The comparative study on agglomerative hierarchical clustering using numerical data

P Praveen1, Ranjith Kumar M2, Mohammed Ali Shaik3, R Ravikumar4, R Kiran5
1, 3, 4 Department of Computer Science & Engineering, SR University, Warangal, Telangana, India
2 Department of Computer Science & Engineering, Sumathi Reddy Institute of Technology for Women, Warangal, Telangana, India
3 Department of Electronics and Communication Engineering, Jyothismthi Engineering College, Karimnagar, Telangana, India
Email: prawin1731@gmail.com

Abstract. In the conventional way of convert data into a singleton or merging has many drawbacks mainly computational complexity. In this context hierarchical clustering method for quantitative measures of similarity among objects that could keep not only the structure of categorical attributes but also relative distance of numeric values. For numeric data the number of clusters can be validated through integral data, the hierarchical and partitioning methods the relationships among categorical items. In this paper we herein investigate linkage criteria in hierarchical clustering algorithm performance calculations using with Euclidian distance measure and some clustering techniques and their applications have been discussed. It also describes the necessities to be calculated for constructing a well-organized to handle the huge data sets. As the study initially investigates distinct issues for creating clusters with numeric attributes. The efficiency is obtained by clustering of datasets that comprises of numeric attributes related to distinct applications. Significant issues such merging object naturally uses Euclidean distance is resolved by using Agglomerative methods.

Keywords: Classification, Clustering, Data Mining, Distance measure, Hierarchical, Linkage criteria, Machine learning

1. 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 [3]. Prevalent ideas of clusters contain groups with diminutive distances between the group objects, data space in dense areas, specific numerical distributions [2]. 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 [20]. The distance function to use, no. of probabilistic clusters or a density threshold are the values used for clustering methods [1]. This research is a stepping stone of optimization as it involves trial and failure concept. Cluster analysis as such is not an Automatic task [5], 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[4]. All the above mentioned approaches ensure its own specialization and traditional methods which are commonly used in the process of data mining real application [8].
1.1. \textit{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 $k \leq n$ 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 \([9][16]\). Objects could also be moved amongst clusters, because the clusters are unit advanced. These strategies usually need variety the amount the quantity of clusters be such that a priori and this number, sometimes doesn't the process. K-Means and Expectation Maximization (EM) and K-means are general algorithms of partitioning method.

1.2. \textit{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 sory 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 \([5][18]\).

1.2.1. \textit{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 an 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).

1.2.2. \textit{Divisive method:}

It’s termed as DIANA (Divide 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 process will continue till objects are in their singleton \([17]\).
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 discrete 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 [15]. 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.

1.4. 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 [14-15][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.

1.5. 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.

2. 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 [11]. 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 [12]. 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 conditions are satisfied \[14\][17]. The up-bottom divisive strategy does the reverse by starting with all of the 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 \[15\] remarked that divisive methods have been largely ignored in the literature mostly due to computational limitations \[16\].

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 \[18\]. 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 \[17\]. They measure the distance between two clusters, \(C_x\) and \(C_y\) have |\(C_x|\) and |\(C_y|\) objects respectively, \(\text{where dist}(O_i,O_j)\) is the distance between two objects \(O_j\) and \(\text{dist}(\mu_{C_x},\mu_{C_y})\) is the distance between the mean values of objects belonging to cluster \(C_x\) and cluster \(C_y\) \[19\].

The single-linkage takes the shortest pair wise distance between objects in two different clusters by using the minimum distance \[20\]. 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 pair wise distances between all pair of objects coming from each of the two clusters \[18\]. 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 method often generates 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 \[4\] 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 \[21\]. 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 \[22\]. Hierarchical clustering algorithms are agglomerative and divisive clustering \[13\]. Table 3 summaries the main differences between these clustering algorithms with regards to the data type they support and the computational cost, and in addition, it includes the shape of clusters they handle as well as the input and output of the algorithms.

3. Hierarchical clustering for numeric items

Many partitioning algorithms require the number of class’s \(k\), as a user-specified parameter; however, \(k\) is not always available in many applications. Hierarchical clustering does not need this priori information. This method creates a sequence of nested partitions. Our method can be divided into two main procedures, these are categorized by two.

3.1. Agglomerative: here initially objects are singleton clusters. First combined two nearest object as a one cluster, iteratively continue the above procedure up to one cluster \[8\], this approach is called as top down approach.

3.2. Divisive: here all data objects have one cluster. Next we have to divide in to clusters by using distance metrics. Finally every object as singleton cluster this approach is called as bottom up approach.
3.3. The HAC (Hierarchical Agglomerative Clustering) takes numeric data as the input and generates the hierarchical partitions as the output.

Our proposed framework starts with the division of Numeric attribute datasets into categorical and numerical attributes sub datasets.

First we dividing attribute type datasets into categorical and numerical attributes sub dataset. We measure the similarity of categorical attribute sub dataset by utilizing entropy based similarity measure using an agglomerative process. Based on the results of the similarity measure, we analyze the changes in total entropy total entropy value while building clusters in agglomerative way and extracting candidate cluster numbers, K (i.e., a list of desirable cluster numbers), for mixed attribute type dataset clustering.

As a criterion function, similarity measure between objects is one of the primary steps in clustering process. Entropy can be used to measure the uncertainty of random variables. Similarity measure for numerical attribute Distance functions such as Euclidean distance are used as since they represent the inherent distance meaning between numerical attributes but they are not for categorical attribute.

3.4. 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.

3.5. Hierarchical algorithm descriptions

Single Linkage: (Figure 4) represents the nearest-neighbor technique that tends to select the attained distance between the nearest or the annotations in groups are closest as shown in below Equation (1).

\[ d_{c_u, c_v} = \arg \min_{(u, v)} \left( \min_{x \in c_u, y \in c_v} D(x, y) \right) \]  

(1)

3.6. Complete linkage: (Figure 5) represent the farthest-neighbor technique that select the farthest observations in clusters using with Euclidean distance measure that as implemented in the Equation (6.2) which is Objective function of complete linkage:

\[ d_{c_u, c_v} = \arg \min_{(u, v)} \left( \max_{x \in c_u, y \in c_v} D(x, y) \right) \]  

(2)

3.7. Average linkage: (Figure 6) 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 (3.3) which is Objective function of average linkage:

\[ d_{c_u, c_v} = \arg \min_{(u, v)} \left( \frac{1}{|c_u|} \sum_{x \in c_u} \frac{1}{|c_v|} \sum_{y \in c_v} D(x, y) \right) \]  

(3)

3.8. Average weighted linkage: (Figure 7) Average linkage calculates the distances with corresponding weights attained among the entire sub group of various annotations generated in groups with weights and the average is implemented from all of these weighted distances attained are as shown in Equation (3.4).

\[ d_{c_u, c_v} = \sum_{x \in X_1, y \in X_2} \frac{d(x, y) \cdot w(x) \cdot w(y)}{w(X_1) \cdot w(X_2)} \]  

(4)

4. Experimental study

4.1. Comparative study of single linkage and complete Link and Average Linkages

In this paper, compared with Agglomerative clustering methods using with Euclidean distance measure 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 prolonged data from a dataset.

4.2. Description of dataset

In the below investigational study we have taken five different datasets obtained from UCI Machine Learning warehouse which can be easily accessed http://archive.ics.uci.edu/ml/datasets/html.

Experimental study clustering algorithms are 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.

4.3. Exercise: agglomerative clustering

In this Exercise we have taken 8 objects for create clusters: \( c_1 = (2, 10), c_2 = (2, 5), c_3 = (8,4), c_4 = (5, 8), c_5 = (7, 5), c_6 = (6, 4), c_7 = (1, 2), c_8 = (4, 9) \).

| Table 1. Linkage: Single |
|--------------------------|
| \( d \) | \( k \) | CLS |
| 0 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 1 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 2 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 3 | 4 | (c4, c8, c1), (c3, c5, c6), (c2), (c7) |
| 4 | 2 | (c1, c3, c4, c5, c6, e8), (c2, c7) |
| 5 | 1 | (c1, c3, c4, c5, c6, e8, c2, c7) |

| Table 2. Linkage: Complete Link |
|--------------------------|
| \( d \) | \( k \) | CLS |
| 0 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 1 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 2 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 3 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 4 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 5 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 6 | 2 | (c8, c1, c3, c5, c6), (c2, c7) |
| 7 | 2 | (c4, c8, c1, c3, c5, c6), (c2, c7) |
| 8 | 1 | (c4, c8, c1, c3, c5, c6, c2, c7) |

| Table 3. Linkage: Average Link |
|--------------------------|
| \( d \) | \( k \) | CLS |
| 0 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 1 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 2 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 3 | 4 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 4 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 5 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 6 | 1 | (c4, c8, c1, c3, c5, c6, c2, c7) |

Distance of Average from (c3, c5, c6) to (c1, c4, c8) is 5.53 and is 5.75 to (c2, c7)

| Table 4. Linkage: Average Weighted Linkage |
|--------------------------|
| \( d \) | \( k \) | CLS |
| 0 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 1 | 8 | (c1), (c2), (c3), (c4), (c5), (c6), (c7), (c8) |
| 2 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 3 | 5 | (c4, c8), (c1), (c3, c5, c6), (c2), (c7) |
| 4 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 5 | 3 | (c4, c8, c1), (c3, c5, c6), (c2, c7) |
| 6 | 1 | (c4, c8, c1, c3, c5, c6, c2, c7) |

Average Distance from (c4, c8, c1) to (c3, c5, c6) is 5.13 and is 5.25 to (c2, c7)
Table 5. Runtime calculation of linkage criteria.

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

Figure 8. Runtime calculation of single linkage, complete linkage and average linkages.

The above Figure 8 represents the time elapsed for creating clusters in milliseconds 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.

5. Conclusion
In this paper four hierarchical clustering methods have been considered where the items using a top down and bottom-up approach that generates clusters that contains only one item and splits into two most related clusters always based on a resemblance metric provided by the algorithm. The linkage methods were analyzed with some examples. In the above Example Analyzes the performance of the 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. In Future we are working on Agglomerative Clustering Algorithms performance calculations on numeric as well as categorical data.

6. References
[1] O. A. Akeem, T. K. Ogunyinka, and B. L. Abimbola 2012A framework for multime- dia data mining in information technology environment International Journal of Computer Science and Information Security 10(5) 69–77
[2] R. A. Baeza-Yates and B. Ribeiro-Neto 1999. Modern information retrieval. Addison- Wesley Longman Publishing Co., Inc., Boston, MA, USA
[3] G. Cormode and A. McGregor 2008 Approximation algorithms for clustering uncertain data In Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on
Principles of Database Systems, PODS08191–200

[4] R Ravi Kumar, M Babu Reddy and P Praveen 2019 Text Classification Performance Analysis on Machine Learning International Journal of Advanced Science and Technology, 28(20) 691 – 697

[5] N. Bhatia and V. Ashev 2010, Survey of nearest-neighbor techniques International Journal of Computer Science and Information Security 8(2) 302–305

[6] Shaik, M.A. 2019 A Survey on Text Classification methods through Machine Learning Methods International Journal of Control and Automation 12(6) 390–396

[7] R. Ravi Kumar, M. Babu Reddy and P. Praveen 2017 A review of feature subset selection on unsupervised learning Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics Chennai 163-167

[8] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu 1998 An optimal algorithm for approximate nearest neighbor searching in fixed dimensions Journal of the ACM 45(6) 891–923

[9] Rajasri I, Guptha AVSSKS and Rao YVD 2011 Influence of Structural Aspects on the Generation Process in Planetary Gear Trains Engineering 3(10) 1018-1021 DOI: 10.4236/eng.2011.310126

[10] Shaik, M.A, T.Sampath Kumar and P Praveen 2019 Research on Multi-Agent Experiment in Clustering International Journal of Recent Technology and Engineering 8(1S4) 1126-1129

[11] M. Sheshikala, D. Rajeswara Rao and R. Vijaya Prakash 2017 Computation Analysis for Finding Co–Location Patterns using Map–Reduce Framework Indian Journal of Science and Technology 10(8)

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

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

[14] P. Indyk and A. Andoni 2006 Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions Foundations of Computer Science 51(1) 117–122

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

[16] B. Rama, P. Praveen, H. Sinha and T. Choudhury 2017 A study on causal rule discovery with PC algorithm International Conference on Infocom Technologies and Unmanned Systems Trends and Future Directions Dubai 616-621

[17] A. Rodriguez and A. Laio 2014 Clustering by Fast Search and Find of Density Peaks Science 344(6191) 1492-1496

[18] Mahender K, Kumar TA and Ramesh KS 2019 Simple transmit diversity techniques for wireless communications Advances in Intelligent Systems and Computing 669 329-342 10.1007/978-981-10-8968-8_28

[19] Praveen., P and Ch. Jayanth Babu 2017 Big Data Clustering: Applying Conventional Data Mining Techniques in Big Data Environment Innovations in Computer Science and Engineering, Lecture Notes in Networks and Systems 74

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

[21] R Ravi Kumar, M Babu Reddy and P Praveen 2019 An Evaluation Of Feature Selection Algorithms In Machine Learning International Journal Of Scientific & Technology Research 8(12) 2071-2074

[22] Sudarshan, E., Satyanarayana, C. and Bindu, C.S., 2017, September. A Parallel RLE Entropy Coding Technique for DICOM Images on GPGPU. In 2017 International Conference on Current
Trends in Computer, Electrical, Electronics and Communication (CTCEEC) (pp. 963-966). IEEE.