Volume 7, Issue 10, October 2022

ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Analysis of Hierarchical Clustering for Single, Average and Complete Linkage

Narendra Suman
M Tech 4th Sem
Computer Science and Engineering Department
Lakshmi Narain College of Technology and Science
Indore M.P. India

Miss Sanjana Sharma
Assistant Professor
Computer Science and Engineering Department
Lakshmi Narain College of Technology and Science
Indore M.P. India

Abstract: In this paper we enhanced hierarchical clustering algorithms like single, complete and average linkage methods by using the concept of cluster ensemble techniques. Hierarchical clustering methods construct the clusters by recursively partitioning the instances in either a top-down or bottom-up fashion. These methods can be subdivided into Agglomerative hierarchical clustering and Divisive hierarchical clustering. The result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change. A clustering of the data objects is obtained by cutting the dendrogram at the desired similarity level. Single linkage method is based on similarity of two clusters that are most similar (closest) points in the different clusters. Complete linkage method based on similarity of two clusters that are least similar (most distant) points in the different clusters. Average linkage method based on In this paper we proposed Association Relation Coefficient (ARC) and calculate versatility. Using ensemble technique we generate cluster with single, complete and average linkage method and find which one generate best cluster. We use original distance matrix with dendrogram value to calculate Association Relation Coefficient. ARC has the value one then method generate perfect cluster for the given data set.

Keywords: Clustering, Single ,Average ,Complete , Hierarchical

I. Introduction

Outliers are unusual values in dataset, and they can distort statistical analyses and we can not violate their assumptions. Analysts will confront outliers and be forced to make decisions about what to do with them. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. It's necessary to understand how outliers occur and whether they strength happen again as a normal part of the process or study area. Outliers increase the variability in data, which decreases statistical power. Consequently, excluding outliers can cause results to become statistically significant. One cannot recognize outliers while collecting data; we won't know what values are outliers until you begin analyzing the data. Many statistical tests are sensitive to outliers and therefore, the ability to detect them is an important part of data analytics. The interpretability of an outlier model is very important, and decisions seeking to tackle an outlier need some context or rationale. Outliers sometimes can be helpful indicators. For example, in some applications of data analytics like credit card fraud detection, outlier analysis becomes important because here, the exception rather than the rule may be of interest to the analyst.

II. Outlier Detection Methods

There are three most common methods for detection of outliers [4,9,10]

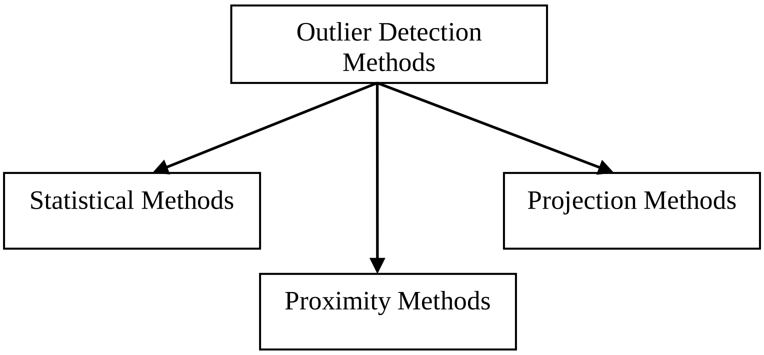

Figure 1 outlier detection methods

1) *Statistical Methods:-* This methods used by starting with visual analysis of the Univariate data by using Box plots, Scatter plots, Whisker plots, etc., can help in finding the extreme values in the data. Assuming a normal distribution, calculate the z-score, which means the standard deviation (σ) times the data point is from the sample's mean. Because we know from the Empirical Another way would be to use Inter Quartile Range (IQR) as a criterion and treating outliers outside the range of 1.5 times from the first or the third quartile.

2) *Proximity-based methods:* Proximity-based methods used clustering techniques to identify the clusters in the data. They assume that an object is an outlier if the nearest neighbors of the object are far away in feature space; that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set. They are classified into two types- Distance-based methods judge a data point based on the distance to its neighbors. Density-based determines the degree of outlines of each data instance based on its local density. DB Scan, k-means, and hierarchical clustering techniques are examples of density based outlier detection methods.

3) *Projection Methods :-* Projection methods utilize techniques such as the PCA to model the data into a lower-dimensional subspace using linear correlations. Post that, the distance of each data point to a plane that fits the sub-space is calculated. This distance can be used then to find the outliers. Projection methods are simple and easy to apply and can highlight irrelevant values. The PCA-based method approaches a problem by analyzing available features to determine what constitutes a "normal" class. The module then applies distance metrics to identify cases that represent anomalies.

III. Impact Of Outliers On Dataset

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set[13,16]

- It increases the error variance and reduces the power of statistical tests
- If the outliers are non-randomly distributed, they can decrease normality
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

To understand the impact deeply, let's take an example to check what happens to a data set with and without outliers in the data set.

Table 1 Data set with outlier

Data set without Outlier	Data set with Outlier
4,4,5,5,5,5,6,6,6,7,7	4,4,5,5,5,5,6,6,6,7,7,300
Mean=5.45	Mean=30.00
Median=5.00	Median=5.50
Mode=5.00	Mode=5.00
Standard Deviation=1.04	Standard Deviation=85.03

IV. Problem Statement

Building a model for finding the data normality is very challenging and may be impossible because it is hard to determine all the behavioral properties of the normal objects. Different applications require different types of data as input and require various modeling and analysis algorithms. Choosing the Outlier detection method depends on the application type. We need to find out the outliers from a vast variety of applications data so the data types of these data sets may vary. There is no unique outlier detection method for all the applications. The noise makes the quality of the data set to be imperfect. Noise often occurs when the data is collected from many resources and applications. Noise in the data sets is caused due to the duplicate tuples, missing values, and deviation of data attributes [4,7,]

V. Literature Survey

In 2011 Akshay Krishnamurthy et al "Efficient Active Algorithms for Hierarchical Clustering". They proposed a general framework for active hierarchical clustering that repeatedly runs an off-the-shelf clustering algorithm on small subsets of the data and comes with guarantees on performance, measurement complexity and runtime complexity. They instantiate framework with a simple spectral clustering algorithm and provide concrete results on its performance, showing that, under some assumptions [5].

In 2012 Dan Wei, Qingshan Jiang et al. proposed "A novel hierarchical clustering algorithm for gene Sequences" .The proposed method is evaluated by clustering functionally related gene sequences and by phylogenetic analysis. They presented a novel approach for DNA sequence clustering, bum, based on a new sequence similarity measure, DMk, which is extracted from DNA sequences based on the position and composition of oligonucleotide pattern.. Proposed method may be extended for protein sequence analysis and meta genomics of identifying source organisms of meta genomic data [6].

In 2013 Yogita Rani and Dr. Harish Rohil “A Study of Hierarchical Clustering Algorithm”. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but these objects are very dissimilar to the objects that are in other clusters. Clustering methods are mainly divided into two groups: hierarchical and partitioning methods. Hierarchical clustering combine data objects into clusters, those clusters into larger clusters, and so forth, creating a hierarchy of clusters. In partitioning clustering methods various partitions are constructed and then evaluations of these partitions are performed by some criterion. They provides a detailed discussion on some improved hierarchical clustering algorithms[7].

In 2014 Varun Chandola et al proposed “Outlier Detection: A Survey”. Outlier detection has been researched within various application domains and knowledge disciplines. This survey provides a comprehensive overview of existing outlier detection techniques by classifying them along different dimensions. Every unique problem formulation entails a different approach, resulting in a huge literature on outlier detection techniques. Several techniques have been proposed to target a particular application domain. The survey can hopefully allow mapping such existing techniques to other application domains. The concept of using a context to detect Type II outliers has not been completely understood. Several techniques unknowingly have adopted a Type II outlier detection approach. Song et al. have shown that using a context improves the outlier detection capability of a technique [8].

In 2015 Shavian P. Patel and Vinita Shah proposed “A Survey of Outlier Detection in Data Mining”. Outlier detection plays an important role in data mining field. Outlier Detection is useful in many fields like Network intrusion detection, Credit card fraud detection, stoke market analysis, detecting outlying in wireless sensor network data, fault diagnosis in machines, etc. They gave a survey on different Outlier detection approaches, which are statistical-based approach, deviation-based approach, distance-based approach, density-based approach. In order to deal with outlier, clustering method is used. For that K-mean is widely used to cluster the dataset then we can apply any technique for finding outliers. They discuss about the concept of outlier. They conclude that k-mean algorithm is most widely used for clustering the dataset. They also describe and compare different approaches of outlier detection which are statistical approach, distance-based approach, density-based approach, deviation-based approach. [9].

In 2016 Kamaljeet Kaurand et al proposed “Comparative Study of Outlier Detection algorithms”. Outlier is considered as the pattern that is different from the rest of the patterns present in the data set. The detection of the outlier in the data set is an important process as it helps in acquiring the useful information that further helps in the data analysis. They covers a study of various outlier detection algorithms like Statistical based outlier detection, Depth based outlier detection, Clustering based technique, Density based outlier detection etc. Comparison study of these outlier detection methods is done to find out which of the outlier detection algorithms are more applicable on high dimensional data. The speed of processing the data is to be increased that helps in the reduction of processing cost of data. They concluded that performance of clustering algorithms is comparatively better than other outlier detection algorithms on huge data sets [10].

In 2017 Zeeshan Ahmad et al proposed “A survey on machine learning and outlier detection techniques”. The machine learning techniques try to understand the different data sets which are given to the machine. The data which comes inside can be divided into two types i.e. labeled data and the unlabeled data. These have to tackle both of the data. Those techniques have been looked upon as well. Then the concept of outlier comes into picture. Outlier Detection is one of the major issues in Data Mining; to find an outlier from a group of patterns is a famous problem in data mining. A pattern which is dissimilar from all the remaining patterns is an outlier in the dataset. Earlier outliers were known as noisy data, now it has become very difficult in different areas of research. A number of surveys, research and review articles cover outlier detection techniques in great details. They discusses and it tries to explain some of the techniques which can help us in identifying or detecting the observation which show such kind of abnormal behavior, and in technical terms called as outlier detection techniques[11].

In 2018 Aurore Archimbaud et al proposed “ ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure”. Many statistical methods are already implemented in R and are briefly surveyed in the present paper. But only a few lead to the accurate identification of potential outliers in the case of a small level of contamination. In this particular context, the Invariant Coordinate Selection (ICS) method shows remarkable properties for identifying outliers that lie on a low-dimensional subspace in its first invariant components. It is implemented in the ICSOutlier package. The main function of the package, ics.outlier, offers the possibility of labeling potential outliers in a completely automated way. Four examples, including two real examples in quality control, illustrate the use of the function. Comparing with several other approaches, it appears that ICS is generally as efficient as its competitors and shows an advantage in the context of a small proportion of outliers lying in a low-dimensional subspace [12].

In 2019 Paulo Jiao et al proposed “Healthcare Outlier Detection with Hierarchical Self-Organizing Map” The amount of data in healthcare repositories along with their high dimensionality nature, requires sophisticated set of analysis capabilities in order to extract new and unexpected patterns, trends and relationships embedded in that data. They proposed the use of Hierarchical Self-Organizing Map (HSOM) algorithm to perform clustering analysis, dimensionality reduction and outlier detection in healthcare data. HSOM provides an appropriate framework to perform the clustering task based on individual types of data and is more powerful and sensitive than standard Self-Organizing Map (SOM) for outlier detection. To solve the problem of the cluster border effect in the SOM, this produces partial clusters not well defined at the edge of the U-mat. Another suggestion for research is the magnification problem that tends to underestimate high probability regions and overestimate low probability areas [13].

In 2021 Najib Ishaq , Thomas J. Howard et al Anomaly and outlier detection in datasets is a long-standing problem in machine learning. In some cases, anomaly detection is easy, such as when data are drawn from well-characterized distributions such as the Gaussian.

However, when data occupy high-dimensional spaces, anomaly detection becomes more difficult. They presented CLAM (Clustered Learning of Approximate Manifolds), a fast hierarchical clustering technique that learns a manifold in a Banach space defined by a distance metric. CLAM induces a graph from the cluster tree, based on overlapping clusters determined by several geometric and topological features. They implemented CHAODA (Clustered Hierarchical Anomaly and Outlier Detection Algorithms), exploring various properties of the graphs and their constituent clusters to compute scores of anomalousness. On 24 publicly available datasets, we compare the performance of CHAODA (by measure of ROC AUC) to a variety of state-of-the-art unsupervised anomaly-detection algorithms. Six of the datasets are used for training. CHAODA outperforms other approaches on 14 of the remaining 18 datasets[14].

VI. ALGORITHM OF PROPOSED APPROACH

Proposed algorithm has following steps

1. Allocate each object as separate cluster and denoted as c_1, c_2, c_3, ..c_n where n is the no. of objects
2. Calculate the distance of each object form other object to create distance matrix D.
3. Check the closest pair of clusters in the current working clustering, and denoted it as pair (x), (y), according to d(x, y) = mind (i, j) { i, is an object in cluster x and j in cluster y}
4. Combine clusters (x) and (s) into a MIN cluster to create a new cluster. Store combine objects with its corresponding distance in Dendrogram distance Matrix.
5. Revise distance matrix, D, by removing the rows and columns corresponding to clusters (x) and (y). Adding a new row and column corresponding to the combine cluster(x, y) and old uster (k) is defined in this way: d[(z), (x, y)] = min d[(z),(x)], d[(z),(y)].For other rows and columns copy the corresponding data from existing distance matrix.
6. If all objects are in one cluster, stop or else, go to step 3.

VII. IMPLEMENTATION ENVIRONMENT

This chapter, we evaluate the performance of proposed algorithm and compare it with MIN linkage, MAX linkage and average linkage methods. The experiments were performed on Intel Core i5-4200U processor2GB main memory and RAM: 4GB Inbuilt HDD: 500GB OS: Windows 8. The algorithms are implemented in using R language. Synthetic datasets are used to evaluate the performance of the algorithms

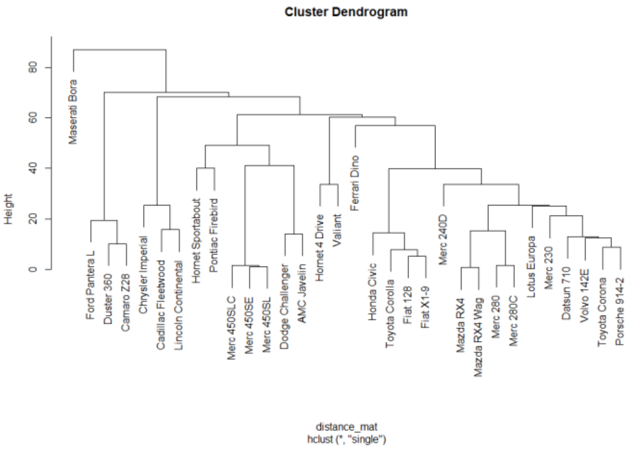

Figure 2 Hierarchical clustering using single linkage

VII. COMPARING BASED ON NUMBER OF OBJECTS IN CLUSTERS

Table 2 number of object in each cluster

Cluster No	No of Objects
1	25
2	3
3	3
4	1

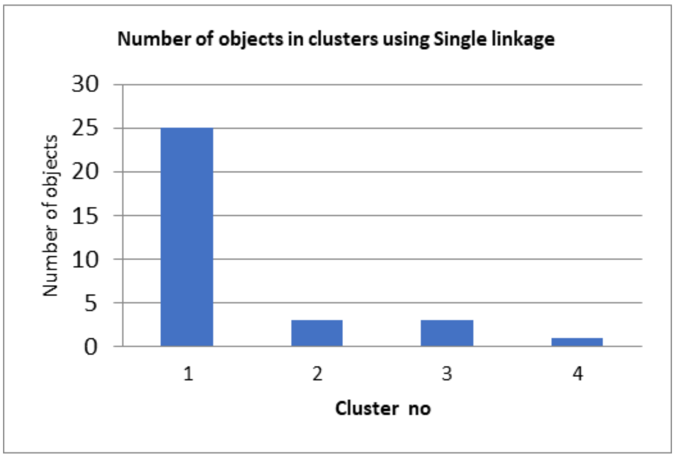

Figure 4 Number of objects in clusters using Single linkage approach

VIII CONCLUSION

There are several algorithms and methods have been developed for clustering problem. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency The most popular agglomerative clustering procedures are Single linkage ,Complete linkage , Average linkage and Centroid .

REFERENCE

1. Revati Raman et al. Fuzzy Clustering Technique for Numerical and Categorical dataset International Journal on Computer Science and Engineering (IJCSE) NCICT 2010 Special Issue.
2. K. Ranjini Performance Analysis of Hierarchical Clustering Algorithm Performance Analysis of Hierarchical Clustering Algorithm” Int. J. Advanced Networking and Applications Volume: 03, Issue: 01, Pages: 1006-1011 (2011).
3. Hussain Abu-Dalbouhet al proposed “Bidirectional Agglomerative Hierarchical Clustering using AVL Tree Algorithm”. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 1, September 2011 ISSN (Online): 1694-0814.
4. Piyush Raiet proposed “Data Clustering: K-means and Hierarchical Clustering” CS5350/6350: Machine Learning October 4, 2011
5. Akshay Krishnamurthy et al “Efficient Active Algorithms for Hierarchical Clustering” Appearing in Proceedings of the 28 the International Conference on Machine Learning, Bellevue, WA, USA, 2011.
6. Dan Wei, Qingshan Jiang et al. A novel hierarchical clustering algorithm for gene Sequences” 2012 Wei et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License.
7. Yogita Rani¹ and Dr. Harish Rohil2 “A Study of Hierarchical Clustering Algorithm” International Journal of Information and Computation Technology. ISSN 0974-2239 Volume 3, Number 11 (2013), pp. 1225-1232 © International Research Publications House
8. Varun Chandola University of Minnesota Arindam Banerjee University of Minnesota and VIPIN KUMAR Outlier Detection : A Survey ACM Computing Surveys · January 2009.
9. Shivani P. Patel Vinita Shah Jay Vala A Survey of Outlier Detection in Data Mining National Conference on Recent Research in Engineering and Technology (NCRRET -2015) International Journal of Advance Engineering and Research Development (IJAERD) e-ISSN: 2348 - 4470 , print-ISSN:2348-6406.
10. Kamaljeet Kaur Atul Gar Comparative Study of Outlier Detection Algorithms International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016.
11. Zeeshan Ahmad Lodhia1† and Akhtar Rasool2††, and Gaurav Hajela3 A survey on machine learning and outlier detection techniques IJCSNS International Journal of Computer Science and Network Security, VOL.17 No.5, May 2017.
12. Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure The R Journal Vol. 10/1, July 2018 ISSN 2073-4859.
13. Paulo João Octavian Postolache Healthcare Outlier Detection with Hierarchical Self-Organizing Map Instituto de Telecomunicações, ISCTE-IUL Lisbon, Portugal e-mail: paulo.joao@lx.it.pt August 2019.
14. Najib Ishaq and Thomas J. Howard “CLUSTERED HIERARCHICAL ANOMALY AND OUTLIER DETECTION ALGORITHMS” arXiv:2103.11774v1 [cs.LG] 9 Feb 2021
15. C. Leela Krishna , C. Kala Krishna Outlier Detection Using Association Rule Mining and Cluster Analysis International Journal of Computer Sciences and Engineering Vol.-6, Issue-6, Jun 2018 E-ISSN: 2347-2693.