Volume 4, Issue 11, November 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Spectral Clustering: Advanced Clustering Techniques

[1]S. V. Suryanarayana (Ph.D), [2]Guttula Rama Krishna (M.Tech), [3]Dr. G. Venkateswara Rao (Ph.D)

[1]Assistant Prof, [3]Associate Professor

[1, 2]Department of CSE, GVVIT Bhimavaram, India

[3]Dept. of. Information Technology, GITAM Visakhapatnam, India

Abstract— Clustering is one of the widely using data mining technique that is used to place data elements into allied groups of "similar behavior". The conventional clustering algorithm called K-Means algorithm has some well-known problems, i.e., it does not work properly on clusters with not well defined centers, it is difficult to choose the number of clusters to construct different initial centers can lead to different resultant clusters.

Now a days,spectral clustering is becoming popular and widely used since its results overcomes the outcomes of the k-means algorithm. Spectral clustering is a more advanced clustering algorithm compared to k-means as it uses several mathematical concepts (i.e. weight matrices,similarity matrices,degree 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.

In this paper we are providing a detailed description on how the spectral clustering works i.e. it describes the steps and also shows the possible implementation. Furthermore, it shows the behavior of the algorithm on a few selected data sets. Finally, it gives a conclusion based on the observations obtained from the experiments.

Keywords—K-Means, Spectral Clustering, Eigenvalues, Eigenvectors, Laplacians.

I. INTRODUCTION

Clustering is the process of grouping data into clusters, so that objects within a cluster are similar while comparing with one another, but are very dissimilar to objects in other clusters. There are several approaches for clustering.

Clustering can be a stand-alone tools and also as a pretreatment process of the model algorithms. Clustering plays an vital role in pattern recognition and image processing. Cluster analysis as a data pre-treatment process is the basis of further analysis and data processing . It is also known as unsupervised learning process, since there is no prior knowledge about the data set. It also acts as an important data processing analysis and techniques which are aimed to complete the exploratory function. It is mainly used to study the individual or similarity distance or similarity measure, according to some clustering rules to divide the data set into different clusters, making the same cluster within high similarity and different clusters have a low similarity. In recent years, cluster analysis has been a hotspot to analyse data and extract information in pattern recognition and data analysis . There are many clustering algorithms, but each algorithm is optimized the certain aspects of data features, such as: minimize the within-class distance, maximize inter-class distance, etc. So far, there is no cluster algorithm can be used to reveal the structure of different multidimensional data sets . Each algorithm imposes some structure on the data set explicitly or implicitly, which results in the difficult clustering validity assessment. The traditional cluster algorithms, such as K-means, Fuzzy C means (FCM) algorithm, etc. most of these algorithms need to assume that cluster objects have some characteristics and it can form a number of different clusters, and based on the convex spherical sample space. But when the sample space is not convex, the algorithm will be trapped in local optimum. To solve this problem, a new clustering algorithm has been proposed, known as Spectral Clustering algorithm. Spectral Clustering algorithm is based on the spectra graph theory. They treat the data clustering as a graph partitioning problem without make any assumption on the form of the data clusters, namely, the clustering of data sets mapped to the Laplacian matrix's row vector. Not only can the n dimensional data sets convert into k-dimensional data sets (many times, k<<n), to achieve the purpose of dimensionality reduction, and have a good clustering results. Spectral Clustering algorithm is a point-to-point clustering algorithm. In recent years, Spectral Clustering algorithm is more increasingly widespread as a clustering analysis algorithm. it was originally used for parallel computing , VLSI design, load balancing and other areas, and now it is beginning to be used in machine learning. Currently, Spectral Clustering is attracted more attention in the field of text mining, information retrieval and image segmentation, and has achieved research results.

II. EXISTING SYSTEM: K-MEANS ALGORITHM

Let $X = \{i\}$, $i=1,...,n$ be the set of n d-dimensional points to be clustered into a set of K clusters, { k= 1,..., k} ,.K-means[1] algorithm finds a partition such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. Let be the mean of cluster. The squared error between and the points in cluster is defined as, $E(C_k) = \Sigma ||X_i - \mu_k||^2$

$x_i \epsilon C_k$

The goal of K-means[1] is to minimize the sum of the squared error over all the K clusters

$$E(C) = \sum_{K=1}^{K} \sum_{x_i \epsilon C_k} ||X_i - \mu_k||^2$$

Minimizing this objective function is known to be an NP-hard problem (even for K = 2). Thus K-means[1], which is a greedy algorithm, can only be expected to converge to a local minimum. K means algorithm starts with an initial partition with K clusters and assign patterns to clusters so as to reduce the squared error. Since the squared error tends to decrease with an increase in the number of clusters *K* (with $E(C) = 0$ when $K = n$), it can be minimized only for a fixed number of clusters.

The main steps of K-means[1] algorithm are as follows .

2.1 Algorithm:

1. Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.
2. Generate a new partition by assigning each pattern to its closest cluster center.
3. Compute new cluster centers.

Following figure shows an illustration of K-means algorithm on a 2-dimensional data set with three clusters.

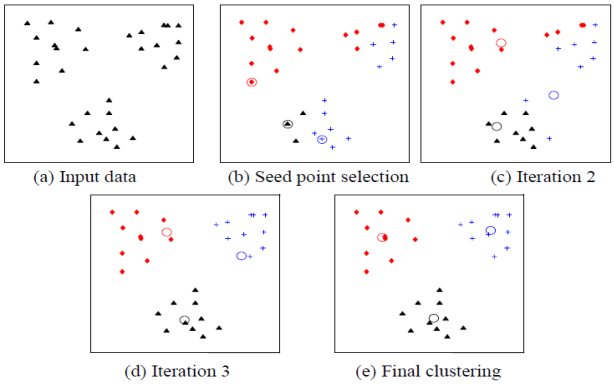

Figure 4 Illustration of K-means[1] algorithm. (a) Two-dimensional input data with three clusters; (b) three seed points selected as cluster centers and initial assignment of the data points to clusters; (c) & (d) intermediate iterations updating cluster labels and their centers; (e) final clustering obtained by K-means[1] algorithm at convergence.

2.2 Parameters of K-means

K-means[1] algorithm requires three user-specified parameters: number of clusters (*K*) ,cluster initialization, and distance measure technique. The most critical choice is K. While no mathematical criterion exists, a number of heuristics are available for choosing K. Typically, K-means[1] is run independently for different values of K and the partition that appears the most meaningful to the domain expert is selected. Different initializations can lead to different final clustering because K-means[1] only converges to local minima. One way to overcome the local minima is to run the K-means[1] algorithm, for a given K, with several different initial partitions and choose the partition with the smallest value of the squared error. K-means[1] is typically used with the Euclidean metric for computing the distance between points and cluster centers. As a result, K means[1] finds spherical or ball-shaped clusters in data. K-means with Mahalanobis distance metric has been used to detect hyper ellipsoidal clusters , but this comes at the expense of higher computation cost.

Drawbacks of K-means[1] algorithm:

1) Difficult to predict K-Value.
2) With global cluster, it didn't work well.
3) Different initial partitions can result in different final clusters.
4) It does not well with clusters (in original data) of different size and different density.

III. PROPOSED SYSTEM: SPECTRAL CLUSTERING ALGORITHM

Spectral clustering[3] 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.

3.1 Preliminaries of Spectral Clustering(according to M. KONG[5]):

3.1.1 Graph notation

Let G= (V, E) be an undirected graph with vertex set $V = \{v_1 \ldots v_n\}$. In the following we assume that the graph G is weighted, that is each edge between two vertices v_i and v_j carries a non-negative weight $w_{ij} \geq 0$. The weighted adjacency matrix of the graph is the matrix $W = (w_{ij})_{i,j=1 \ldots n}$. If $w_{ij} = 0$ this means that the vertices v_i and v_j are not connected by an edge. As G is undirected we require $w_{ij} = w_{ji}$. The degree of a vertex $v_i \in V$ is defined as,

$$d_i = \sum_{j=1}^{n} w_{ij}$$

3.1.2 Different similarity graphs

There are several popular constructions to transform a given set $x_1,...,x_n$ of data points with pairwise similarities s_{ij} or pairwise distances d_{ij} into a graph. When constructing similarity graphs the goal is to model the local neighbourhood relationships between the data points.

The *ε-neighbourhood graph* here we connect all points whose pairwise distances are smaller than ε. As the distances between all connected points are roughly of the same scale (at most ε), weighting the edges would not incorporate more information about the data to the graph. Hence, the ε- neighbourhood graph is usually considered as an unweighted graph.

K-nearest neighbour graphs Here the goal is to connect vertex v_i with vertex v_j if v_j is among the k-nearest neighbours of v_i. However, this definition leads to a directed graph, as the neighbourhood relationship is not symmetric. There are two ways of making this graph undirected. The first way is to simply ignore the directions of the edges, that is we connect v_i and v_j with an undirected edge if v_i is among the k-nearest neighbours of v_j or if v_j is among the k-nearest neighbours of vi. The resulting graph is what is usually called the k-nearest neighbour graph.

3.1.3 Graph Laplacians and their basic properties

The main tools for spectral clustering are graph Laplacian matrices. There exists a whole field dedicated to the study of those matrices, called spectral graph theory. In this section we want to define different graph Laplacians and point out their most important properties. We will carefully distinguish between different variants of graph Laplacians. Note that in the literature there is no unique convention which matrix exactly is called "graph Laplacian". Usually, every author just calls "his" matrix the graph Laplacian. Hence, a lot of care is needed when reading literature on graph Laplacians. In the following we always assume that G is an undirected, weighted graph with weight matrix W, where $w_{ij} = w_{ji} \geq 0$. When using eigenvectors of a matrix, we will not necessarily assume that they are normalized. For example, the constant vector 1 and a multiple a1 for some $a \neq 0$ will be considered as the same eigenvectors. Eigen values will always be ordered increasingly, respecting multiplicities. By "the first k eigenvectors[4]" we refer to the eigenvectors[4] corresponding to the k smallest eigenvalues[4].

3.2 Spectral Clustering Algorithm[3] :

- Construct a similarity. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- Compute the first k generalized eigenvectors [4]$u_1,...,u_k$ of the generalized Eigen problem $Lu=\lambda Du$.
- Let $U \in R^{n \times k}$ be the matrix containing the vectors $u_1,...,u_k$ as columns.
- For i = 1,...,n, let $y_i \in R^k$,be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,...,n}$ in R^k with the k-means[1] algorithm into clusters $C_1,...,C_k$. Output: Clusters $A_1,...,A_k$ with $A_i = \{j | y_j \in C_i\}$.

IV. RESULTS

To implement this Spectral clustering Algorithm we have used the open source tool WEKA. In original WEKA there is no tab called SpectralClusterer. We developed a java source code for this Spectral Clustering[3] and added to the WEKA library. We applied both Basic K-Means[1] algorithm and Spectral Clustering algorithm on the data sets which are available in UCI Machine Learning library and observed the number of clusters formed along with the number of instances within each cluster. We summarised our results in the form the following table.

Table: Results observed from WEKA for Spectral Clustering and K-Means

Dataset	K-Means			Spectral Clustering		
	No.of Clusters	No.of Instances	% of points in a cluster	No.of Clusters	No.of Instances	% of points in a cluster
Cars	2	185	46%	3	215	53%
		221	54%		108	27%
					83	20%
Contact-lenses	2	12	50%	4	2	8%
		12	50%		13	54%
					4	17%
					5	21%
Cpu	2	171	82%	1	209	100%
		38	18%			
Iris	2	100	67%	3	50	33%
		50	33%		50	33%
					50	33%
Iris-2D	2	100	67%	2	100	67%
		50	33%		50	33%
Prawn-crabs	2	106	53%	4	34	17%
		94	47%		36	18%
					87	44%
					43	22%
Boxing	2	55	46%	4	61	51%
		65	54%		23	19%
					17	14%
					19	16%
Birthday	2	117	32%	7	52	14%
					52	14%
					52	14%
					52	14%
		248	68%		52	14%
					52	14%
					53	15%
Aids	2	23	46%	4	5	10%
		27	54%		5	10%
					20	40%
					20	40%
Devils	2	54	66%	1	82	100%
		28	34%			

V. CONCLUSION & FUTURE SCOPE

Spectral Clustering[3] algorithm binds data mining, pattern recognition, mathematics, image processing and many other areas of research. From the above table it is clear that the Basic K-Means[1] algorithm always forms two clusters from all the data sets all the times. But, in case of spectral clustering it forms the different number of clusters corresponding to the given data set by computing eigen values. A variety of Spectral Clustering algorithms have their own advantages and disadvantages. Because of the actual complexity and the data diversity, each algorithm only can solve a set of problems. Therefore, the users should base on specific problems to select suitable clustering algorithm. In recent years, along with the development of conventional methods and the new emerging technologies in the field of data mining, machine learning and artificial intelligence, Spectral Clustering[3] algorithm has been considerable development.

REFERENCES

[1] Y. ZHU, H. YANG, L. SUN. Data mining technology. Nanjing: Southeast University Press, 2006.

[2] Z. TIAN, X. LI, Y. JU. The perturbation analysis of the Spectral clustering. Chinese Science, 2007, vol.37(4),pp.527-543

[3] R. GU, B. YE, W. XU. An improved spectral clustering algorithm. Computer Research and Development, 2007, vol.44,pp.145-149

[4] Y. Weiss. Segmentation using eigenvectors: A unified view. International Conference on Computer Vision, 1999

[5] M. KONG. Spectral Analysis and Clustering Relational Graphs.Hefei: Anhui University, 2006

[6] Z. WANG, G. LIU, E. CHEN. A spectral clustering algorithm based on fuzzy K-harmonic means. CAAI Transactions on Intelligent Systems, 2009, vol.4(2),pp.95-99

[7] Von Luxburg, U., *A Tutorial on Spectral Clustering*, Max-Planck-Institute for Biological Cybernetis, TR-149, August 2006.

[8] Von Luxburg, U., *Lecture slides on Clustering*, Max-Planck-Institute for Biological Cybernetis, http://velblod.videolectures.net/2007/pascal/bootcamp07_vilanova/luxburg_von_ulrike/luxburg_clustering_lectures.pdf, July 2007.

