International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 7, July 2025

Clustering Based Efficient Approach for Detecting Multivariate Outliers

Ghamandi Yadnyesh Rajendra[1],Virendra Verma[2]
M. Tech 4th Semester LNCT Bhopal Indore Campus Indore MP India [1]
Asst Professor CSE department LNCT Bhopal Indore Campus Indore MP India [2]
yadneshghamandi7@gmail.com[1], virendraverma.v@gmail.com[2]

Abstract: *Outlier detection plays a critical role in various domains such as fraud detection, quality control, and data preprocessing for machine learning tasks. While distance and density-based approaches are widely used for this purpose, hierarchical clustering offers an intuitive alternative for identifying outliers based on the natural groupings in data. This paper presents a method for outlier detection using hierarchical clustering. We explore how dendrogram structures reveal sparsely connected points, propose threshold-based pruning, and evaluate the approach on synthetic and real-world datasets. The results demonstrate that hierarchical clustering can effectively detect global and local outliers while providing an interpretable structure for data analysts.*

Keywords:*Outlier detection, hierarchical clustering, dendrogram, anomaly detection, unsupervised learning.*

1. Introduction

Outliers, in the context of information evaluation, are information points that deviate significantly from the observations in a dataset. These anomalies can show up as surprisingly high or low values, disrupting the distribution of data. For instance, in a data set of monthly sales figures, if the income for one month are extensively higher than the sales for all of the different months, that high sales determine would be considered an outlier. An outlier is an observation that lies outside the overall pattern of a distribution (Moore and McCabe 1999). Usually, the presence of an outlier indicates some sort of problem. This can be a case which does not fit the model under study or an error in measurement. An outlier may also be explained as a piece of data or observation that deviates drastically from the given norm or average of the data set. An outlier may be caused simply by chance, but it may also indicate measurement error or that the given data set has a heavy-tailed distribution [1].

Consider a simple example to determine the outliers of the data set. Also, evaluate the mean of the data set including the outliers and excluding the outliers.
35, 75, 20, 25, 15, 30, 30, 15, 45, 40, 110
First, arrange the data set in order.
15, 15, 20, 25, 30, 30, 35, 40, 45, 75, 110
Find the mean of the data:
Mean = {Sum of the data values}/{Number of data values}
= [15 + 15 + 20 + 25 + 30 + 30 + 35 + 40 + 45 + 75 + 110]/ 11
= 40
Now to find the mean without the outlier,
Evaluating the mean of the data set excluding the outliers, remove the values far off the middle (i.e. 75 and 110):
Mean = Sum of the data values/Number of data values
= {15 + 15 + 20 + 25 + 30 + 30 + 35 + 40 + 45}/9
=20.45
The mean of the given data set is 40 when outliers are included, however, it is 20.45 when outliers are not included.

2. Impact of Outliers on a 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[2,3]:

- 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.

Example:

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

3. Most common causes of outliers on a data set

Outliers, data points significantly different from other values in a dataset, commonly arise from human errors, measurement inaccuracies, experimental flaws, intentional manipulation, or natural variations within the data[4,5].

1. Human or Data Entry Errors:

- Typographical errors: Mistakes while inputting data into a system.
- Incorrect recording: Mistakes during manual data collection.
- Incorrect assumptions: Misunderstanding the data or its context, leading to errors in recording or interpretation.

2. Measurement Errors:

- Faulty instruments: Malfunctioning or improperly calibrated equipment can produce inaccurate readings.
- Improper use of instruments: Incorrect setup or usage of measurement tools.
- Environmental factors: Unforeseen conditions (e.g., temperature, humidity) affecting measurements.

3. Experimental Errors:

- Errors in experimental design: Flaws in how the experiment is set up or carried out.
- Errors in data extraction: Mistakes in extracting data from experiments.
- Errors in data processing: Errors during the cleaning, transformation, or analysis of data.

4. Intentional Outliers:

- Dummy data: Artificial outliers added to test outlier detection methods.
- Fraudulent data: Outliers deliberately introduced to mislead.

5. Sampling Errors:

- Unrepresentative samples: When the sample data doesn't accurately reflect the population.
- Combining data from different sources: Inconsistent data from various sources or locations.

6. Natural Variation:

- Genuine outliers: Some outliers represent actual, extreme values within a natural distribution.
- Rare events: Outliers can occur due to infrequent or unusual occurrences in the data.

4. literature survey

In 2015 Shavian P. Patel et al proposed "A Survey of Outlier Detection in Data Mining". 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 also describe and compare different approaches of outlier detection which are statistical approach, distance-based approach, density-based approach, deviation-based approach. Most of researchers use distance-based approach and density-based approach to detect the outlier [4].

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. They cover a study of various outlier detection algorithms like Statistical based

outlier detection, Depth based outlier detection, Clustering based technique, Density based outlier detection etc. 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 [5].

In 2017 Dipannita Kar et al 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.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 [6].

In 2018 Aurore Archimbaud, Klaus et al proposed " ICSOutlier: Unsupervised Outlier Detection for Low-Dimensional Contamination Structure". But only a few lead to the accurate identification of potential outliers in the case of a small level of contamination. It is implemented in the ICSOutlierpackage. 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 [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. Such networks make it possible to generate multiple auto encoders with different neural network connection structures.. They proposed two auto encoder ensemble frameworks based on sparsely connected recurrent neural networks for unsupervised outlier detection in time series. [9]

In 2020 Bhadri Naarayanan et al proposed Comparing the Performance of Anomaly Detection AlgorithmsAn Anomaly is a data point which differs in characteristics from other data points in the dataset. The detection of anomaly plays an important role in machine learning. But most of the algorithms provide anomaly detection only with limited generalization capacity. In this paper, In this paper, we compare the efficiency of anomaly detection methods which has better robustness. The three outlier detection algorithms used are Local Outlier Factor, Isolation Forest and Autoencoders. Based on the accuracy, recall, precision, F1 score of the algorithms, the comparison graph is constructed for the three datasets and the efficient algorithm is determined[10]

In 2021 Jintao Song et al proposed Outlier Detection with Multivariable Panel Data using K-Means Clustering. They arranged the original spatiotemporal monitoring data into the multivariable panel data format. They the proposed model and applied to the Jinping-I Arch Dam in China which is the highest dam in the world, and results indicate that the detection method has high accuracy detection ability, which is valuable in dam safety monitoring application. multivariable panel data theory and K-means clustering algorithm are combined to construct an outlier detection model for dam deformation monitoring data[20].

In 2022 B.Angelin et al proposedAn Outlier Detection Using Clustering Algorithms. They remove data which necessary for carrying out to speed up the applications like classification, data perturbation and compression They provides a brief survey on clustering techniques and outlier detection techniques. Particularly the K-means clustering algorithm for outlier detection is discussed. They conduct various experiments, by using limited size datasets are used. Because of the explosive growth of available information, a series of experiments and investigations are necessary to establish the potential utility of the proposed methods in real time datasets [21].

In 2023 Sanae Borrohou proposed Data cleaning survey and challenges–improving outlier detection. They provide critical facets of data cleaning and the utilization of outlier detection algorithms. They introduced an innovative algorithm cantered on the fusion of Isolation Forest and clustering techniques. By leveraging the strengths of both methods, this proposed algorithm aims to enhance outlier detection outcomes. They endeavors to elucidate the multifaceted importance of data cleaning, underscored by its symbiotic relationship with ML models. [22].

In 2024 Qi Li et al proposed Detecting outliers by clustering algorithms. Clustering and outlier detection are two important tasks in data mining. Outliers frequently interfere with clustering algorithms to determine the similarity between objects, resulting in unreliable clustering results. Currently, only a few clustering algorithms (e.g., DBSCAN) have the ability to detect outliers to eliminate interference. By experiments they showed that ODAR is robust to diverse datasets. Compared with baseline methods, the clustering algorithms achieve the best on 7 out of 10 datasets with the

help of ODAR, with at least 5% improvement in accuracy. The proposed ODAR, method detect outliers for clustering algorithms [23].

5. Problem Statement

there are number of problems that need to be consider due to outliers some of them are

1. 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.
2. 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.
3. Noise in the data sets is caused due to the duplicate tuples, missing values, and deviation of data attributes.
4. Many machine learning algorithms are sensitive to outliers, which can disproportionately influence model training.
5. In regression, outliers can distort the regression line, leading to poor predictions.
6. In clustering, outliers can shift cluster centers, affecting the accuracy of grouping.

6. Objectives

Numerically and graphically, we need to identify the point as an outlier. We need to examine the data for this point to see if there are any problems with the data. If there is an error, we should fix the error if possible or delete the data. If the data is correct, we would leave it in the data set.

1. Need to determine if there is an outlier preent in the data set or not.
2. We use divisive clustering methods to find out global outliers
3. We apply all three (single, average and complete) approaches of divisive clustering to check that the outlier presents in the data set or not.
4. We used real life data set for detection of outlier's implementation.
5. If there is an outlier, delete it and fit the remaining data to a new line

7. Proposed Approach

1. Assign each object as individual cluster like c1, c2, c3, ..cn 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.

8. Implementation details

We evaluate the performance of proposed algorithms 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

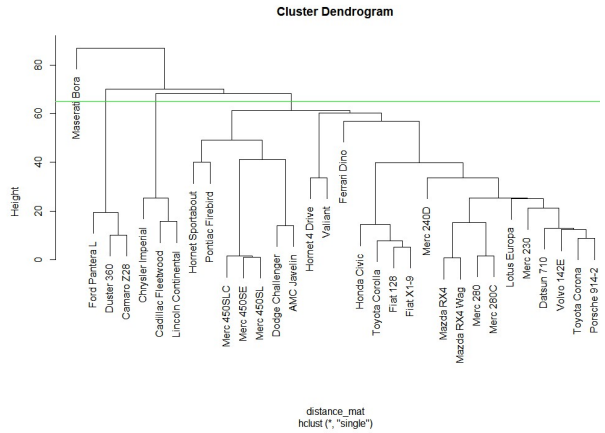

Figure 1 Number of objects in clusters using Single linkage approach

Table 2 Number of objects in clusters using Single linkage approach

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

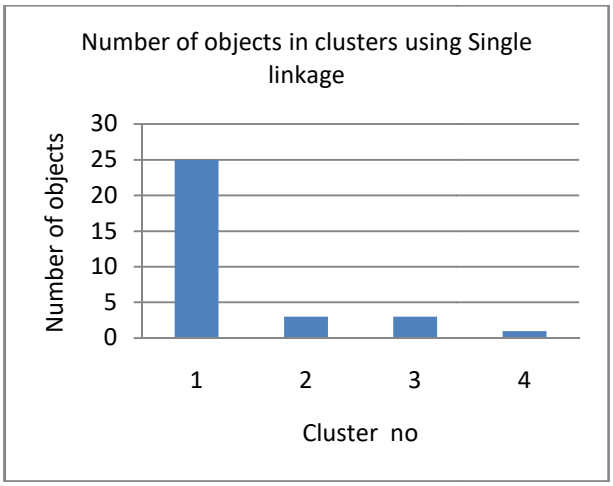

Figure 2 Number of objects in clusters using Single linkage approach

9. Conclusion

There are several algorithms and methods that have been developed for clustering problems. But problems always arise 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. In the proposed work 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 presents in the data set or not. We use real life data mtcars and implemented all three basic approaches. We also found a number of cluster for each approach. We used abline methods do decide the correct number of clusters. We 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.

References

[1] J. Han, M. Kamber, Data mining, Concepts and techniques, Academic Press, 2003.
[2] Arun K. Pujari, Data mining Techniques, University Press (India) Private Limited, 2006.
[3] D. Hand, H. Mannila, P. Smyth, Principles of Data Mining, Prentice Hall of India, 2004
[4] 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
[5] Kamaljeet Kaur Atul Gar Comparative Study of Outlier Detection Algorithms International Journal of Computer Applications (0975 – 8887) Volume 147 – No. 9, August 2016
[6] 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
[7]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
[8] 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
[9] 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.
[10]Bhadri Naarayanan Comparing the Performance of Anomaly Detection Algorithms International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 http://www.ijert.org IJERTV9IS070532 www.ijert.org Vol. 9 Issue 07, July-2020.
[11] Jintao Song ,1,2 Shengfei Zhang Outlier Detection Based on Multivariable Panel Data and K-Means Clustering for Dam Deformation Monitoring Data Advances in Civil Engineering Volume 2021, Article ID 3739551, 11 pages https://doi.org/10.1155/2021/3739551
[12] B.Angelin An Outlier Detection Using Clustering Algorithms and Its Techniques International Journal of Scientific Research & Engineering Trends Volume 8, Issue 2, Mar-Apr-2022, ISSN (Online): 2395-566X
[13] Sanae Borrohou *, Rachida Fissoune and Hassan Badir Data cleaning survey and challenges –improving outlier detection algorithm inmachine learning Journal of Smart Cities and Society 2 (2023) 125–140 125DOI 10.3233/SCS-230008
[14]Qi Li, Shuliang WangDetecting outliers by clustering algorithmsarXiv:2412.05669v1 [cs.LG] 7 Dec 2024