Improved Clustering using Hierarchical Approach

Megha Gupta, M.Tech Scholar,RTU
Rajasthan, India

Vishal Shrivastava, Professor, RTU
Rajasthan, India

Abstract: **Clustering is the process of partitioning a set of data so that the data can be divided into subsets. Clustering is implemented so that same set of data can be collected on one side and other set of data can be collected on the other end. Clustering can be done using many methods like partitioning methods, hierarchical methods, density based method. Hierarchical method creates a hierarchical decomposition of the given set of data objects. In each successive iteration, a cluster is split into smaller clusters, until eventually each object is in one cluster, or a termination condition holds.**

In this paper, partitioning method has been used with hierarchical method to form better and improved clusters. We have used various algorithms for getting better and improved clusters.

***Keywords: Clustering, Hierarchical, Partitioning methods*.**

*

I. Introduction

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful

information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.

Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.

Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra-class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

II. Related Work

Performance guarantees for hierarchical clustering

Hierarchical clustering is the recursive partitioning of n data points into 2, 3, 4… and further converting them into clusters. Each intermediate cluster is made more refined by dividing them into furtherclusters. There must always exist ahierarchical clustering in which, for every k, the induced k-clustering is close to the optimal k-clustering under some reasonable cost function. The optimal based k-clustering cannot be obtained by merging clusters of the (k+1)-clustering.

Theorem 1: Take the cost of a clustering to be the largest radius of its clusters. Then, any data set in any metric space has a hierarchical clustering in which, for each k, the induced k-clustering has cost at most eight times that of the optimal k-clustering.

An algorithm for constructing such a hierarchy which is similar in simplicity and efficiency to standard heuristics for hierarchical clustering has been used. The algorithm that has been implemented is known as farthest first traversal of a set of points, used by Gonzalez [1] as an approximation for closely-related k-center problem.

Theorem 2: In the setting of the previous theorem, there is a randomized algorithm which produces a hierarchical clustering such that, for each k, the induced k-clustering has expected cost at most 2e=5.44 times that of the optimal k-clustering.

The most common heuristic for hierarchical clustering work bottom-up, starting with a separarte cluster for each point, and then it is merged into two closest clusters until a single cluster is left.

In single linkage clustering, the distance between two clusters is the distance between their closest pair of points. In complete-linkage clustering, it is the distance between their farthest pair of points. Average linkage has many variants in which the distance between clusters is the distance between their means [2].

Limitations

I. Small values of k:

It is sufficient to guarantee good k-clustering just for small values of k, say in hundred or say on or in more smaller values. The lower bound and upper bound on the approximation ratio of average and complete linkage would be Ώ (log k).

II. Efficiency

Can this algorithm or any standard heuristics that have been considered, be implemented in o(n^2) time for data sets of size n? Results of Borodin et al.[3] and Thoroup [4] offer some hope here.

Evaluation of Agglomerative Hierarchical Clustering Methods

Traffic data are the foundation of highway transportation planning and are used to assist highway engineers in maintaining and designing safe, efficient, and cost effective facilities [5]. It is well known that traffic variations occur at different time scales e.g. Time of day, day of week, and season (month) of the year [6]. So, it is important to accurately interpret the temporal variation effects on collected traffic data in order to achieve better design decisions.

The following agglomerative hierarchical clustering methods are available in SAS (Statistical Analysis System) Version 8 [7] for quantifying the distance (or dissimilarity) between two clusters.

1. Average Linkage (AVE)
2. Centroid method (CEN)
3. Flexible-beta Method (FLE)
4. McQuitty’s Similarity Analysis (MCQ)
5. Median Method (MED)
6. Single Linkage (SIN)
7. Ward’s Minimum-Variance Method (WAR)

Motivations

Using hierarchical clustering, better clusters will be formed. The clusters formed will appear in better way and there will be tight bonding in between them. It means that the clusters formed will be refined using the various algorithm of hierarchical clustering.

III. Problem Statement

The objective of the proposed work is to perform hierarchical clustering to obtain the more refined clusters with strong relationship between members of same cluster

IV. Proposed Approach

In this paper, we have used K-means algorithm and CFNG to find better and improved clusters.

K-means Algorithm

Suppose a data set, D, contains n objects in Euclidean space. Partitioning methods distribute the objects into k clusters, Ci…..$C_{k,}$ that is, $C_i \subset D$ and $C_i \cap C_j$=Ø for ($1\leq$ i, j$\leq$k). An objective function is used to access the partitioning quality so that objects within a cluster are similar to one another but dissimilar to objects in other clusters. This is, the objective function aims for high intracluster similarity and low intercluster similarity [8].

A centroid-based partitioning technique uses the centroid of a cluster, $C_{i,}$ to represent that cluster. The centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoids of the objects assigned to the cluster. The difference between an object p and $C_{i,}$ the representation of the cluster, is measured by dist(p,C_i), where dist(x,y) is the Euclidean distance between two points x and y.

CFNG

Colored farthest neighbor graph shares many characteristics with SFN (shared farthest neighbors) by Rovetta and Masulli [9]. This algorithm yields binary partitions of objects into subsets, whereas number of subsets obtained by SFN can vary. The SFN algorithm can easily split a cluster where no natural partition is necessary, while the CFNG often avoids such splits.

V. Results and Analysis

In figure 1, cluster 1 represents the values of strong clusters that have been formed using K-Means algorithm whereasthe weak cluster formed is further divided into two clusters using CFNG. The values of strong and weak cluster formed in CFNG are represented with cluster 2 and cluster 3 respectively.

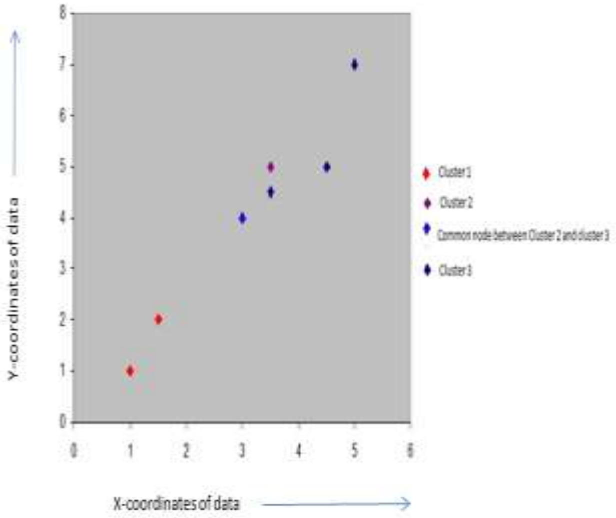

Figure 1: Final Results of Hierarchical Clustering

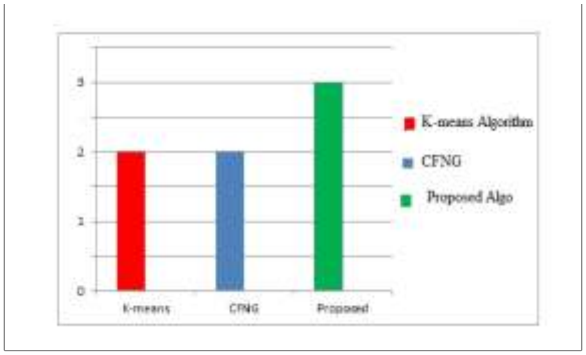

Figure 2: Comparison of Proposed Algorithm with K-means and CFNG

Above figure represents that using K-means we have obtained 2 clusters in which one is very strong and other one is very weak. The using CFNG, we have obtained further 2 clusters. In final result we got 3 clusters, one strong cluster from K-means algorithm and two clusters from CFNG.

VI. Conclusion and Future Scope

We have obtained better and improved clusters using K-means and CFNG algorithms hierarchically. The final clusters obtained are tightly bonded with each other.

In this paper, we have used 2 different algorithms for hierarchical clustering.Instead of using CFNG, we could have used other hierarchical clustering algorithm.

References

[1] T.F. Gonzalez, "Clustering to minimize the maximum interclusterdistance."Theoretical Computer Science,38:293-306,1985.

[2] M.B.Eisen, P.T.Spellman, P.D.Brown, &D.Botstein. "Cluster analysis & display of Genome-wide expression patterns.Proceedings" of National Academy of Science, 95:14863-14868, 1998.

[3] A.Borodin, R.Ostrovsky& Y. Rabani. "Subquadratic approximation algorithm for clustering problem in high dimensional spaces."Proceedings of 31st ACM Symposium on Theory of Computation, 1999.

[4] M.Thoroup, "Quick K-median, k-center & facility location for sparse graphs. International Colloquim on Automata, Languages & Programming," 2001.

[5] Project Traffic Forecasting Handbook. Florida Department of Transportation. Tallahassee. FL.April 2002.

[6] Traffic Monitoring Guide, Federal Highway Administration. U.S Department of Transportation, Washington D.C May 2001.

[7] SAS- Online DOC, Version 8, SAS Institute Inc., Cary, NC, 1999.

[8] Han and Kamber, "Data Mining and Concepts".

[9] S.Rovetta, F.Masulli, "Shared farthest neighbor approach to clustering of high dimensionality, low cardinality data, Pattern Recognition " 39 (12) (2006) 2415-2425.