EasyChair Preprint

№ 2523

Compare Spectral and Hierarchical Clustering using different parameters

Anand Khandare and Roshankumar R. Maurya

February 1, 2020

Compare Spectral and Hierarchical Clustering using different parameters

Prepared by

Dr. Anand Khandare
Associate Professor
Thakur College of Engineering and Technology
Mumbai, India

Roshakumar R. Maurya
Department of Computer Engineering
Thakur College of Engineering and Technology
Mumbai, India

***Abstract*-- Clustering is an automatic learning technique which aims at grouping a set of objects into clusters so that objects in the same clusters should be similar as possible, whereas objects in one cluster should be as dissimilar as possible from objects in other clusters. Clustering aims to group in an unsupervised way, a given dataset into clusters such that dataset within each clusters are more similar between each other than those in different clusters. Cluster analysis aims to organize a collection of patterns into clusters based on similarity.**

This report show a comparison between two clustering techniques: Spectral Clustering, and Hierarchical Clustering. Given the dataset that was used for this report, the most accurate techniques were found to be Spectral Clustering (when using lobpcg as the eigen solver, and considering 25 neighbors when constructing the affinity matrix), and Hierarchical Clustering (when computing the linkage using the cosine method, and using an 'average' as a linkage cretirion).

***Keywords*-- Clustering, Hierarchical Clustering, Spectral Clustering.**

I. INTRODUCTION

Clustering is the most interesting topics in data mining which aims of finding intrinsic structures in data and find some meaningful subgroups for further analysis. It is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. Thus a cluster could also be defined as the "methodology of organizing objects into groups whose members are similar in some way."

In this paper, we compare between two clustering techniques: Spectral Clustering, and Hierarchical Clustering. Given the dataset includes 2645 samples. That was used for this report, the most accurate techniques were found to be Spectral Clustering algorithm and Hierarchical Clustering algorithm, in this algorithms we used different parameters compare their time complexity and error rate to display which is better such as when spectral using 'lobpcg' as the eigen solver, and considering 25 neighbours when constructing the affinity matrix, and Hierarchical Clustering using computing the linkage using the cosine method, and using an 'average'' as a linkage criterion.

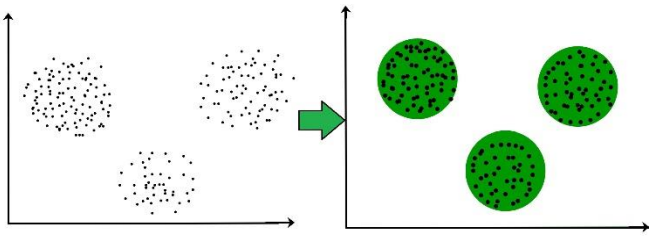

Figure.1. Clustering

II. BACKGROUND AND MOTIVATION

A. Background

Clustering is one of the challenging mining techniques in the knowledge data discovery process. Managing huge amount of data is a difficult task since the goal is to find a suitable partition in an unsupervised way (i.e. without any prior knowledge) trying to maximize the intra-cluster similarity and minimize inter-cluster similarity which in turn maintains high cluster cohesiveness. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups.

The instances are thereby organized into an efficient representation that characterizes the population being sampled. Thus the output of cluster analysis is the number of groups or clusters that form the structure of partitions, of the data set. In short clustering is the technique to process the data into meaningful group for statistical analysis. The exploitation of Data Mining and Knowledge discovery has penetrated to a variety of Machine Learning Systems.

B. Motivation

As the amount of digital documents over the years as the Internet grows has been increasing dramatically, managing information search, and retrieval, etc., have become practically important problems. Developing methods to organize large amounts of unstructured text documents into a smaller number of meaningful clusters would be very helpful as clustering such as indexing, filtering, automated metadata generation, population of hierarchical catalogues of web resources and, in general, any application requiring document organization.

Also there are large number of people who are interested in reading specific news so there is necessity to cluster the news articles from the number of available articles, since the

large number of articles are added each data and many articles corresponds to same news but are added from different sources. By clustering the articles, we could reduce our search domain for recommendations as most of the users are interested in the news corresponding to a few number of clusters.
This could improve the result of time efficiency to a greater extent and would also help in identification of same news from different sources. The main motivation is to compare different types of unsupervised algorithm to study their behaviour, advantage, and disadvantage and study how you choose unsupervised learning algorithm based on the dataset type.
This paper projected a some common clustering algorithm (Spectral and Hierarchical Clustering) for compare and analysis their behavior on different types of dataset such as structural dataset and un-structural dataset etc. and also implement the different parameter of unsupervised learning algorithm to observed error rate, correctness etc. by compare different unsupervised learning algorithm we get their advantage and disadvantage, what types of dataset we used that algorithm to increase the application performance.

III SPECTRAL CLUSTERING

Spectral clustering is a technique with roots in graph theory, where the approach is used to identify communities of nodes in a graph based on the edges connecting them. The method is flexible and allows us to cluster non graph data as well. Spectral clustering uses information from the eigenvalues (spectrum) of special matrices built from the graph or the data set. We'll learn how to construct these matrices, interpret their spectrum, and use the eigenvectors to assign our data to clusters.

Eigenvectors and Eigenvalues

Critical to this discussion is the concept of eigenvalues and eigenvectors. For a matrix A, if there exists a vector x which isn't all 0's and a scalar λ such that Ax = λx, then x is said to be an eigenvector of A with corresponding eigenvalue λ. We can think of the matrix A as a function which maps vectors to new vectors. Most vectors will end up somewhere completely different when A is applied to them, but eigenvectors only change in magnitude. If you drew a line through the origin and the eigenvector, then after the mapping, the eigenvector would still land on the line.
The amount which the vector is scaled along the line depends on λ. Eigenvectors are an important part of linear algebra, because they help describe the dynamics of systems represented by matrices. There are numerous applications which utilize eigenvectors, and we'll use them directly here to perform spectral clustering.

Basic Spectral Algorithm

1. Create a similarity graph between our *N* objects to cluster.
2. Compute the first k eigenvectors of its Laplacian matrix to define a feature vector for each object.
3. Run k-means on these features to separate objects into k classes.

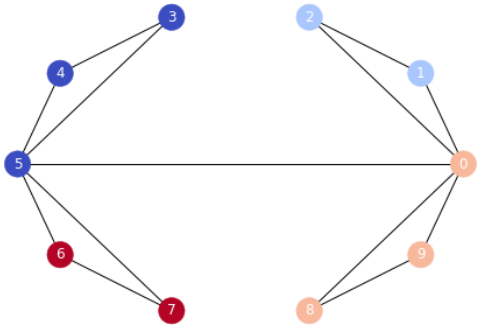

Figure 2: Spectral Clustering for Four Cluster

1) **Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG):** Locally Optimal Block Preconditioned Conjugate Gradient (LOBPCG) is demonstrated to efficiently solve eigenvalue problems for graph Laplacians that appear in spectral clustering. For static graph partitioning, 10–20 iterations of LOBPCG without preconditioning result in ˜10x error reduction, enough to achieve 100% correctness for all Challenge datasets with known truth partitions.
LOBPCG methods do not require storing a matrix of the eigenvalue problem in memory, but rather only need the results of multiplying the matrix by a given vector. Such a matrix-free characteristic of the methods makes them particularly useful for eigenvalue analysis problems of very large sizes, and results in good parallel scalability on multi-threaded computational platforms to large matrix sizes processed on many parallel processors.
LOBPCG is a block method, where several eigenvectors are computed simultaneously as in the classical subspace power method. Blocking is beneficial if the eigenvectors to be computed correspond to clustered eigenvalues, which is a typical scenario in multi-way spectral partitioning, where often a cluster of the smallest eigenvalues is separated by a gap from the rest of the spectrum. Blocking also allows taking advantage of high-level BLAS3-like libraries for matrix-matrix operations, which are typically included in CPU-optimized computational kernels.

Advantages

1. Does not make strong assumptions on the statistics of the clusters - Clustering techniques like K-Means Clustering assume that the points assigned to a cluster are spherical about the cluster centre. This is a strong assumption to make, and may not always be relevant. In such cases, spectral clustering helps create more accurate clusters.
2. Easy to implement and gives good clustering results. It can correctly cluster observations that actually belong to the same cluster but are farther off than observations in other clusters due to dimension reduction.
3. Reasonably fast for sparse data sets of several thousand elements.

Disadvantages

1. Use of K-Means clustering in the final step implies that the clusters are not always the same. They may vary depending on the choice of initial centroids.
2. Computationally expensive for large datasets this is because eigenvalues and eigenvectors need to be computed and then we have to do clustering on these

vectors. For large, dense datasets, this may increase time complexity quite a bit.

IV HIERARCHICAL CLUSTERING

Hierarchical clustering algorithm that groups similar objects into groups called clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other. Time Complexity O(n^3)

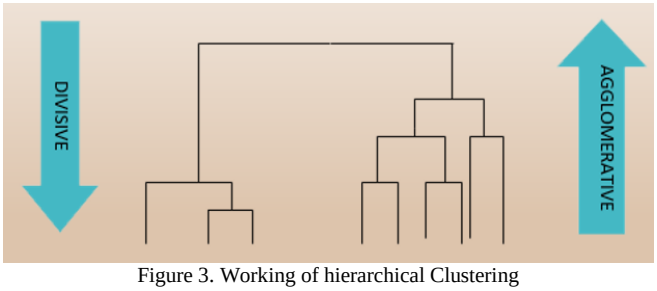

Figure 3. Working of hierarchical Clustering

There are two basic approaches for generating a hierarchical clustering:
Agglomerative: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.
Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step and how to do the splitting.
Hierarchical clustering techniques are by far the most common. A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram, which displays both the cluster sub-cluster. Many agglomerative hierarchical clustering techniques are variations on a single approach: starting with individual points as clusters, successively merge the two closest clusters until only one cluster remains.

Basic hierarchical algorithm

1. Compute the proximity matrix, if necessary.
2. **Repeat**
3. Merge the closest two clusters.
4. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.
5. **Until** Only one cluster remains.

There are various parameter we use in hierarchical clustering.

1) Single Link Hierarchical Clustering: For the single link of hierarchical clustering, the proximity of two clusters is defined as the minimum of the distance (maximum of the similarity) between any two points in the two different clusters. Using graph terminology, if you start with all points as singleton clusters and add links between points one at a time, shortest links first, then these single links com- bine the points into clusters. The single link technique is good at handling non-elliptical shapes, but is sensitive to noise and outliers.

2) Complete Link (CLIQUE) Hierarchical Clustering: The complete link of Hierarchical clustering, the proximity of two clusters is defined as the maximum of the distance between any two points in the two difference clusters. Using graph terminology, if you start with all points as singleton clusters and add links between points one at time, shorted links first, then a group of points is not a cluster until all the points in it are completely linked, i.e. from a clique. Complete link is less susceptible to noise and outliers, but it can break large clusters and it favors globular shapes.

3) Group Average Hierarchical Clustering: The group average of hierarchical clustering, the proximity of two clusters is defined as the average pairwise proximity among all pairs of points in the different clusters. This is an intermediate approach between single and complete link approaches.

4) Centroid Hierarchical Clustering: Centroid methods calculate the proximity between two clusters by calculating the distance between the centroids of clusters. These technique may seem similar to K-means, but as we have remarked, ward's method is the correct hierarchical analog. Centroid methods also have a characteristic often considered bad that is not possessed by the hierarchical clustering techniques.

Advantages

1. Hierarchical clustering outputs a hierarchy, i.e. a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram.
2. Easy to implement.

Disadvantages

1. It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
2. Time complexity: not suitable for large datasets.
3. Initial seeds have a strong impact on the final results.
4. The order of the data has an impact on the final results.
5. Very sensitive to outliers.

Clustering Method	α_A	α_B	β	γ
Single Link	1/2	1/2	0	−1/2
Complete Link	1/2	1/2	0	1/2
Group Average	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	0	0
Centroid	$\frac{m_A}{m_A+m_B}$	$\frac{m_B}{m_A+m_B}$	$\frac{-m_A m_B}{(m_A+m_B)^2}$	0
Ward's	$\frac{m_A+m_Q}{m_A+m_B+m_Q}$	$\frac{m_B+m_Q}{m_A+m_B+m_Q}$	$\frac{-m_Q}{m_A+m_B+m_Q}$	0

Table 1: Hierarchical clustering parameter statistic

Sr no.	Hierarchical Cluster	Spectral Cluster
1	It is can't handle larger dataset well	It is can't handle larger dataset well
2	hierarchical algorithm by interpreting the dendrogram	Spectral algorithm by interpreting the Eigenvalues and Eigenvector
3	Hierarchical clustering, especially are globular.	Spectral clustering also a tighter cluster
4	It work well with clusters of different	It work well with clusters of different size

	size and different density	and different density
5	Easy to implement every dataset	Easy to implement in graph dataset
6	Time complexity O (kn^2).	Time complexity O(n(n+k))
7	There are various type of parameter to execute	There are various types of parameters to execute

Table 2: Compare

V. EXPERIMENT

In this report we use the two clustering algorithm. The algorithm are spectral clustering and hierarchical clustering algorithm with their different types of parameter. Use both cluster parameter we analysis and compare which algorithm is better.

The dataset used for this study is **'manhatta-ndof. csv'**, which was made available to us by NYU. The dataset includes 2645 samples. The attributes that were used from this dataset are the following:

- BldClassif - Building class. Used as a cluster indicator for validation purposes.
- GrossSqft, MarketValueperSqft - Indipendent variables that were used to generate the prediction model.
- I Seems that the data that was very hard to cluster, and in most cases, the mode; that was generated by the algorithms, which are being describes below, was not accurate in describing the raw data.

Neighborhood	BldClassif	YearBuilt	GrossSqFt	GrossIncomeSqFt	MarketValueperSqFt
0	0	1926	239121	34	158
0	1	1909	51387	36	170
0	1	1911	167448	35	167
1	1	1910	9530	41	201
2	0	1923	182200	31	148
2	0	1923	182200	31.8	148
3	1	1925	66492	25	104
1	1	1918	43339	38	177

Table 3: Sample dataset

VI. RESULT

Since the data that was used was not very easy to cluster, it is easy to assume that different dataset could yield different results. Compare between different spectral cluster using with their parameter. And also compare hierarchical clustering using their parameter. At the end we compare both the algorithm to see which algorithm se better using their different parameter.

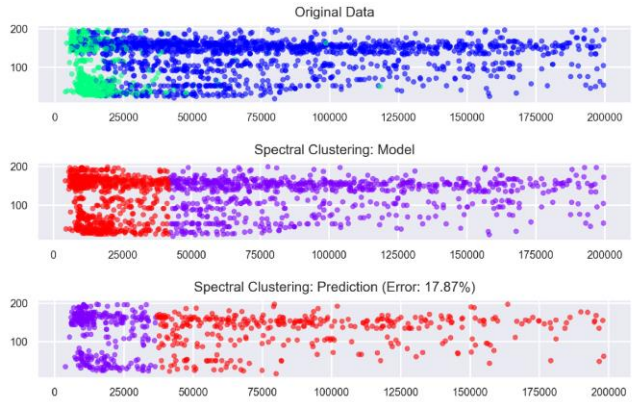

Figure 4. Simple Spectral clustering result

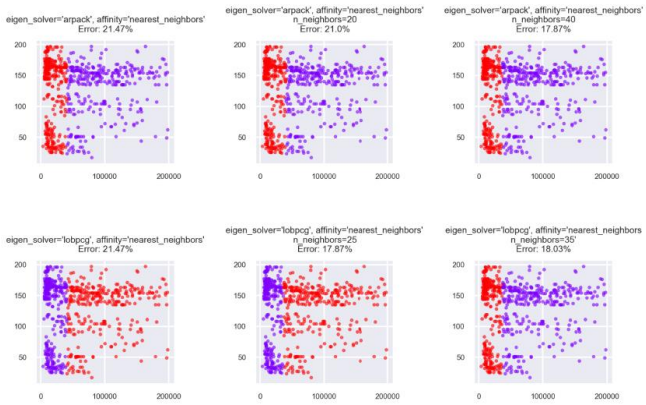

Figure 5. A comparison between different configurations of the Spectral Clustering algorithm.

After spectral clustering us analysis the hierarchical clustering parameter and then compare both the clustering algorithm.

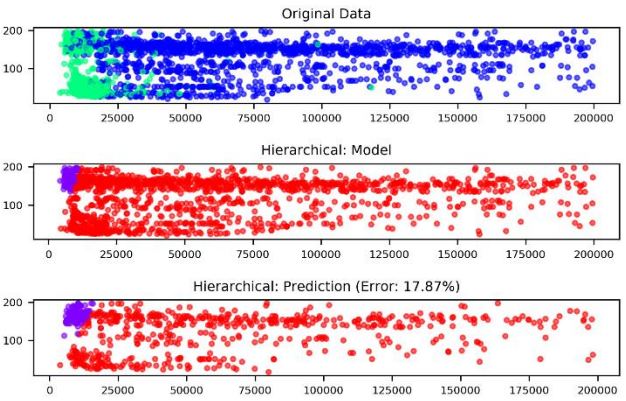

Figure 6. Simple Spectral clustering result

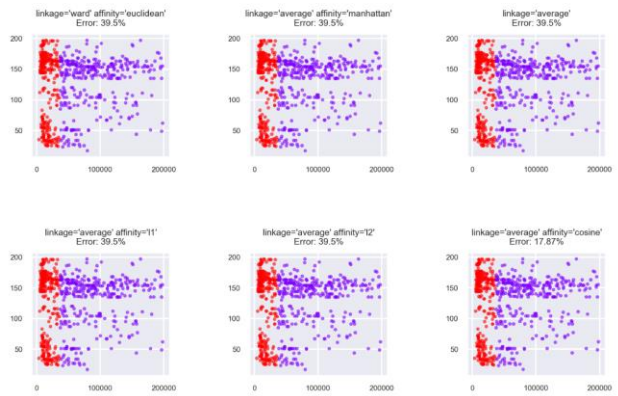

Figure 7. A comparison between different configurations of the Spectral Clustering algorithm.

The result of running a comparison between all the tested algorithms shows a similarity in accuracy between the Spectral Clustering algorithm and the Hierarchical Clustering algorithm (tied at **17.87%** miss clustered samples)

Dataset	Algorithm	Parameter	Result	
			Error Rate	Time Complexity
Real State Dataset	Spectral	arapack	21.47%	O(log n)
	Hirarchical	Lickage: Average	39.50%	O(log n)
	Spectral	LOBPCG	10%	O(2n)
	Hirarchical	Singlelink	15%	O(log n)
IRIS	Spectral	arapack	12.67%	O(log n)
	Hirarchical	Lickage: Average	10.09%	O(log n)
	Spectral	LOBPCG	29%	O(2n)
	Hirarchical	Singlelink	39%	O(log n)
	Hirarchical	Completelink	18%	O(n log n)

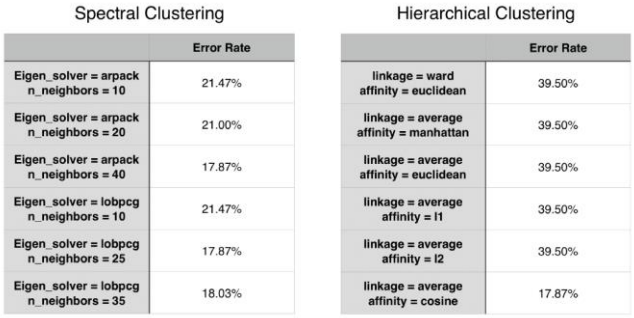

Figure 5. A comparison between error rates of different clustering methods.

VII. CONCLUSTION

We conclude that it is hardly possible to get a general clustering algorithm, which can work the best in clustering all types of datasets. Thus we tried to implement spectral clustering and hierarchical clustering algorithms which can work well in different types of datasets and compare that clustering algorithm with their parameter to analyse which clustering algorithm is better than other algorithm. In which the required classes are related to each other and we require a strong basis for each cluster with that result we can understand which clustering algorithm is better which types of dataset.

REFERENCES

[1] Khaled M. Hammouda, Mohamed S. Kamel , "Efficient phrase-based document indexing for web document clustering" , IEEE transactions on knowledge and data engineering, October 2004

[2]Clustering with multi-viewpoint based Similarity measure." IEEE transaction on knowledge and Data Engineering, vol. XX, No. YY2011.

[3] Active Clustering: Robust and Efficient Hierarchical Clustering using Adaptively Selected Similarities, Proceedings of 14th international conference on Artificial Intelligence and Statistics, 2011.

[4] Hierarchical Clustering, IEEE trans. on Knowl and Data Eng., April 2016.

[5] E.M. Voorhees. "Implementing agglomerative hierarchical clustering algorithms for use in document retrieval. Information Processing and Management", 22(6):465–476, 1986.

[6] Sun Dafei,Chen Guoli,Liu Wenju. The discussion of maximum likehood parameter estimation based on EM algorithm. Journal of HeNan University. 2002, 32(4):35~41

[7] Haojun sun, zhihui liu, lingjun kong, "A Document Clustering Method Based On Hierarchical Algorithm With Model Clustering, 22nd international conference on advanced information networking and applications.

[8] Shi zhong, joydeep ghosh, Generative Model-Based Document Clustering: A Comparative Study, The University Of Texas.

[9] T. Soni Madhulatha, "An overview of Clustering Methods",IOSR Journal of Engineering, Apr.2012, Vol. 2(4) pp: 719-725

[10] Prof. Neha Soni, Dr.Amit Ganatra. "Comparitive Study of Several Clustering Algorithms", International Journal of Advanced Computer Research,Volume-2,Number-4,Issue-6 December 2018.

[11] Yogita Rani and Dr.Harish Rohil, "A Study of Hierarchical Clustering Algorithm", International Journal of Information and Computation Technology.ISSN 0974-2239 Volume 3, Number 11 (2015), pp. 1225-123

[12] K.A.V.L.Prasanna and Mr. Vasantha Kumar, "Performance Evaluation of multiview-point based similarity measures for data clustering", Journal of Global Reasearch in Computer Science,

[13] K.Sathiyakumari, V.Preamsudha , "A Survey on Various Approaches in Document Clustering", Int. J. Comp. Tech. Appl., Vol 2 (5), 1534-1539