Volume 9, Issue 09, September 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

IDENTIFYING OUTLIERS USING DIVISIVE HIERARCHICAL CLUSTERING MIN, AVERAGE AND MAX LINKAGE FUNCTION

Nimesha Mandal
M Tech CSE 4th semester
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Mr. Kamlesh Patidar
Assistant Professor CSE department
Jawaharlal Institute of Technology Borawan
Khargone (M.P.) 451228

Abstract: ***Outliers are unusual values in dataset, and they can distort statistical analyses and violate their assumptions. Analysts will confront outliers and be forced to make decisions about what to do with them. Given the problems they can cause, might think that it's best to remove them from data. Removing outliers is legitimate only for specific reasons. Outliers can be very informative about the subject-area and data collection process. It's essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. Unfortunately, resisting the temptation to remove outliers inappropriately can be difficult. 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 we 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. We have used R language to implement the Hierarchical clustering approach. We found that the global cluster is generate by all three basic approaches are same but the number of object in clusters are in each are different. By the experimental analysis we found that all three methods has different approach but by the Dendrogram we easy find the global***

Keywords: Outlier, Single, Average, Complete, Clusters, Global

I. INTRODUCTION

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together. For example, consider a dataset of vehicles is given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner. Now our task is to convert the unlabelled data to labelled data and it can be done using clusters. The main idea of cluster analysis is that it would arrange all the data points by forming clusters like cars cluster which contains all the cars, bikes clusters which contains all the bikes, etc. Simply it is partitioning of similar objects which are applied on unlabelled data. The clustering methods can be classified into the following categories:[9,10]

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

II. PROPERTIES OF CLUSTERING

1. Clustering Scalability: Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable if it is not scalable, then we can't get the appropriate result and would lead to wrong results.[8,9,11]

2. High Dimensionality: The algorithm should be able to handle high dimensional space along with the data of small size.

3. Algorithm Usability with multiple data kinds: Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

5. **Dealing with unstructured data:** These would be some databases that contain missing values, noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data give it some structure to the data by organizing it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. Interpretability: The outcomes of clustering should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

HIERARCHICAL METHOD

In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:[11,12,13]

Hierarchical clustering method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called **Dendrogram** (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or cluster are break up (top-down view).

The basic methods to generate hierarchical clustering are:

A. Agglomerative:

Initially consider every data point as an **individual** Cluster and at every step, **merge** the nearest pairs of the cluster. (It is a bottom-up method). At first every data set is considered as individual entity or cluster. At every iteration, the clusters merge with different clusters until one cluster is formed. Algorithm for Agglomerative Hierarchical Clustering is:

- Calculate the similarity of one cluster with all the other clusters (calculate proximity matrix)
- Consider every data point as a individual cluster
- Merge the clusters which are highly similar or close to each other.
- Recalculate the proximity matrix for each cluster
- Repeat Step 3 and 4 until only a single cluster remains.

We have six data points A, B, C, D, E, F.

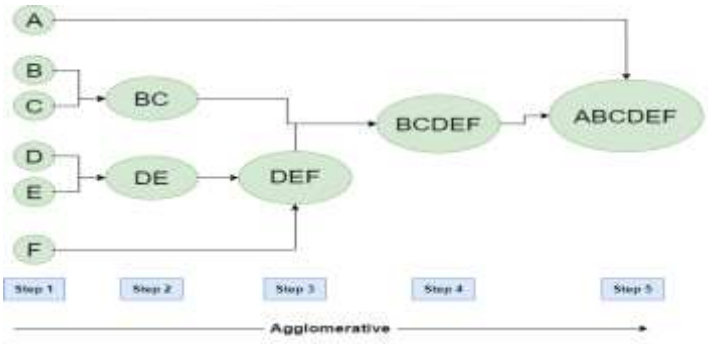

Fig 1 – Agglomerative Hierarchical clustering

B. Divisive:

We can say that the Divisive Hierarchical clustering is precisely the **opposite** of the Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, we take into account all of the data points as a single cluster and in every iteration, we separate the data points from the clusters which aren't comparable. In the end, we are left with N clusters.

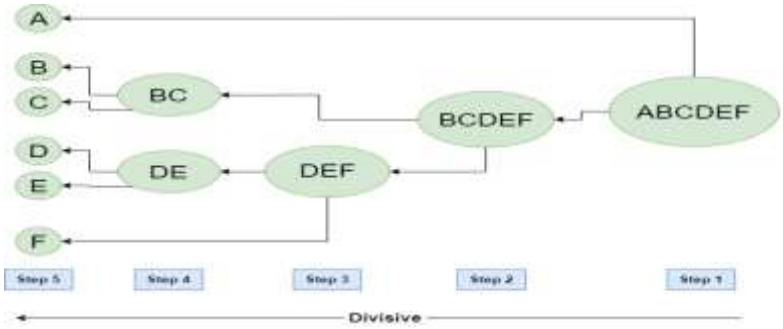

Fig2 – Divisive Hierarchical clustering

THE LINKAGE FUNCTION

The linkage function tells us to measure the distance between clusters. Again, there are many choices. Typically we consider either a new item that summarizes the items in the cluster, or a new distance that summarizes the distance between the items in the cluster and items in other clusters. Here is a list of four methods. In each example, x is in one cluster and y is in the other.

TABLE 1

Linkage function

Name of linkage function	Form of linkage function
Single	f=min(d(x,y))
Complete	f=max(d(x,y))
Average	f=average(d(x,y))

III. LITERATURE SURVEY

In 2017 Dipannita Kar, Mr. Haresh Chande and Mr. Rajendra Gaikwad proposed "A Study Paper on Outlier Detection on Time Series Data". Time series data are observations collected sequentially over time. Consider as example weather prediction. When observations are collected at frequent time intervals, large data sets in the form of time series are generated. Outlier detection is a primary step in many data mining. An outlier is a piece of data or observation that deviates from other observations, outliers is not noise They gave comparative study has been performed on the k-means, Density based, EM and Cobweb algorithm. Comparison is performed on AWS data using WEKA tool. Comparative results are shown in the form of table. The comparative study is performed on the basis of time. The best algorithm is Density Based algorithm. It takes less time than other algorithms [2].

In 2017 Zeeshan Ahmad Lodhiaand Akhtar Rasool proposed "A survey on machine learning and outlier detection techniques" computer science having different types of techniques such as supervised learning, unsupervised learning, reinforcement learning and the various techniques which are lying under them, so in order to understand these different machine learning techniques a survey on these machine learning techniques has been done and tried to explain these few techniques. Finding an outlier is useful in detecting the data which can't be predicted and that which can't be identified. 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[3].

In 2018 Aurore Archimbaud, Klaus Nordhausen, and Anne Ruiz-Gazen 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.. It is implemented in the ICSOutlierpackage. 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 [4].

In 2018 C. Leela Krishna and C. Kala Krishna proposed "Outlier Detection Using Association Rule Mining and Cluster Analysis". An object whose behavior is found to be different from others in a dataset is said to be an outlier. They proposed two different approaches for outlier detection.. They considered outlier detection problem in two varieties of databases, one involving data streams, where data arrives continuously and also in a time based order, and the other involving static databases. They provided two algorithms for the problem. For streamed data, a sliding window is considered to make the data items in a database bounded. Then for all the transactions in the bounded data set, a prefix-tree is taken and items are added to it, and it serves as a stack. Association rules are derived from the transactions data set and the items which do not obey the association rules are declared as outliers [5].

In 2018 Remi Dominguesa and Maurizio Filippone proposed "A comparative evaluation of outlier detection algorithms: experiments and analyses". They survey unsupervised machine learning algorithms in the context of outlier detection. This task challenges state-of-the-art methods from a variety of research fields to applications including fraud detection, intrusion detection, medical diagnoses and data cleaning. Each method is then submitted to extensive scalability, memory consumption and robustness tests in order to build a full overview of the algorithms' characteristics. In the context of outlier detection, we benchmarked the average precision, robustness, and computation time and memory usage of 14 algorithms on synthetic and real datasets. They suggested that this algorithm is more suitable than ride in production environment as the latter is much more computationally expensive and memory consuming [6].

In 2019 Paulo Jiao Octavian Postulate proposed "Healthcare Outlier Detection with Hierarchical Self-Organizing Map" Historically, healthcare entities in both public and private sectors have relied upon business rules and outlier models to mine claims data for fraud, waste and abuse patterns. 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. Further research is necessary to fully explore the potential of the HSOM. This includes the standardization of vocabulary and data share across organizations, to enhance the benefits of this kind of applications. [7].

In 2019 Tung Kiel, Bin Yang and Chinua Guo proposed "Outlier Detection for Time Series with Recurrent Auto encoder Ensembles". They propose two solutions to outlier detection intimae series based on recurrent auto encoder ensembles. This ensemble-based approach aims to reduce the effects of some auto encoders being overfitted to outliers, this way improving overall detection quality. Experiments

with two real-world time series data sets, including univariate and multivariate time series, offer insight into the design properties of the proposed ensemble frameworks and demonstrate that the proposed frameworks are capable of outperforming both baselines and the state-of-the-art methods. They proposed two auto encoder ensemble frameworks based on sparsely-connected recurrent neural networks for unsupervised outlier detection in time series. One of the framework strains multiple auto encoders independently while the other framework trains multiple auto encoders jointly in a multi-task learning fashion [8]

In 2020 Harry Braga, S. Praia and K. Aditya proposed "Outlier Detection Based on Machine Learning Techniques". Outliers are being researched in many fields of research and various domains. They analyses and bring together various outlier detection techniques.. They come out with perfect results to find anomaly using various methods in order to decrease the fraud pricing of housing by several agents. So, this technique has made the work easier to find the price whether it's suitable for that society or not and moreover decreased the human interaction and manual calculations. Exception location plans to discover designs in information that don't adjust to anticipated conduct. It will have a broad use in a wide assortment of uses, for example, military reconnaissance for foe exercises, interruption location in digital security, extortion discovery for charge cards, protection or medicinal services and flaw recognition in wellbeing basic frameworks. Their significance in information is because of the way that they can convert into noteworthy data in a wide assortment of utilizations. A strange traffic design in a PC system could imply that a hacked PC is conveying touchy information to an unapproved goal [10].

IV. PROPOSED APPROACH

1. Assign each object as individual cluster like c_1, c_2, c_3, ..c_n where n is the no. of objects
2. Find the distance matrix D, using any similarity measure
3. Find the closest pair of clusters in the current clustering, say pair (r), (s), according to d(r, s) = mind (i, j) { i, is an object in cluster r and j in cluster s}
4. Merge clusters (r) and (s) into a MIN cluster to form a merged cluster. Store merged objects with its corresponding distance in Dendrogram distance Matrix.
5. Update distance matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s). Adding a new row and column corresponding to the merged cluster(r, s) and old uster (k) is defined in this way: d[(k), (r, s)] = min d[(k),(r)], d[(k),(s)].For other rows and columns copy the corresponding data from existing distance matrix.
6. If all objects are in one cluster, stop. Otherwise, go to step 3.
7. Find association relation coefficient value with Single, Average and Complete linkage methods.

V. EXPERIMENTAL ANALYSIS

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.
Choosing no. of clusters form Dendrogram with Single linkage

abline(h = 74, col = "green")

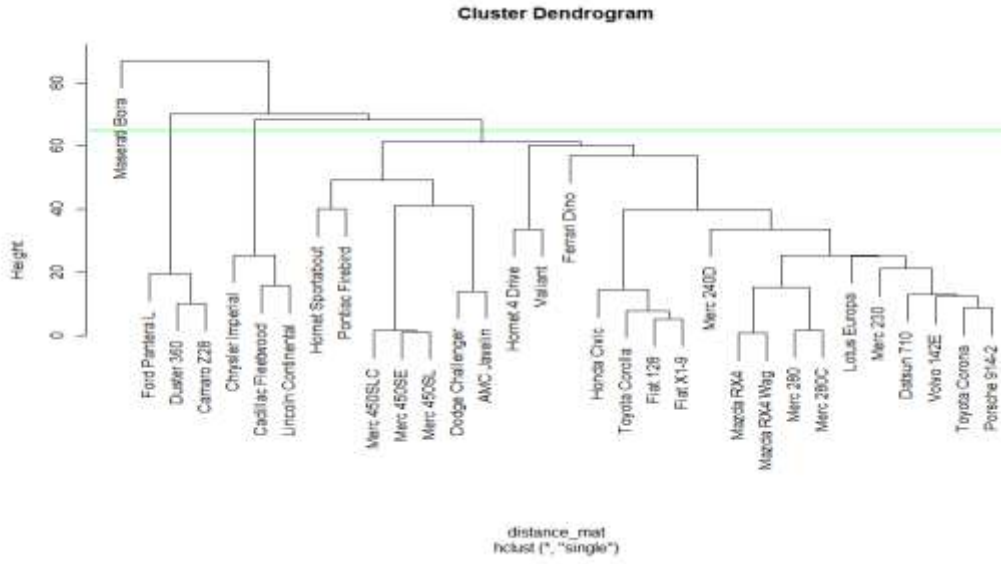

Fig3: Choosing no. of clusters form Dendrogram with Single linkage

CONCLUSION

In this paper we analysis the Hierarchical clustering approach with three basic techniques Single linkage, complete linkage, Average linkage. Our objective is to find whether the global outlier present in the data set or not. We use real life data mtcars and implemented all three basic approaches. We also found number of cluster for each approach. We used abline methods do decide correct number of cluster. We used R language to implement the Hierarchical clustering approach. We found that the global cluster is generate by all three basic approach are same but the number of object in clusters are in each are different.

REFERENCE

1. Kamaljeet Kaur Atul Gar Comparative Study of Outlier Detection Algorithms International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016
2. Dipannita Kar, Mr. Haresh Chande, Mr. Rajendra Gaikwad A Study Paper on Outlier Detection on Time Series Data International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org Volume 5, Issue 4 December 2017 | ISSN: 2320-2882
3. 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
4. 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
5. 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
6. Remi Dominguesa , Maurizio Filipponea , Pietro Michiardia , Jihane Zouaouib A comparative evaluation of outlier detection algorithms: experiments and analyses Preprint submitted to Elsevier August 20, 2018
7. 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
8. Tung Kieu , Bin Yang∗ , Chenjuan Guo and Christian S. Jensen Outlier Detection for Time Series with Recurrent Autoencoder Ensembles Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19.
9. Stefan Mandic-Rajcevic and Claudio Colosio Methods for the Identification of Outliers and Their Influence on Exposure Assessment in Agricultural Pesticide Applicators: A Proposed Approach and Validation Using Biological Monitoring Department of Health Sciences, University of Milan and International Centre for Rural Health of the Saints Paolo and Carlo Hospital, 20142 Milan, Italy
10. Harry Bhagat, S.Priya, K. Aditya Outlier Detection Based on Machine Learning Techniques IJAST Vol. 29 No. 06 (2020): Vol. 29 No. 06 (2020)
11. Kamal Malik H.Sadawarti , Member IEEE, 3Kalra G.S., Member IEEE Comparative Analysis of Outlier Detection Techniques International Journal of Computer Applications (0975 – 8887) Volume 97– No.8, July 2014.
12. Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad A Comparative Study for Outlier Detection Techniques in Data Mining Faculty of Science and Technology University College of Science and Technology 21030 Kuala Terengganu, Malaysia {zuriana, rosmayati}@kustem.edu.my, 1-4244-0023-6/06/$20.00 ©2006 IEEE
13. 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