an equivalent relationship with weighted K-means clustering that dramatically decreases the time and space complexity. Based on weighted K-means, the intrinsic consensus objective function of SEC is derived, then authors investigated its robustness and generalizability, and extend it to incomplete basic partitioning. Experimental results demonstrated that SEC produces high quality, efficient clustering compared with other state-of-the-art methods. Later authors in [9] expanded the SEC and proved the spectral clustering of the co-association matrix is equivalent to the binary matrix, which adopts k-means clustering algorithm. Then authors enhanced the method to support incomplete partitions. Most of the existing works focused on the process of the clustering on the modified co-association matrix, while the SEC method exactly decomposes the co-association matrix into a binary matrix and provides a high efficient solution for consensus clustering.

2. LITERATURE SURVEY

2.1 SPECTRAL CLUSTERING APPLICATIONS

From the past decade, Spectral Clustering algorithm is evolved as more powerful algorithm in the field of data mining. In the recent research, the Spectral Clustering is extensively used in text mining, information retrieval and image segmentation domains. The obtained results of spectral clustering are very notable.

(i) Image segmentation

Authors in the paper [10] proposed a spectral clustering method applicable for massive images using a mixure of block wise transform and stochastic ensemble consent. In digital image refinement, distribution is important for image description and classification. According to their mechanism the clusters formed for images established on the attributes called intensity of the pixel, color, texture, location, and mix of these. “Major functionality of the spectral clustering is the eigen decomposition of pair wise affinity matrix, which is very complex for high dimensional datasets. The basic idea of this mechanism used by the author is to execute an over-partition of the image with respect to the pixel level using spectral clustering, and hence combine the partitions by using a mix of stochastic ensemble consensus and an iterative approach of spectral clustering at individual segment level. To determine the pixel classifications they used stochastic ensemble in cooperation with both global and local image quality. The current step also removes block wise processing artifacts. Tung et al.[86] also presented the empirical outcomes on a collection of uniform scene images of the normalized cut, the self-tuning spectral clustering. They conclude that "the illustrated mechanism produce partition outcomes that are identical to or superior to the rest of the two techniques”.

(ii) Educational Data Mining

Authors in [11] proposed a methodology to describe Educational Data Mining (EDM) tasks as well, such as making an in-tutor prediction on the KDD Cup 2010 dataset. The immensely inter-disciplinary terrain of Educational Data Mining (EDM) has emerged from a fusion of numerous different areas, some of which covers Machine Learning, Cognitive Science, and Psychology. The major task in EDM is to build computational models and tools to mine data that explored in an educational setting. With rapidly growing data archive from various educational contexts (paper tests, e-learning, Intelligent Tutoring Systems etc.), best tradition in EDM can potentially answer significant research issues about student learning. The task of EDM is justifying instrumental in combining the knowledge derived from the data to combine with theories from cognitive psychology to formulate the best learning settings and methodologies. The results have been very encouraging. In the same vein, this methodology was also tried on the Performance Factor Analysis (PFA) task, the only difference being that the predictor used to train on each cluster was a logistic regression model. Preliminary work has indicated an improvement in the prediction accuracy. The objective of this work was to introduce to the domain of Educational Data Mining the great utility of using spectral clustering methods.

Authors in [12] used spectral clustering to enhance the performance of a new ensemble method proposed in an earlier work. While the objective was to introduce the use of spectral clustering, a very significant result of the work is proving the efficiency of Dynamic Assessment as compared to static assessment. These results show that an Intelligent

DOI10.17148/IJARCCE.2017.6938

Tutoring System that can assess as it assists offers a significant advantage to students and teachers. This is because it can not only save time that is wasted on assessment for instruction, but it can also be a better predictor of their performance in post-tests.

(iii) Entity resolution

Authors in [13] proposed an efficient spectral neighborhood (SPAN) algorithm based on spectral clustering. In numerous telecom and web utilization, the demand of entity resolution is getting bigger and bigger. The main aim of Entity resolution is to check the correspondence between the objects in the related source and the identical entity in the real world. The similar problem rises often in the field of data integration when there lacks a specific attribute across several data authority to serve as a real world entity. Blocking is a necessary technique for developing the computational performance of the algorithms for entity resolution. To solve the entity resolution issue, an efficient spectral neighborhood (SPAN) algorithm constitute on spectral clustering is then proposed. SPAN is an unsupervised and unconstrained algorithm and it is suitable in several applications where the number of blocks is unexplored beforehand. SPAN uses the vector space model in the way of characterize every record by a vector of qgrams.

2.2 RECENT IMPROVEMENTS IN SPECTRAL ENSEMBLE CLUSTERINGALGORITHM

From the past few years, spectral clustering techniques have attained reputation as a mechanism to implement data clustering one of the largest primitive functions of machine learning. All the spectral techniques face few relevant improvements, being the capability to cluster scalar data, and generally give exceptional experimental performance. Furthermore, they are well-studied and backed hypothetically. In our literature survey we identified the major advances in the spectral clustering algorithm. Here we are notifying some of the improvements.

(i) Improvement in time complexity

In paper [14] recommend and review a new spectral clustering algorithm with improvement in time complexity. This algorithm works well with linear time complexity in the order of given input information points. Hence for the massive data sets also this algorithm works accurately in linear time. This fast and improved algorithm is implemented based on the combination of the two important algorithms spectral clustering algorithms and Nystrom methods which are widely used mechanisms in machine learning. Usually these algorithms applied to attain best quality low rank similarities of massive matrices. The suggested algorithm employs the Nystrom similarity on the graph Laplacian to implement clustering. We commenced the hypothesis study of the administration of the algorithm and prove that the oversight limit it attains and we also mention the algorithm performance circumstances corresponding to spectral clustering with the initial graph Laplacian.

(ii) Time and Space Efficient Spectral Clustering via Column Sampling

In the paper [15] authors discussed the performance of the Spectral clustering only certain eigenvectors plays the major role. To investigate those crucial eigenvectors is a typical problem. A simple approach to resolve this issue is with the use of low-rank matrix similarities. One of the simple approach which is most adequate for this problem is the Nystrom technique. In this technique, first it partition m x n columns from the initial n × n matrix, after that regulates a low-rank approximation of the complete matrix by using the correspondence between the sampled columns and the rest of the n - m columns. It is clear that in the performance of the Nystrom technique only some portion of complete matrix is analyzed and stored, hence it automatically reduce the time and space complexities greatly. Fowlkes et al. satisfyingly enforced this to spectral clustering for image partition. Along with the Nystrom technique has also been regularly applied for the issues such as Gaussian processes and manifold training.

(iii) Spectral Clustering on a Budget in paper [16] authors introduced the spectral clustering that performs on the bottom of the budget restraint. They worked under a constraint direction in which they are concentrating only on some specific entries to generate clusters from the affinity matrix even though; the complete matrix is available in our hand. Authors introduced two algorithms for this issue. These two algorithms are explained theoretically as well as practically. The first algorithm is a elementary and it uses randomized procedure to produce the satisfied results. The hypothesis clearly points out that once data is clustered in simple manner then the performance improves automatically. Specifically, for a given simple n × n affinity matrix the clustering is performed within the budget of $\tilde{O}(n)$ (i.e., linear time with respect to input data points). The next algorithm is flexible, and has improved experimental performance. On the other side, second algorithm involves higher computational complexity, and the results are not matched with hypothesis.

(iv) Active Spectral Clustering

Authors in the paper [17] introduced a unique training algorithm for Spectral clustering which iterative approach. It measures the affinities in an incremental procedure through intermediate results produced from each and every

DOI10.17148/IJARCCE.2017.6938

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 9, September 2017

iteration. For particular applications, measuring the affinity is very complex and ambiguous. They implement this algorithm to preserve execution assessment of the pure affinities and also to measure the accuracy. Based on this information, the algorithm upgrades some specific measures which are approximately unreliable and whose upgradation would useful for removing the uncertainty in clusters. They study these methods on several datasets, including a real world example where affinities are more expensive and ambiguous. From the outcome it is very clear that these methods improve the performance of the clustering relative to other available methods.

(v) Parallel Spectral Clustering proposed in the paper [18] in the case of larger datasets traditional spectral clustering experience several difficulties as both memory usage and time complexity increases corresponding to the increase in the data size. To apply clustering on massive datasets, however, the parallel spectral clustering affected by the scalability issue.

(vi) 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 [19]. 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 cannot 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.

2.3 MAJOR CHALLENGES IN CLUSTERING

The process of clustering has many common and specific issues, which are not represented in the literatures, are summarized below. This summarization helps us to select optimal clustering methods.

(i) Data representation: clustering algorithm performance is considered as one of the important factor of data representation, the clusters are compact and simple if the representation is good as for K-means finds them. But unfortunately there is no such good representation. Domain knowledge is used to guide the choice of representation. **(ii) Number of Clusters:** In data clustering the most difficult problems is the finding of clusters automatically. on running a clustering algorithm for variable values of K the best value of K is traced out by criterion function. MML, minimum message length criteria is used b authors in conjunction with Gaussian mixture model in order to estimate the values of K. They started the approach with large number of clusters, and slowly merged the clusters which made a decrease in the MML criterion. Another commonly used approach is Gap Statistics. Here the main assumption is that when dividing data into an optimal number of clusters. The obtained partition is more similar to the accidental perturbation. For the number of clusters a non parametric was introduced by Dirichlet Process (DP). It is used in probabilistic models in the derivation of distribution of posterior which are most likely computed clusters. A number of clusters of Bayesian prior of non parametric are introduced with a key idea. It is very difficult to decide the exact value of K for more meaningful clusters.

2.4 Summary:

While new clustering algorithms continue to be developed, some issues still have to be resolved. Some problems and research directions as pointed in the literature have to be addressed:

There is a need to achieve tighter integration between clustering algorithms and application requirements. Each application has its own requirements: some of them just need a global partition of the data while others need to have the best partition with great precision. Generally, in mining applications, the goal is not to provide all the clusters of the search results but a summarized list of the different topics of the query. Users can after easily figure out what they are exactly searching for by selecting the target topic. Showing images from the target category in which the user is truly interested is much more effective and efficient than returning all the clusters or all the mixed images. There is a need for clustering algorithms that lead to computationally efficient solutions for large-scale data. Not all clustering algorithms can deal with large scale issues. There is a need for stable and robust clustering algorithms that lead to stable solutions even in the presence of noisy data. There is a need to use any available a priori information concerning the nature of the dataset and the goal/domain of the application in order to decide which data

DOI10.17148/IJARCCE.2017.6938

representation is the most suitable and which clustering method is the most appropriate. There is a need to have generic clustering that can be applied to any type of data. There is a need for benchmark data with available ground truths and diverse data sets from various domains to evaluate any kind of clustering algorithm because current benchmarks are limited to a small dataset that can be applied only for a limited choice of clustering methods. As said above, the growing amount of data leads to diverse data (both structured and unstructured). Raw images, text, video are considered as unstructured data because they do not follow a specific format, in contrast to structured data where there is a semantic relationship between objects. Generally, clustering approaches are applied without taking into account the structure of the data. It is precisely for these reasons, that new algorithms are being developed. In the literature several authors presents an overview of clustering techniques and highlights some emerging, and useful, trends in data clustering, some of which are presented below.
Another problem in clustering is the scalability, which is stated as Large-scale clustering: large size datasets are being handled by the clustering algorithms. Some of them are based on efficient nearest neighbor’s search and use trees as in literature or random projections as in. When clustering algorithms are summarized then large data set are converted to small data set. Dataset as with the BIRCH algorithm in contrast to sampling based methods like CURE algorithm which creates a sub-sample, when clustering is performed on a small data set then it is transferred to large data set.
It is more difficult when a cluster is formed by the mixture of heterogeneous components in the multi way. A classical clustering method leads to poor performances. Co-clustering treats this problem, and has been successfully applied to document clustering (at the same time both the word and documents are clustered together, this leads to multi way clustering here a set of objects are being clustered simultaneously by the heterogeneous components.

3. CONCLUSION

As mentioned before, thousands of clustering algorithms have been proposed in the literature in many different scientific disciplines. This makes it extremely difficult to review all the published approaches. Nevertheless, clustering methods differ on the choice of the objective function, probabilistic generative models, and heuristics. We will briefly review some of the major approaches. Clusters can be defined as high density regions in the feature space separated by low-density regions.

REFERENCES

[1]. Sidhu, Nimrat Kaur, and Rajneet Kaur. "Clustering in data mining." International Journal of Computer Trends and Technology (IJCTT) 4, no. 4 (2013): 710-714.
[2]. Aggarwal, Charu C., and Chandan K. Reddy, eds. Data clustering: algorithms and applications. CRC press, 2013.
[3]. Von Luxburg, Ulrike. "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.
[4]. Ghosh, Joydeep, and Ayan Acharya. "Cluster ensembles." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 1, no. 4 (2011): 305-315.
[5]. Topchy, A. Jain, and W. Punch, “Combining multiple weak clusterings,” in Proceedings of ICDM, 2003.
[6]. Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE transactions on pattern analysis and machine intelligence 24, no. 7 (2002): 881-892.
[7]. Bradley, Paul S., Usama Fayyad, and Cory Reina. Scaling EM (expectation-maximization) clustering to large databases. Redmond: Technical Report MSR-TR-98-35, Microsoft Research, 1998.
[8]. Liu, Hongfu, Tongliang Liu, Junjie Wu, Dacheng Tao, and Yun Fu. "Spectral ensemble clustering." In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp. 715-724. ACM, 2015.
[9]. Liu, H., Wu, J., Liu, T., Tao, D., & Fu, Y. (2017). Spectral ensemble clustering via weighted k-means: Theoretical and practical evidence. IEEE Transactions on Knowledge and Data Engineering, 29(5), 1129-1143.
[10]. Tung, Frederick, Alexander Wong, and David A. Clausi. "Enabling scalable spectral clustering for image segmentation." Pattern Recognition 43.12 (2010): 4069-4076.
[11]. Trivedi, S., Pardos, Z., Sárközy, G., & Heffernan, N. (2010, June). Spectral clustering in educational data mining. In Educational Data Mining 2011.
[12]. Trivedi, S., Pardos, Z. A., & Heffernan, N. T. (2011, June). Clustering students to generate an ensemble to improve standard test score predictions. In International Conference on Artificial Intelligence in Education (pp. 377-384). Springer Berlin Heidelberg.
[13]. Shu, L., Chen, A., Xiong, M., & Meng, W. (2011, April). Efficient spectral neighborhood blocking for entity resolution. In Data Engineering (ICDE), 2011 IEEE 27th International Conference on (pp. 1067-1078). IEEE.
[14]. Choromanska, Anna, et al. "Fast spectral clustering via the nyström method." International Conference on Algorithmic Learning Theory. Springer, Berlin, Heidelberg, 2013.
[15]. Li, Mu, Xiao-Chen Lian, James T. Kwok, and Bao-Liang Lu. "Time and space efficient spectral clustering via column sampling." In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 2297-2304. IEEE, 2011.
[16]. Shamir, Ohad, and Naftali Tishby. "Spectral clustering on a budget." In International Conference on Artificial Intelligence and Statistics, pp. 661-669. 2011.
[17]. Wauthier, Fabian L., Nebojsa Jojic, and Michael I. Jordan. "Active spectral clustering via iterative uncertainty reduction." Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.
[18]. Chen, Wen-Yen, et al. "Parallel spectral clustering in distributed systems." IEEE transactions on pattern analysis and machine intelligence 33.3 (2011): 568-586.
[19]. Wu, Rui, Jianhua Huang, Xianglong Tang, and Jiafeng Liu. "A Text Image Segmentation Method Based on Spectral Clustering." Computer and Information Science 1, no. 4 (2008): 9.

DOI10.17148/IJARCCE.2017.6938