An Efficient Linkage Criterion for Creating Clusters in Hierarchical Method

P Praveen

Associate Professor in Department of CSE, S R Engineering College, Warangal, Telangana State, India.

Abstract

The Clustering process has been performed in statistics over many years. Many algorithms were developed based on the conception of similarity or distance, so item cluster that are connected or on the subject of each alternative one another and totally different or far from other objects could also be place organized in a very cluster. The tactic might seem simple, however it is somewhat difficult as a result of to find objects that are dissimilar between an out sized cluster of objects desires comparison every object with each alternative object which can be very comfortable for giant knowledge sets. evaluates mean values attained in between the pair of clusters by considering k mean observations attained from them that is using single linkage (nearest neighbor method), complete linkage (furthest neighbor method), average linkage method and Average weighted linkage method were analyzed. Here some clustering techniques and their applications have been discussed. It also describes the necessities to be calculated for constructing an well-organized clustering algorithm to handle the huge data sets.

Keywords: Classification, Clustering, Data Mining, Hierarchical, Linkage criteria.

I INTRODUCTION

Various algorithms can be used for Cluster analysis that be different expressively in their perception of creates a group also how to proficiently discover them. Prevalent ideas of clusters contain groups with diminutive distances between the group objects, data space in dense areas, specific numerical distributions. The groups can thus expressed as a multi-objective optimization difficulty [6].To get an anticipated results and individual dataset, a suitable clustering algorithm and parameter settings should be used. The distance function to use, no. of probabilistic clusters or a density threshold are the values used for clustering methods. This research is a stepping stone of optimization as it involves trial and failure concept. Cluster analysis as such is not an

automatic task, but repetitious process of knowledge finding or interactive multi objective optimization [7]. To achieve the result with expected properties, the modification of data preprocessing and model parameters is necessary.All the above mentioned approaches ensure its own specialization and traditional methods which are commonly used in the process of data mining real application.
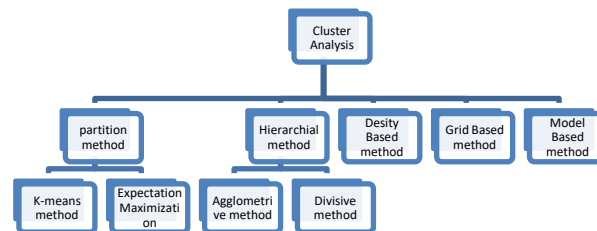

Figure 1.1 Categorization of Cluster Algorithms

Partitioning methods: This method is used to find a single level partition of objects. To get a local optimum solution, these methods are used repeatedly and are generally based on greedy heuristics. Given n objects, these strategies create k<=n clusters of knowledge and use a repetitive relocation technique. it's

assumed that every cluster has a minimum of one object and every object belongs to just one cluster[9]. Objects could also be moved amongst clusters, because the clusters area unit advanced. These strategies usually need variety the amount the quantity of clusters be such that a priori and this number, sometimes doesn't amendment throughout the process. K-Means and Expectation Maximization (EM) and K-means are general algorithms of partitioning method.

Hierarchical methods: This method attains a nested partition of the objects future in an exceedingly cluster tree. These strategies either begin with one cluster or then split into smaller and smaller clusters or begin with every object in a private cluster and so try and merge similar clusters into larger and bigger clusters (called clustered or bottom up). During this approach, in distinction to partitioning, tentative clusters could also be unified or split supported some criteria [45]. A hierarchical algorithm yields A tree of clusters referred to as dendrogram representing the nested grouping of objects and similarity levels at that groupings modification. Figure 1.3 shows a dendrogram representing Agglomerative and Divisive clustering process applied to a set of data objects,

Agglomerative method: It's additionally called as AGNES (Agglomerative Nesting). It works in a bottom-up manner. i.e, every item is at initially as a independent group (leaf). At every step the items are merged which are two clusters that neighborhood unit the foremost similar neighborhood unit conquered into a replacement larger cluster (nodes). This process is recurring till whole items are in a single huge cluster (root) (see figure below).

Divisive method: It's termed as DIANA (Divise Analysis) it works on the principle of top down approach. It is a reverse approach of AGNES. It starts with the root with all item are involved in a lone cluster. During every iteration, the most diverse cluster is divided into two groups here one group is called left cluster and another group is called as right cluster. This processes will continue till objects are in their singleton.
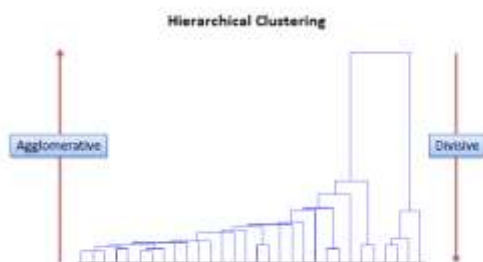

Figure 1.2Dendrogram obtained using Agglomerative and Divisive clustering algorithm

Divisive clustering algorithms are less time consuming than agglomerative clustering algorithms. However, in normal practice, they are used less often owing to the complex problems of choosing a cluster split and then determining the optimal subdivision of the selected cluster [1-3].

Density-based methods: These methods are distinctive to identify an object in the cluster at a least no. of objects should occur inside a given radius. These strategies will contend with discretional form clusters since the most demand of such strategies is that every cluster may be a dense region of points encircled by regions of density[2]. DBSCAN is the most general density based clustering method is DBSCAN, it is centered on connecting points within distance threshold and its complexity is fairly low.

Grid-based Methods: This method works in the object space as opposed to the information is partitioned keen on a lattice. Grid partitioning is depends on the qualities of the information and such strategies can manage non-numeric information all the more effortlessly. The data order will not affect this method[14][5]. This methodology uses a multi resolution grid system. It quantizes the expose address into

a restricted variety of cells that define a grid arrangement on which all of the clustering performance are done.

Model-based methods: it is a prediction method based on a prospect allocation. Basically, this algorithm attempts to form groups by way of elevated similarity stage inside a cluster and a near to the ground similarity stage among clusters[10][12]. These measurements of similarity level are built on the mean values and the algorithm attempts to reduce the squared error function.

II Related Work

Hierarchical algorithms create clusters recursively by dividing a database D of N objects into a number of levels of nested partitioning, denoted by a dendrogram[1]. A dendrogram is a two-dimensional diagram or tree and gives a complete hierarchical description of how objects are similar to each other on different levels. It can be examined at a particular level to represent a different clustering of the data[13]. There are two types of hierarchical algorithms: agglomerative algorithms and divisive algorithms. Agglomerative algorithms build the tree bottom-up, i.e. merging the N objects into groups. Divisive algorithms build the tree up-bottom by separating the N objects into finer clusters [7]. Bottom-up or agglomerative clustering, the more commonly used technique, treats each object as a cluster of size 1.

Then, it merges the two nearest objects in a cluster of size two and so on to reach one cluster combining all the objects unless other termination condition is Satisfied [6][8]. The up-bottom or divisive strategy does the reverse by starting with all of the N objects in one cluster and subdividing them into smaller groups until a termination condition is met such as a desired number of clusters or it stops when each object forms a cluster. This strategy of hierarchical algorithms, up-bottom, is used less often. Kaufman and Rousseeuw [11] remarked that divisive methods have been largely ignored in the literature mostly due to computational limitations.

The computational demands of these techniques is $O(2^N)$ so grow exponentially as the number of objects, N, raises. Hierarchical algorithms differ in the ways they determine the similarity between two clusters. There are three main ways to consider the distance between the two clusters: the single-linkage method, the complete-linkage method and the average- linkage method. The following formulas define four distance measures required to distinguish between the three linkages. They measure the distance between two clusters, C_x and C_y that have $|C_x|$ and $|C_y|$ objects respectively, where $dist(O_i, O_j)$ is the distance between two objects O_i and O_j and $dist(\mu_{C_x}, \mu_{C_y})$ is the distance between the mean values of objects belonging to cluster C_x and cluster C_y[14].

The single-linkage takes the shortest pair wise distance between objects in two different clusters by using the minimum distance. In contrast, complete-linkage takes the longest distance between the objects by using the maximum distance, while the average-linkage takes the average of the pairwise distances between all pair of objects coming from each of the two clusters[9]. The latter type of linkage, average-linkage, may use the mean or the average distance. Whereas the mean distance is simpler to calculate, the average distance is advantageous as it can be used to deal with categorical data.

The complete-linkage methods often generate more compact clusters and more useful hierarchical structure than the single-linkage methods. Having said that, the latter methods are more versatile[15].

Guha et al[47] have discussed the disadvantages of single-linkage and average-linkage methods. They stated that chaining effect is the main drawback of single-linkage clustering. This happens when a few points form a bridge between two clusters which enforce this type of methods to unify the two clusters. Elongated clusters mislead average-linkage clustering according to Guha et al.

Most of the hierarchical algorithms join two clusters or divide a cluster into two sub- clusters. However, some algorithms work in a similar manner but with more than two clusters or sub-clusters.

Thus, hierarchical clustering merges smaller clusters into larger ones or splits larger clusters into smaller ones recursively. hierarchical clustering algorithms are agglomerative and divisive clustering [13], Table 3.3 summaries the main differences between these clustering algorithms with regards to the data type they support and the computational cost, where N is the number of objects in the dataset. In addition, it includes the shape of clusters they handle as well as the input and output of the algorithms.

III Linkage Criteria

In the present day scenario some of the well-known and extensively used clustering algorithms are hierarchical algorithms which is considered in this paper as a major topic for the comparative study as the hierarchical algorithm consists of four types: Linkage single, Complete, Average and Average weighted The choices that are made in this paper is implemented to make comparisons between different linkages that are possible as the average and complete are two of the most commonly used approaches where furthermore, since the inception of Euclidean distance it is most widely used distance measure algorithm as the same is also chosen in this paper..

Hierarchical Algorithm descriptions

Single Linkage: (Figure 3.a) represents the Single linkage which is also called as the nearest-neighbor technique that tends to select the attained distance between the nearest or closest observations in clusters as shown in below Equation (3.1) represents objective linkage for single linkage.

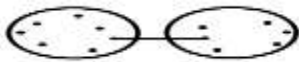$$d_{C_u,C_v} = \arg\min_{(u,v)}\left(\min_{x\in C_u, y\in C_v} D(x,y)\right) \tag{3.1}$$

Figure 3.a Single Linkage

3. 2. Complete Linkage: (Figure 3.b) complete linkage is also called as the furthest-neighbor technique that selects various distance between the farthest observations in clusters as implemented in the Equation (6.2) which is Objective function of complete linkage:

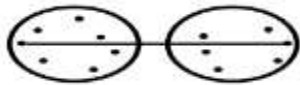$$d_{C_u,C_v} = \arg\min_{(u,v)}\left(\max_{x\in C_u, y\in C_v} D(x,y)\right) \tag{3.2}$$

Figure 4.b Complete Linkage

.3 Average Linkage: (Figure 4.c) Average linkage calculates the distances attained between all the pairs of various observations generated in clusters and the average implemented from all of these distances attained are as shown in Equation (4.3) which is Objective function of average linkage:

$$d_{C_u,C_v} = \arg\min_{(u,v)}\left(\frac{1}{|C_u|}\frac{1}{|C_v|}\sum_{x\in C_u}\sum_{y\in C_v} D(x,y)\right) \tag{3.3}$$
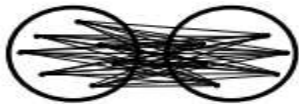

Figure 4c. Average Linkage

3.4 Average Weighted Linkage: (Figure.3.d) Average linkage calculates the distances with corresponding weights attained between all the pairs of various observations generated in clusters with weights and the

average is implemented from all of these weighted distances attained are as shown in Equation (6.4) which is Objective function of average linkage:

$$d_{C_u \mathcal{L}_v} = \frac{\sum_{x \in X_1, y \in X_2} d(x,y) \cdot w(x) \cdot w(y)}{w(X_1) \cdot w(X_2)} \qquad (3.4)$$
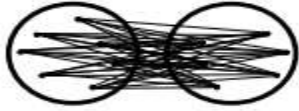

Figure 3.c Average Weighted Linkage

IV Experimental study

Comparative study of single linkage and complete Link and Average Linakges
In this paper, the proposed method was compared with traditional linkage methods such as single , complete , average and average weighted methods where the data set will be loaded and expanded as a dataset that can be considerably accessed from UCI repository where the possible application will access and implement the agglomerative hierarchical clustering using the JAVA programming language as the expanded application reads data from a dataset and includes five different methods to implement clustering of data such as: single, complete, average and average weighted with the proposed MB_Divclues algorithm.

Dataset Description
In the experimental study all the five different datasets which are well-known and broadly used in most of the data mining algorithms are being selected to demonstrate the capabilities as the dataset name obtained from UCI Machine Learning warehouse which can be easily accessed from the web site: http://archive.ics.uci.edu/ml/datasets/html.

Experimental study methods have been used for the first time to improve the clustering validation and quality that tend to measure the cluster validity with the cluster labels that match with the externally or previously known class labels that are attached and evaluated with regards of performance while rating the clustering results or in simple terms the performance is the total computational complexity that is attained while creating the singleton clusters.

In the process of estimating the proposed method the experiences that are gained towards the selection or choice of datasets preferred over selection of single linkage and complete linkage method in terms of performance rate that is time complexity is represented in the below figure. The obtained outcome illustrates that the projected normally more victorious than that of the single linkage method with respect to the performance speed though single linkage clustering may appear preferable as an optimal with admiration to the wrong principle in a lot of clustering applications as the single linkage clustering and complete algorithms reduces the appraisal of cluster superiority as complete linkage takes more time than that of single linkage but completely checks the data set, by combining the measures attained rigorously for identifying the overall allotment of the clusters.

Table: 4.1 Runtime calculation of linkage criteria.

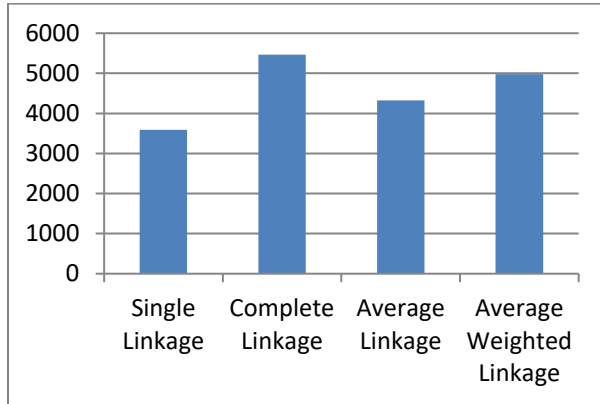| Linkage Criteria	Execution Time
Single Linkage	3591
Complete Linkage	5463
Average Linkage	4322
Average Weighted Linkage	4978

Figure 4.2 Runtime calculation of single linkage, complete linkage and Average Linkages.

The above figure (4.3) represents the time elapsed for creating clusters in mille seconds where the cluster numbers are also represented using the Single cluster strategy using attribute X. where the start time is 31216 and finish time is 34807 and the total elapsed time is 3591.

the cluster numbers are also represented using the Complete cluster strategy using attribute X. where the start time is 56588 and finish time is 62051 and the total elapsed time is 5463. Complete linkage takes more time than that of single linkage. As well as we complete the average and average weighted linkage criteria these also takes more than that of single linkage.

V Conclusion

In this paper four hierarchical clustering methods have been considered where the objects using a bottom-up approach that generates clusters that contains only one item and splits into two most similar clusters continuously based on a similarity metric provided by the algorithm. that evaluates mean values attained in between the pair of clusters by considering k mean observations attained from them that is using single linkage (nearest neighbor method), complete linkage (furthest neighbor method), average linkage method and Average weighted linkage method were analyzed. Analyzes the performance of the above clustering algorithms works as nearly twice as fast as that of complete linkage hierarchical algorithms. So finally single linkage is efficient than other linkage clustering approaches.

References

[1] O. A. Akeem, T. K. Ogunyinka, and B. L. Abimbola. A framework for multime- dia data mining in information technology environment. International Journal of Computer Science and Information Security (IJCSIS), 10(5):69–77, 2012.

[2] R. A. Baeza-Yates and B. Ribeiro-Neto. Modern information retrieval. Addison- Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[3] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Sym- posium on Principles of Database Systems, PODS '08, pages 191–200, New York, NY, USA, 2008. ACM.

[4] D. C. Duro, S. E. Franklin, and M. G. Dubé. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the clas- sification of agricultural landscapes using spot-5 {HRG} imagery. Remote Sensing of Environment, 118(3):259 – 272, 2012.

[5] Praveen P., Rama B. (2018) A Novel Approach to Improve the Performance of Divisive Clustering-BST. In: Satapathy S., Bhateja V., Raju K., Janakiramaiah B. (eds) Data Engineering and Intelligent Computing. Advances in Intelligent Systems and Computing, vol 542. Springer,

Singapore

[6]   N. Bhatia and V. Ashev, "Survey of nearest-neighbor techniques," International Journal of Computer Science and Information Security, vol. 8, no. 2, pp. 302–305, 2010.

[7]   R. Ravi Kumar, M. Babu Reddy and P. Praveen, "A review of feature subset selection on unsupervised lea*rning,"* 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)*,* Chennai,2017,pp.163-167.doi: 10.1109/AEEICB.2017.7972404.

[8] J P. Praveen and B. Rama, "An empirical comparison of Clustering using hierarchical methods and K-means," 2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics(AEEICB),Chennai,2016,pp.445-449.doi: 10.1109/AEEICB.2016.7538328

[9]   S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithmforapproximate nearest neighbor searching in fixed dimensions,"Journal of the ACM, vol. 45, no. 6, pp. 891–923, 1998.

[10]   M. Sheshikala, D. Rajeswara Rao and R. Vijaya Prakash, Computation Analysis for Finding Co–Location Patterns using Map–Reduce Framework, Indian Journal of Science and Technology, Vol 10(8), DOI: 10.17485/ijst/2017/v10i8/106709, February 2017.

[11]   C. Silpa-Anan and R. Hartley, "Optimised KD -trees for fast image descriptor matching," inIEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008,   pp. 1–8.

[12]   Mohammed Ali Shaik, P.Praveen, Dr.R.Vijaya Prakash, "Novel Classification Scheme for Multi Agents", Asian Journal of Computer Science and Technology, ISSN: 2249-0701 Vol.8 No.S3, 2019, pp. 54-58.

[13]   P. Praveen, B. Rama and T. Sampath Kumar, "An efficient clustering algorithm of minimum Spanning Tree," 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), Chennai, 2017, pp. 131-135. doi: 10.1109/AEEICB.2017.7972398

[14]   P. Indyk and A. Andoni, "Near-optimal hashing algorithms for approximate nearest   neighbor   in high dimensions," in Foundations of Computer Science, 2006, vol. 51,   no. 1, pp. 117–122.

[15]   SallauddinMohmmad,Dr.M.Sheshikala,Shabana,"Software   Defined   Security   (SDSec):Reliable centralized security system to decentralized applications in SDN and their challenges",Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 10-Special Issue, 2018, pp. (147-152).