Volume 9, Issue 03, March 2024 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Hierarchical Clustering Better Approach for Handling Outliers

Krati Sanduke
M Tech (CSE) 4th semester
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus Indore M.P.
kratisanduke05@gmail.com

Mr Nilesh Avinash Joshi
Assistant Professor
Computer Science and Engineering Department
L.N.C.T. (Bhopal) Indore Campus
Indore M.P.

Abstract: ***Outlier detection is a key consideration within the development and deployment of any model and also data analysis. Identifying and dealing with outliers is an integral part of working with data, and data analysis. In this data-rich environment, organisations can expect to have to deal with outlier data. Outliers can skew trends and have a serious impact on the accuracy data analysis. The presence of outliers can be a sign of concept drift, so ongoing outlier analysis is needed. Outliers can skew results, and anomalies in training data can impact overall model effectiveness. Outlier detection is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analysed once identified. Outlier detection is an important part of each stage of the machine learning process. Accurate data is integral during the development and training of algorithms, and outlier detection is performed after deployment to maintain the effectiveness of models. This guide explores the basics of outlier detection techniques in machine learning, and how they can be applied to identify different types of outlier. In this paper we proposed various outlier detection methods and also there advantages and disadvantages***

Keywords: component; formatting; style; styling; insert (Minimum 4 to 7 key words must be phrases)

I. INTRODUCTION

The presence of outliers can be a sign of concept drift, so ongoing outlier analysis in machine learning is needed. Machine learning models learn from data to understand the trends and relationship between data points. Outliers can skew results, and anomalies in training data can impact overall model effectiveness. Outlier detection is a key tool in safeguarding data quality, as anomalous data and errors can be removed and analysed once identified. Outlier detection is an important part of each stage of the machine learning process. Accurate data is integral during the development and training of algorithms, and outlier detection is performed after deployment to maintain the effectiveness of models. This guide explores the basics of outlier detection techniques in machine learning, and how they can be applied to identify different types of outlier. An outlier is an individual point of data that is distant from other points in the dataset. It is an anomaly in the dataset that may be caused by a range of errors in capturing, processing or manipulating data. Outliers can skew overall data trends, so outlier detection methods are an important part of statistics. Outliers will be a consideration for any area that uses data to make decisions. If an organisation is gaining insight from data, outliers are a real risk. Outlier detection is particularly important within machine learning. Models are trained on huge arrays of training data. The model understands the relationship between data points to help predict future events or categorise live data. Outliers in the training data may skew the model, lowering its accuracy and overall effectiveness. Outlier analysis and resolution can lengthen the training time too. Outliers can be present in any data or machine learning use case, whether that's financial modelling or business performance analysis[9,10].

II. OUTLIER DETECTION METHODS

There are various methods of outlier detection is as follows −

Supervised Methods − Supervised methods model data normality and abnormality. Domain professionals tests and label a sample of the basic data. Outlier detection can be modeled as a classification issue. The service is to understand a classifier that can identify outliers. The sample can be used for training and testing. In various applications, the professionals can label only the normal objects, and several objects not connecting the model of normal objects are documented as outliers. There are different methods model the outliers and consider objects not connecting the model of outliers as normal.

Unsupervised Methods − In various application methods, objects labeled as "normal" or "outlier" are not applicable. Therefore, an unsupervised learning approach has to be used. Unsupervised outlier detection methods create an implicit assumption such as the normal objects are considerably "clustered." An unsupervised outlier detection method predict that normal objects follow a pattern far more generally than outliers. Normal objects do not have to decline into one team sharing large similarity. Instead, they can form several groups, where each group has multiple features. This assumption cannot be true sometime. The normal objects do not send some strong patterns. Rather than, they are uniformly distributed. The collective outliers, share large similarity in a small area.

Unsupervised methods cannot identify such outliers efficiently. In some applications, normal objects are separately distributed, and several objects do not follow strong patterns. For example, in some intrusion detection and computer virus detection issues, normal activities are distinct and some do not decline into high-quality clusters.

Some clustering methods can be adapted to facilitate as unsupervised outlier detection methods. The main idea is to discover clusters first, and therefore the data objects not belonging to some cluster are identified as outliers. However, such methods deteriorate from two issues. First, a data object not belonging to some cluster can be noise rather than an outlier. Second, it is expensive to discover clusters first and then discover outliers. [7,8].

III. Effect 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. Literature Survey

In 2012 Karanjit proposed ,Outliers once upon a time regarded as noisy data in statistics, has turned out to be an important problem which is being researched in diverse fields of research and application domains. Many outlier detection techniques have been developed specific to certain application domains, while some techniques are more generic. Some application domains are being researched in strict confidentiality such as research on crime and terrorist activities. They attempt to bring together various outlier detection techniques, in a structured and generic description. With this exercise, we hope to attain a better understanding of the different directions of research on outlier analysis for ourselves as well as for beginners in this research field who could then pick up the links to different areas of applications in details.[1]

A.

In 2013 Ji Zhang proposed, Outlier detection is an important research problem in data mining that aims to discover useful abnormal and irregular patterns hidden in large datasets. They proposed comprehensive survey is presented to review the existing methods for detecting point outliers from various kinds of vector-like datasets. The outlier detection techniques that are primarily suitable for relatively low-dimensional static data, which serve the technical foundation for many of the methods proposed later, are reviewed first. We have also reviewed some of recent advancements in outlier detection for dealing with more complex high-dimensional static data and data streams.
It is important to be aware of the limitation of this survey. As it has clearly stated in Section 2, we only focus on the point outlier detection methods from vector-like datasets due to the space limit. Also, outlier detection is a fast-developing field of research and more new methods will quickly emerge in the foreseeable near future [2].

In 2014 Manish Gupta, Jing Gao proposed "Outlier Detection for Temporal Data: A Survey". They presented an organized overview of the various techniques proposed for outlier detection on temporal data. Modeling temporal data is a challenging task due to the dynamic nature and complex evolutionary patterns in the data. In the past, there are a wide variety of models developed to capture different facets in temporal data outlier detection. This survey organized the discussion along different data types, presented various outlier definitions, and briefly introduced the corresponding techniques. This survey provides a number of insights and lessons as follows. The methods for different data types are not easy to generalize to one another, though some of them may have similarity in the framework at the broader level. For example, change detection in continuous time series and discrete time series both require forecasting methods. While the number of formulations of the temporal outlier detection problem are diverse, they are generally motivated by the most common applications which are encountered in the literature. Many recent applications, especially those corresponding to novel data types in the context[3].

In 2015 Christy.Aa ,proposed the concept of data preprocessing for outlier reduction. They propose two algorithms namely Distance-Based outlier detection and Cluster-Based outlier algorithm for detecting and removing outliers using a outlier score. By cleaning the dataset and clustering based on similarity, can remove outliers on the key attribute subset rather than on the full dimensional attributes of

dataset. Experiments were conducted using 3 built-in Health care dataset available in R package and the results show that the cluster-based outlier detection algorithm providing better accuracy than distance-based outlier detection algorithm. The goal of the algorithms presented is to improve the quality of data processing and capture the underlying patterns in the data by reducing the effect of outliers at the pre-processing stage. This outlier may be due to the unavailability or distortions in the data collection stage that consists of irrelevant or weakly relevant data objects. From the algorithms, it is shown that by choosing a valid outlier score, the overall performance of the algorithm can be improved[4].

In 2016 Atul Garg proposed the dimension of the data is increasing day by day, outlier detection is emerging as one of the active areas of research. Finding of the outliers from large data sets is the main problem. Various algorithms have been proposed till date for the detection of the outliers. 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. There is no single universally applicable outlier detection approach of the current techniques. They presented the study of different existing outlier detection techniques and the way in which they are categorized. It is concluded that performance of clustering algorithms is comparatively better than other outlier detection algorithms on huge data sets. There is need of developing some new algorithms or improvement in the existing one is required [5].

In 2018 Simon Kojo Appiah, Doris Arthur Proposed study examined the performance of six outlier detection techniques using a non-stationary time series dataset. Two key issues were of interest. Scenario one was the method that could correctly detect the number of outliers introduced into the dataset whiles scenario two was to find the technique that would over detect the number of outliers introduced into the dataset, when a dataset contains only extreme maxima values, extreme minima values or both. Air passenger dataset was used with different outliers or extreme values ranging from 1 to 10 and 40. The six outlier detection techniques used in this study were Mahalanobis distance, depth-based, robust kernel-based outlier factor (RKOF), generalized dispersion, Kth nearest neighbors distance (KNND), and principal component (PC) methods. Mahalanobis method could identify more outliers than the others, making it the "best" method for the extreme minima category. The kth nearest neighbor distance method was the "best" method for not over-detecting the number of outliers for extreme minima[8].

In 2019 Lee Jonathan proposed Identifying outliers can lead to better datasets by (1) removing noise in datasets and (2) guiding collection of additional data to fill gaps. However, the problem of detecting both outlier types has received relatively little attention in NLP, particularly for dialog systems. They introduce a simple and effective technique for detecting both erroneous and unique samples in a corpus of short texts using neural sentence embeddings combined with distance-based outlier detection. They introduce the first neural outlier detection method for short text and demonstrates its effectiveness across multiple metrics in multiple experiments. We also propose a way to integrate outlier detection into data collection, developing and evaluating a novel crowdsourcing pipeline. This pipeline supports the creation of higher quality datasets to yield higher quality models by both reducing the number of errors and increasing the diversity of collected data. While the experiments discussed herein are concerned with components of dialog systems[9].

V. DIVISIVE HIERARCHICAL CLUSTERING

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.

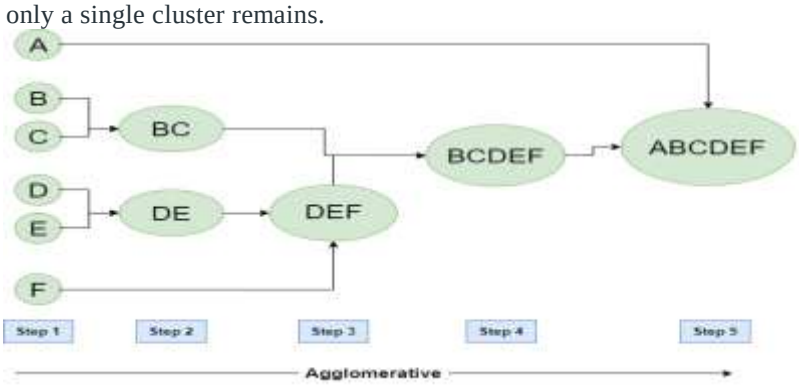

Fig. 1 Agglomerative Hierarchical clustering

VI. EXPAEIMETNAL 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.

Number of objects in clusters using Average linkage approach

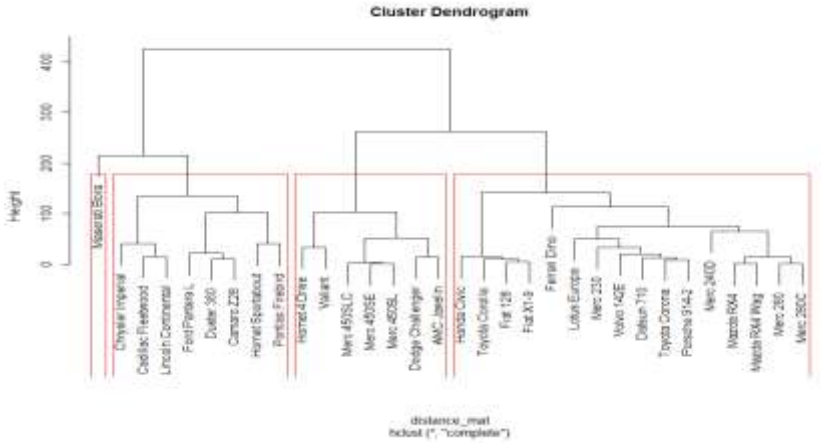

Fig.2 Agglomerative Hierarchical clustering with Average linkage approach

TABLE 2
Number of objects in clusters using Average linkage approach

Cluster No	No of Objects
1	16
2	7
3	8
4	1

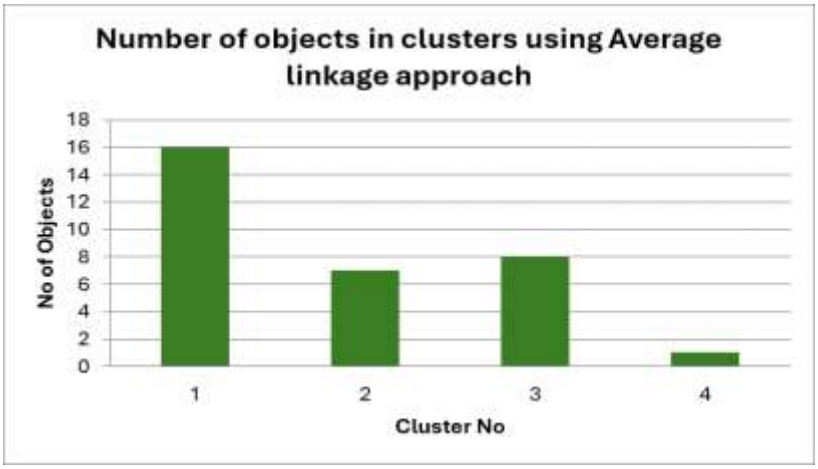

Fig 3 Number of objects in clusters using Single linkage approach

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. Karanjit Singh and Dr. Shuchita Upadhyaya Outlier Detection: Applications and Techniques IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012 ISSN (Online): 1694-0814 www.IJCSI.org
2. Ji Zhang Department of Mathematics and Computing University of Southern Queensland, Australia Advancements of Outlier Detection: A Survey ICST Transactions on Scalable Information Systems Research Article.
3. Manish Gupta, Jing Gao, Member, IEEE, Charu C. Aggarwal, Fellow, IEEE, and Jiawei Han, Fellow, IEEE Outlier Detection for Temporal Data: A Survey IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 1, JANUARY 2014.
4. Christy.Aa , MeeraGandhi.Gb , S. Vaithyasubramanianc Cluster Based Outlier Detection Algorithm For Healthcare Data © 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer-review under responsibility of scientific committee of 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15) doi: 10.1016/j.procs.2015.04.058.
5. Atul Garg Comparative Study of Outlier Detection Algorithm International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016.

6. R. Muthukrishnan and G. Poonkuzhal A Comprehensive Survey on Outlier Detection Method American-Eurasian Journal of Scientific Research 12 (3): 161-171, 2017 ISSN 1818-6785 © IDOSI Publications, 2017 DOI: 10.5829/idosi.aejsr.2017.161.171
7. Sampson Twumasi-Ankrah, Simon Kojo Appiah Comparison of Outlier Detection Techniques In Non-Stationary Time Series Data Global Journal of Pure And Applied Sciences Vol. 27, 2021: 55-60 Copyright© Bachudo Science Co. Ltd Printed In Nigeria ISSN 1118-0579.
8. Lee Hongzhi Wang , Mohamed Jaward Bah , And Mohamed Hammad Progress in Outlier Detection Techniques: A Survey Received July 14, 2019, accepted July 29, 2019, date of publication August 2, 2019, date of current version August 19, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2932769.
9. Stefan Larson Anish Mahendran Ann Arbor, MI, USA Outlier Detection for Improved Data Quality and Diversity in Dialog System Proceedings of NAACL-HLT 2019, pages 517–527 Minneapolis, Minnesota, June 2 - June 7, 2019. c 2019 Association for Computational Linguistics .
10. Hongzhi Wang , Mohamed Jaward Bah , And Mohamed Hammad Progress in Outlier Detection Techniques: A Survey Received July 14, 2019, accepted July 29, 2019, date of publication August 2, 2019, date of current version August 19, 2019. Digital Object Identifier 10.1109/ACCESS.2019.2932769.
11. Clement Franklin D C Comparing the Performance of Anomaly Detection Algorithms International Journal of Engineering Research & Technology (IJERT) http://www.ijert.org ISSN: 2278-0181 IJERTV9IS070532 (This work is licensed under a Creative Commons Attribution 4.0 International License.) Published by : www.ijert.org Vol. 9 Issue 07, July-2020.