A Study of Different Clustering Techniques

Ambika K Biradar
Assistant Professor
Department of Mathematics
Dr. D.Y .Patil Institute of Engineering & Technology, Pimpri, Pune-18

Abstract: Clustering is the process that groups the objects into clusters, such that objects within a cluster are more “similar” to each other than they are to the objects in the other clusters. Clustering huge amount of data is a difficult task, in order to handle large amount of data various clustering methods have been developed. This paper is intended to study and compare different data clustering algorithms: spectral clustering, density based clustering and complete gradient clustering algorithm.

Keywords: Clustering Techniques, Kernel Density estimators.

I. INTRODUCTION

Clustering is the most important technique for data mining which aims at partitioning a given population into groups or clusters with common characteristics, similar objects are grouped together, while dissimilar objects belong to different groups. As a result of this some new categories are been discovered which describes the population in a meaningful form. There are various clustering methods which are used based on similarity or dissimilarity between objects or clusters or based on size of the data or on density of the data.

In this paper we will be discussing about Spectral clustering method and complete gradient clustering algorithm. During cluster analysis one the important task is to find good clusters for further analysis. This process of cluster extraction can be done well using Gradient clustering algorithm whose basic idea depends on inflexion point which is the start or end point of any sub cluster.

II. SPECTRAL CLUSTERING ALGORITHM

Spectral clustering is a technique which is based on eigenstructure of the affinity matrix , matrix which partitions the data into clusters. It is one of the powerful tool which can easily separate non convex groups of the data. The algorithm constructs the affinity matrix from the data and then finds the eigen decomposition of the matrix and uses the clustering techniques to subspace of these eigenvectors. If there are L clusters then subspace is formed by considering first L independent eigenvectors thus it is necessary to have an prior knowledge of the number of clusters .

Procedures of the SCA:

Spectral clustering first constructs the affinity matrix A_{nxn} encoding the distance between these data points and then this matrix A is normalised to a new matrix N which is formed by conjugating matrix A with matrix $D^{1/2}$where D is the diagonal matrix whose diagonal elements are the sum of the row elements of matrix A.Then find L eigenvectors of the matrix N those with largest eigenvalue, and arrange these eigenvectors in the columns and form a matrix B. The rows of matrix B are normalised to unit length, and considered to be L-dimensional space, so that each of data points can be assigned to a location in a *L*-dimensional space using K mean clustering. This L- dimensional space is the clustering space,which has formed *L* convex clusters.

Spectral clustering technique relies on the structure of the eigenvectors of the Laplacian of the similarity matrix. In order to do these computation it is necessary to have n(n-1)/2
similarities between the n objects in a cluster and Also many times computing spectral decomposition, on large datasets is difficult.Spectral clustering algorithm
provide concrete results on itsperformance, showing that, under some assumptions ,this algorithm recovers all clusters of size $\Omega(\log n)$ using $O(n \log(\log n)))$ similarities and runs in $O(n \log(\log(\log n)))$ time for a dataset of n objects.

III. DENSITY BASED CLUSTERING ALGORITHM

In density-based clustering, clusters are defined as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed.(2)
These algorithms group objects according to specific density objective functions. Density is usually defined as the number of objects in a particular neighbourhood of a data objects. In these approaches a given cluster continues growing as long as the number of objects in the neighbourhood exceeds some parameter. The most popular density based clustering method is DBSCAN(Density-Based Spatial Clustering of Applications with Noise)which uses the concept of **density reachability** and **density connectivity**.[**Density Reachability** - A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbours which are within distance ε. **Density Connectivity** - A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".]

Time complexity of DBSCAN is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times.[10] The runtime of the algorithm is of the order O(n log n) if region queries are efficiently supported by spatial index structures, i.e. at least in moderately dimensional spaces. OPTICS is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter, and produces a hierarchical result related to that of linkage clustering. (2).Density-Link-Clustering combines ideas from single-linkage clustering and OPTICS, eliminating the parameter entirely and offering performance improvements over OPTICS by using a tree index. Since DBSCAN algorithm fails in case of varying density clusters, DBSCAN and OPTICS expect some kind of density drop to detect cluster borders. Moreover they cannot detect an intrinsic cluster structure which is prevalent in majority of real life data.(for example: it fails in case of neck type of dataset).

IV. COMPLETE GRADIENT CLUSTERING ALGORITHM

In this algorithm each cluster is identified by the humps of Kernel density estimator of the data. (1) Kernel density estimation: Suppose x_1, x_2, x_3, x_4, ……. x_m are m elements in n-dimensional space with probability density function f then kernel density estimator $\overline{f} : R^n \rightarrow [0,\infty)$ is defined as $\overline{f} = \frac{1}{mh^n}\sum_{i=1}^{i=m} K\left(\frac{x - x_1}{h}\right)$
Where h>0 is called smoothing parameter and K is called as kernel which is measurable function $K : R^n \rightarrow [0,\infty)$ with unit integral and is unimodal and symmetrical with respect to zero [Measurable function:] Most popular Kernel used in practice is standard normal kernel $K(x) = \frac{1}{2\pi^{n/2}} e^{-\frac{x^T x}{2}}$ and the choice of smoothing parameter h is made with the criterion of mean integrated square. The value of h is chosen such that it minimizes the mean integrated square error

MISE $\overline{f} = E\int(\overline{f}(x) - f(x))^2 dx$

Procedures of the CGCA(1)

Consider the data set containing m samples x_1, x_2, . . . , xm in n-dimensional space.
Using the methodology of the kernel density estimator $\overline{f}$ may be constructed.

given the start points: $x^0{}_j = x_j$ for j = 1,2, . . . ,m,
each point is moved in an uphill gradient direction using the following iterative formula:

$$x^{k+1}{}_j = x^k{}_j + b\frac{\nabla\widehat{f}(x^k{}_j)}{\widehat{f}(x^k{}_j)} \quad \text{for j = 1,2, . . . ,m, and k = 0,1, . . .}$$

where $\nabla\widehat{f}$ denotes the gradient of kernel estimator and parameter $b = h^2/(n+2)$

To complete the algorithm the following two aspects need to be specified: a termination criterion of the algorithm and procedure of creating clusters.
The algorithm will be stopped when the following condition is fulfilled: |Dk −Dk−1| < aD0, where D0 and Dk−1, Dk denote sums of distances between particular elements of set x1, x2, . . . , xm before starting the algorithm as well as after the (k−1)-th and k-th step, respectively. The positive parameter a is taken arbitrary and the value 0.001 is recommended. This k-th step is the last one and will be denoted by k*. Finally, after the k*th step of algorithm the set:
$x^{k*}{}_{1,}\ x^{k*}{}_2\ ,\ x^{k*}{}_{3,}\ldots\ldots\ldots\ldots\ldots\ x^{k*}{}_m$

considered as the new representation of all samples x1, x2, . . . , xm, is obtained. Following this, the set of mutual distances of the above elements: is defined. Then for these new sample points Auxiliary Kernel estimator is found according to the definition mentioned above.local minimum of this new Kernel estimator belonging to (0,D) is found where D means the maximum value of the points of the set.(6)

This local minimum will be denoted as xd, and it can be interpreted as half the distance between “centers” of potential clusters lying closest together. Finally, the clusters will be created. First, the element of set will be taken; it initially create a one-element cluster containing it. An
element of set is added to the cluster if the distance between it and any element
belonging to the cluster is less than xd. .

V. CONCLUSION AND FUTURE WORK

The CGCA clustering algorithm, based on kernel estimator methodology, is expected to be an effective technique in comparision to other techniques . It behaves equally or better than the spectral clustering algorithm. Further research is needed for the ability to identify good kernels.

REFERENCES

[1] ‘A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images’ Małgorzata Charytanowicz, Jerzy Niewczas, Piotr A. Kowalski, Piotr Kulczycki,Szymon Łukasik, and Sławomir ˙ Zak‘Density-Based Data Analysis and SimilaritySearch’

[2] Hans-Peter Kriegel, Stefan Brecheisen, Peer Kr¨oger, Martin Pfeifle, Matthias Schubert, and Arthur Zimek

[3] Information cut for clustering using a gradient descent approach’ Robert Jenssena,∗, Deniz Erdogmusb, Kenneth E. Hild IIc, Jose C. Principed, TorbjZrn Eltofta

[4] Amandeep Kaur Mann and Navneet Kaur “ Survey Paper on Clustering Techniques” International Journal of Science, Engineering and Technology Research Vol. 2 Issue 4, April 2013

[5] Murtagh, F. “A survey of recent advances in hierarchical clustering algorithms which use cluster centers”. Comput. J. 26 354-359, 1984

[6] Fast Clustering based on KernelDensity Estimation’Alexander Hinneburg1 and Hans-Henning Gabriel

[7] On Data Clustering Analysis: Scalability,Constraints and Validation’Osmar R. Za¨ane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang

[8] Efficient Active Algorithms for Hierarchical Clustering’ Akshay Krishnamurthy ,Sivaraman Balakrishnan, Min Xu min,Aarti Singh ,Carnegie Mellon University 5000 Forbes Avenue, Pittsburgh, PA 15213