STUDY OF OUTLIERS IN CLUSTER ANALYSIS

Jana Cibulková – Phuong Ngoc Vu

Abstract

Outliers are pervasive in data and can significantly influence the outcomes of statistical analyses. This paper addresses the impact of outliers on hierarchical cluster analysis by introducing different types of outliers across various data types. Through the analysis of simulated data, we examine the effects of outliers on cluster analysis outcomes. The simulated data allow us to control both the type and number of outliers present. Three hierarchical cluster analysis methods (single-linkage, complete-linkage, and average-linkage) and one non-hierarchical method (the k-means algorithm) are applied to the datasets with varying outlier compositions. The clustering solutions obtained are then evaluated using evaluation criteria such as silhouette index, and adjusted Rand index. The results of the simulated study offer further insights into the impact of different types of outliers on the effectiveness of cluster analysis methods. This research contributes to a deeper understanding of outlier behaviour in cluster analysis and informs the development of robust clustering algorithms capable of handling outlier-rich datasets.

Key words: outliers, cluster analysis, hierarchical clustering, evaluation criteria.

JEL Code: C38, C63, C88.

Introduction

We live in an era of big data. Every day, we encounter vast amounts of information stored for further processing and analysis. These datasets are often so extensive that extracting valuable insights from them can be challenging. Clustering becomes useful in such situations. By grouping similar objects into clusters, we obtain representatives of individual categories that capture the typical properties and characteristics of each group. This enables us to quickly and easily identify objects of interest without having to sift through the entire dataset.

However, when analyzing real-world data, it is not always possible to create high-quality clusters. One of the factors complicating this process is the presence of outliers. These objects differ significantly from the rest in their properties and can greatly influence the results of the analysis, leading to distorted interpretations and erroneous conclusions. To better understand the impact of outliers on the clustering process, we conduct simulations with varying occurrences and types of outliers. We also focus on their detection and assess the impact of removing outliers on the quality of clustering.

The motivation for writing this paper stems from the insufficient exploration of this issue in the scientific literature. To our knowledge, only one study has previously addressed a similar topic, namely the article published by Nowak-Brzezińska and Gaibei (2022). However, their work focuses solely on the impact of outliers on cluster quality, not on the accuracy of group classification. As we will discover, the quality of clusters does not necessarily imply correct classification. Moreover, the authors of that study worked exclusively with quantitative data and used only outlier detection methods suitable for this data type, whereas our work also introduces methods for identifying outliers in qualitative data, an area that is also underexplored.

1 Theoretical foundation

We assume a dataset is composed of a collection of observations and their associated variables. In this work, we consider all variables to be either quantitative or qualitative. We represent the dataset with a data matrix ($\boldsymbol{X} \in R^{n \times p}$), where n is the number of observations and p is the number of variables.

1.1 Clustering and its evaluation

We focus on well-known methods of hierarchical clustering: *single-linkage, complete-linkage, average-linkage,* and *Ward's method*. Additionally, we will use one non-hierarchical method: the *k-means algorithm*. A more detailed description of these methods can be found in the book of Gan, Ma, and Wu (2021).

The quality of the identified clusters will be evaluated using both *internal* and *external evaluation criteria*. External criteria compare the clustering result with the true known groups of observations. Internal criteria do not require knowledge of the actual clusters and are based on the properties of the resulting clusters, in our case, measuring cohesion and separation.

- The *silhouette score* ranges from -1 to 1, with higher values indicating better clustering quality. The silhouette score is calculated as the average silhouette index across all observations in the dataset. The silhouette index, which is an internal evaluation criterion, was first introduced by Rousseeuw (1987).

- The *Rand index* is an external criterion that measures the similarity between two clustering solutions. Rand (1971) defined it as the percentage of correctly classified observations into clusters. The main drawback of the Rand index is its sensitivity to randomness. Therefore, in this work, we use the Adjusted Rand Index (ARI), which normalizes the index value to a range from -1 to 1. The greater the similarity between two clusterings, the higher the index value. If two clustering solutions are identical, the index reaches a value of 1. For random clustering, the expected index value is 0.

1.2 Outliers and their detection methods

Barnett and Lewis (1978) define an outlier as an observation (or a subset of observations) that appears to be inconsistent with the remainder of the dataset. Chandola, Banerjee, and Kumar (2009) distinguish three types of outliers:

1. *Point outliers*: These are specific values that differ from the rest of the entire dataset. This type is the simplest and most commonly studied. An example might be a single unusually high transaction in a person's account. Point outliers can be further divided into local outliers (those that are outliers with respect to a single variable) and global outliers (those that are outliers with respect to all variables).
2. *Contextual (or conditional) outliers*: These outliers are characterized by their deviation depending on a specific context. An observation is identified as an outlier only if certain conditions are met; otherwise, it is considered normal. A typical example is a low temperature recorded during the summer, which would be considered normal in winter.
3. *Collective anomalies*: These are groups of observations that, as a whole, deviate from the norm, even though individual objects within the group may not be considered outliers on their own. An example might be an ECG recording where a short-term decrease in heart rate may be normal, but regular repetition could indicate issues such as cardiac arrhythmia.

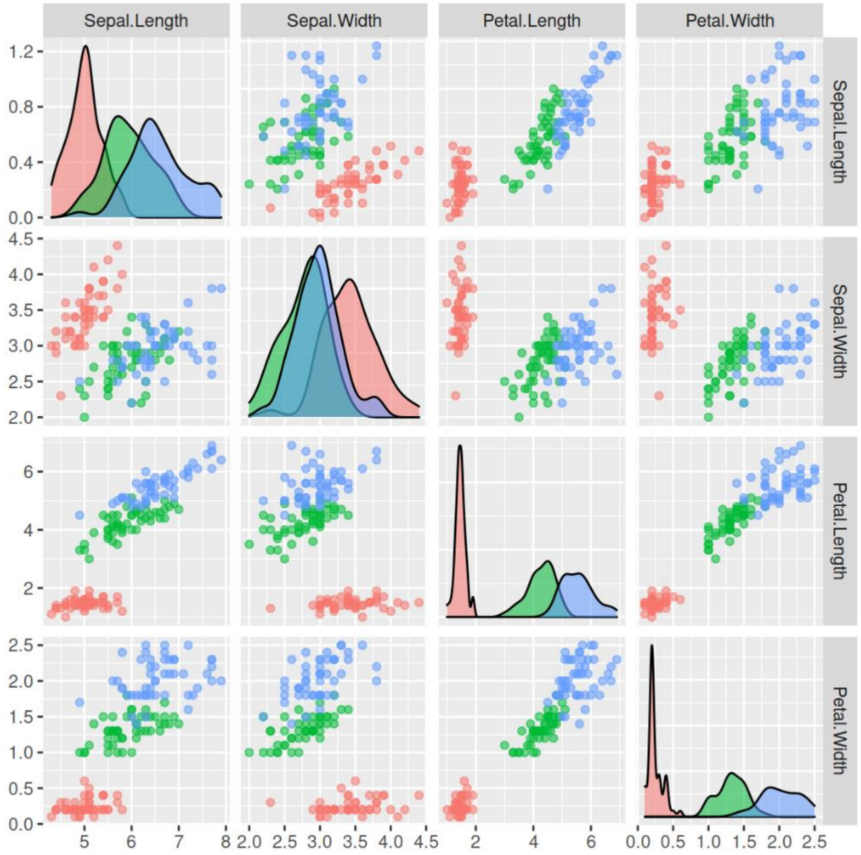
There is no universal method for identifying outliers that can be applied in all cases. Different methods are suitable for detecting different types of outliers, as mentioned above, and specific methods are better suited for different data types. In this work, we analyze the following methods for detecting outliers in numerical variables: Local Outlier Factor (LOF), Isolation Forest (iForest), k-means clustering. For detecting outliers in categorical data, we analyze: Couple Biased Random Walk (CBRW), Frequent Pattern Outlier Factor (FPOF), k-means clustering.

- The *Local Outlier Factor* (LOF) method, proposed by Breunig et al. (2000), utilizes the concept of density in a given space. The method aims to identify outliers in the data by comparing the local density of a point to that of its neighbors. The main idea behind LOF is that an outlier will have a significantly lower local density than its neighbors, while a point within a cluster will have a similar local density to the surrounding points. LOF assigns each observation a score that reflects its outlier status relative to the local density of neighboring points.
- In contrast to traditional approaches, the *Isolation Forest* (iForest) method focuses on isolating abnormal points. Liu et al. (2008) designed this method based on the premise that outliers are "few" and "different," making them easier to isolate from typical observations. The iForest method is based on tree structures known as isolation trees (a specialized type of binary tree).
- The *Couple Biased Random Walk* (CBRW) method focuses on both the relationships between pairs of observations within individual variables (intra-feature value coupling) and the interactions between different variables (inter-feature value coupling). CBRW determines an outlier score for all variable values, which is then used to calculate an outlier score for individual observations or for selecting variables for subsequent outlier detection (feature selection). This method was proposed and further described by Pang et al. (2016).
- The *Frequent Pattern Outlier Factor* (FPOF) method is based on the theory of knowledge discovery in databases and uses the frequency of itemsets (combinations of categories of a certain length) to identify outliers. The method involves searching for frequent items and patterns, which are commonly occurring subsets of values. Based on these insights, as described by He et al. (2015), outliers are identified as those containing patterns that occur infrequently in the data.
- The k-means clustering method can be utilized for outlier detection. This method considers observations with distances from their centroid above a threshold to be outliers. The threshold can be set based on statistical methods or domain knowledge.

2 Experiment methodology

In the simulation study, various types of outliers are introduced into the *Iris* dataset (Fisher, 1936), which contains measurements of sepal and petal dimensions for three species of iris flowers (Iris setosa, Iris virginica, and Iris versicolor), as shown in Figure 1. The dataset was chosen for its cleanliness—there are no missing or erroneous values, and the number of clusters is already known. The dataset contains 150 observations, with each of the three iris species represented by 50 records. Python programming language is used for all data processing and data analysis presented in this paper.

Fig. 1: Dataset Iris



Source: Authors

2.1 Quantitative data

We propose the following algorithm to quantify the impact of outliers in quantitative data.

For each clustering method described in the first chapter:

1. We apply the clustering method to the original Iris dataset to detect three clusters.
2. We calculate and record the evaluation metrics[1] (silhouette index, adjusted Rand index).
3. We introduce outliers into the dataset. In the simulation study, we select three different proportions of artificially created outliers:

[1] To calculate the adjusted Rand index, where knowledge of the true group is required, we will use only the clustering results for the original 150 observations.

   a. 1% (1 case),
   b. 5% (7 cases), and
   c. 10% (15 cases).

   Each scenario is repeated 50 times for each of the four types of outliers: point local outliers, point global outliers, contextual outliers, and collective outliers. Vu (2024) describes the process of outlier generation in detail.

4. We perform clustering on the datasets with added outliers and record the evaluation metrics.
5. For each outlier detection method for numerical variables (LOF, iForest, k-means):
   a. We apply the method to the datasets with added outliers.
   b. We evaluate the performance of the outlier detection methods using accuracy and precision, which can be easily derived from the confusion matrix.
   c. We remove the identified outliers from the datasets and then apply the clustering method to the cleaned dataset to detect three clusters.
   d. We calculate and record the evaluation metrics exactly like we did in step 2.

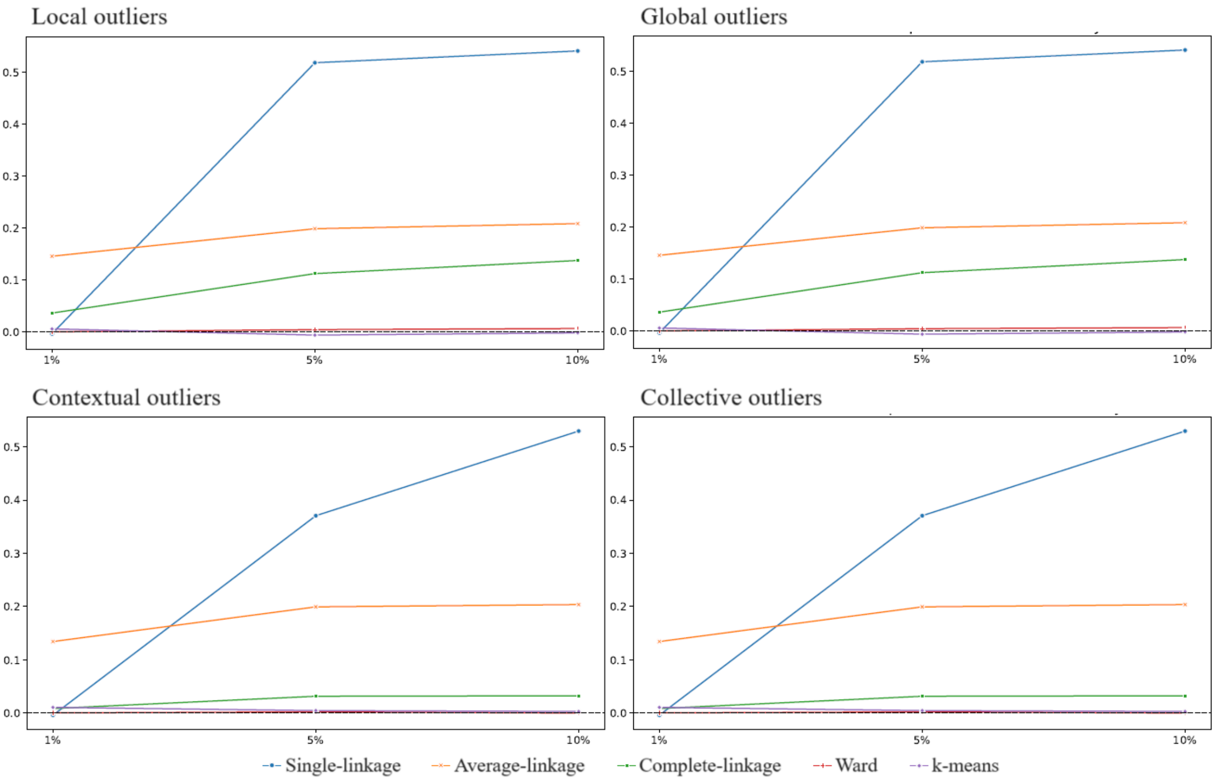
2.2 Qualitative data

To quantify the impact of outliers in categorical data, we will proceed similarly. Given the limited availability of clean reference datasets containing categorical variables and even fewer with a known number of clusters, and considering the non-trivial task of adding synthetic outliers to categorical data, we will discretize the Iris dataset to obtain categorical variables. This process will be applied to both the original dataset and all modified datasets with artificially added outliers from step 3. In the discretization process, we choose an approach where each interval has the same length.

3 Experiment outcomes

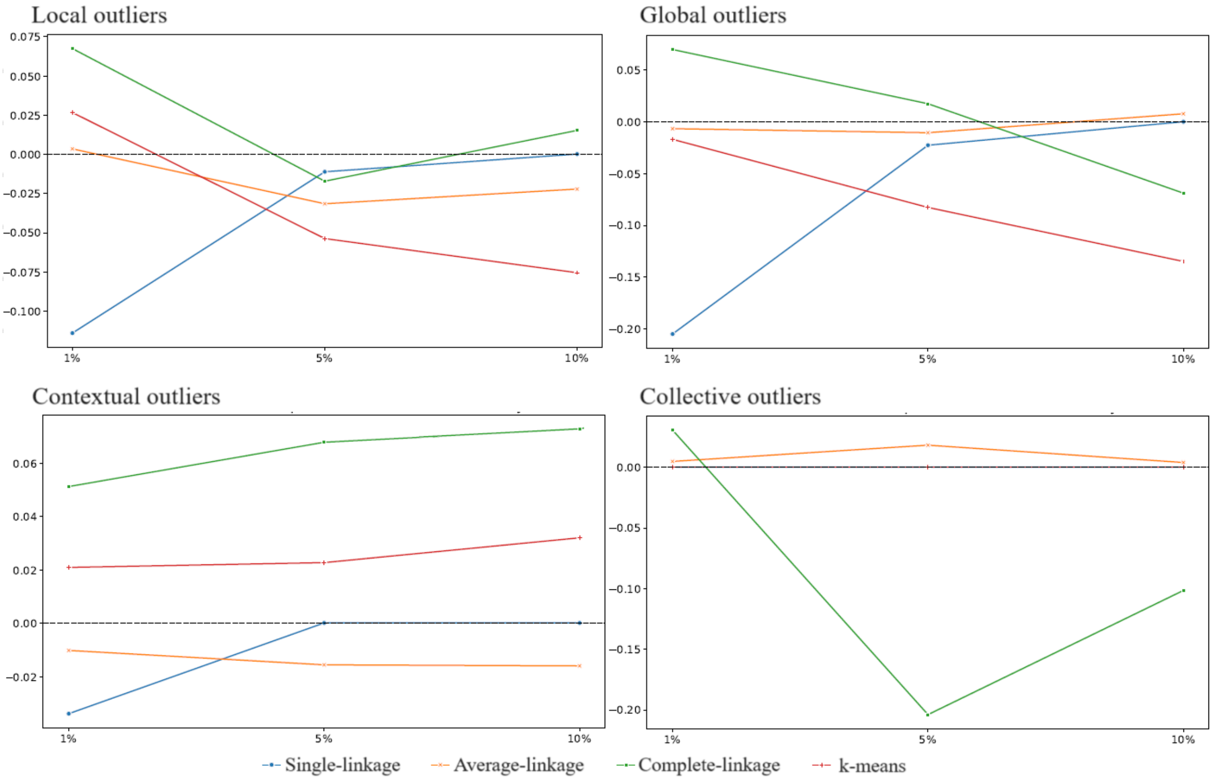
To compare the correctnes of clustering solutions before and after the addition of outliers to the dataset, we created a set of four graphs, shown in Figures 2 and 3. These graphs display the differences in the adjusted Rand index values, which were calculated in steps 2 and 5 of algorithm for the experiment. The graphs illustrate the different clustering methods (indicated by different line colors) and the varying frequencies of outlier presence (as shown on the x-axis in percentages). Each graph addresses a different type of outlier. Positive values on the y-axis indicate a worsening in clustering accuracy cause be the introduction of outliers.

Fig. 2: The correctness of clustering solutions comparison; ARI; quantitative data



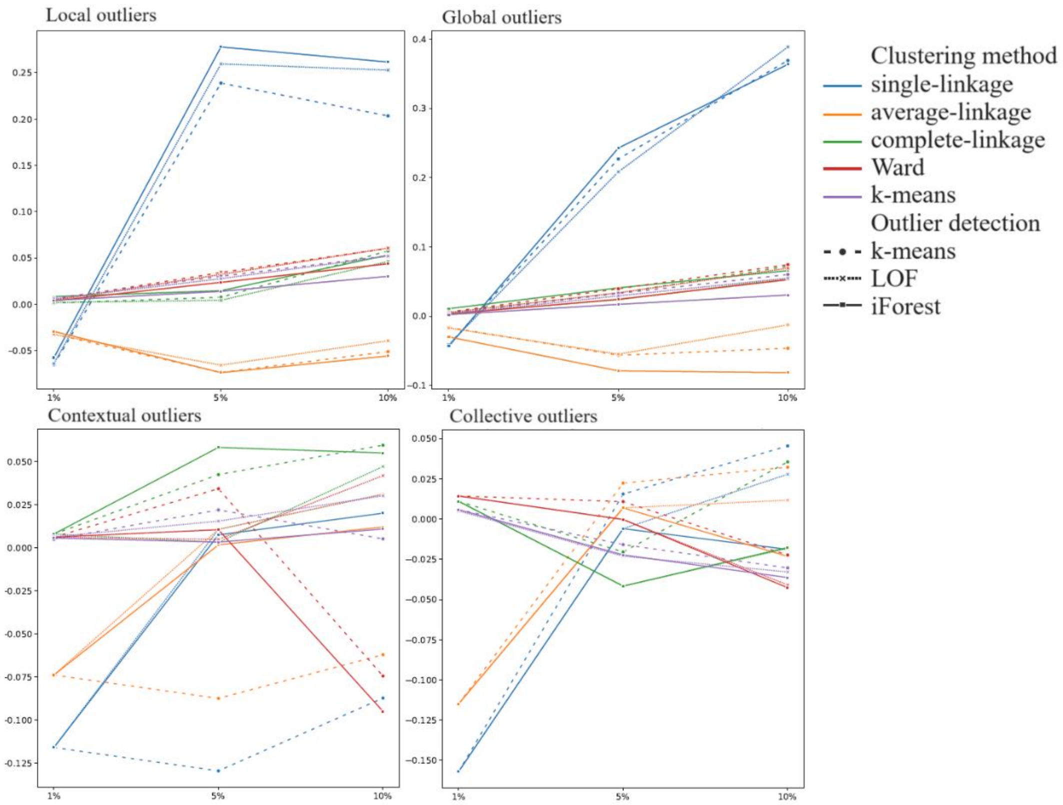
Source: Authors

Fig. 3: The correctness of clustering solutions comparison; ARI; qualitative data



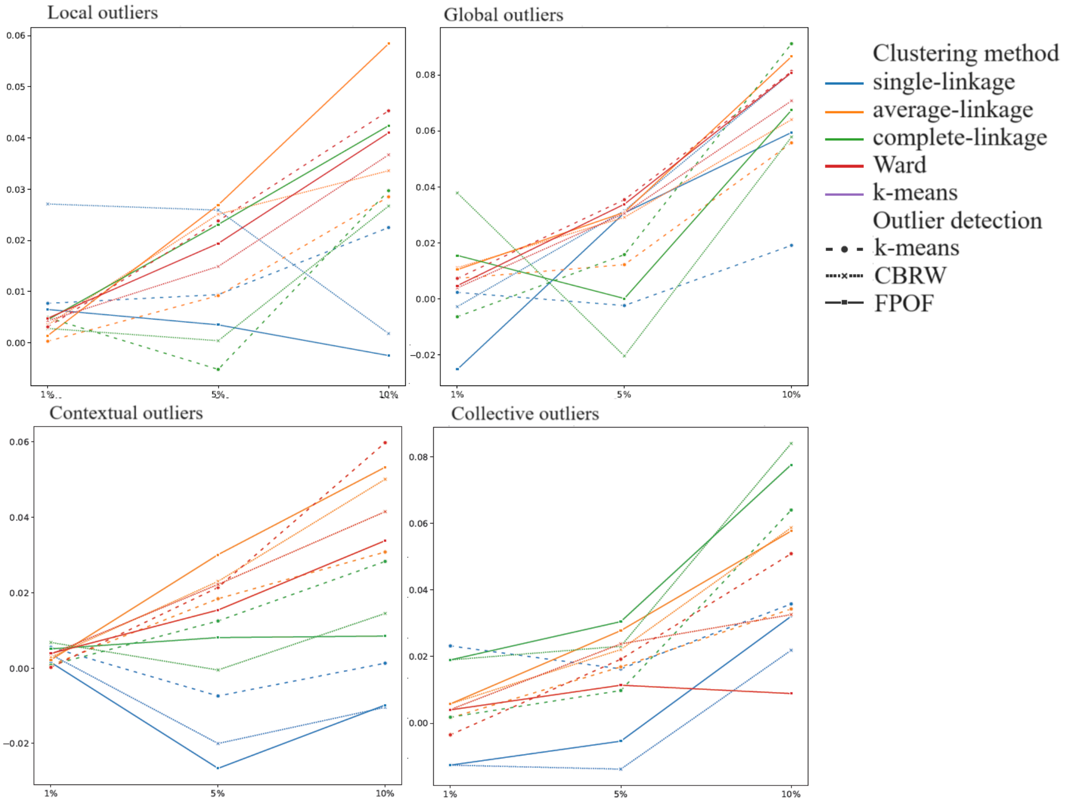
Source: Authors

Fig. 4: The quality of clustering solutions comparison; Silhouette index; quantitative data



Source: Authors

Fig. 5: The quality of clustering solutions comparison; Silhouette index; qualitative data



Source: Authors

The difference in Silhouette scores before and after the removal of outliers from the dataset is shown in a set of four graphs in Figures 4 and 5. Each graph addresses a different type of outlier. The graphs display the various clustering methods (different line colors) and the different frequencies of outliers (as shown on the x-axis). Additionally, the results for all outlier detection methods (line type) are visualized. Positive values on the y-axis indicate an improvement in cluster characteristics (greater compactness and better separation) after the removal of outliers.

Conclusion

The results for quantitative data can be summarized in a few key points. The quality of clusters, as measured by the Silhouette index, does not necessarily correspond with the accuracy of object classification into groups, as measured by the Adjusted Rand Index (ARI). The single-linkage method, along with the average-linkage method, was among the most sensitive to the presence of outliers. In contrast, the k-means method and Ward's method proved to be the most robust in our experiment, maintaining high classification accuracy even in the presence of outliers. The most significant negative impact on classification accuracy (based on ARI) across all tested clustering algorithms was observed with global outliers. The application of methods for identifying and removing outliers may or may not lead to increased compactness and separability of clusters.

In the context of qualitative data, it should be noted that the approach used to derive categorical variables for our analysis is experimental in nature, and the results may not be as robust as in the case of quantitative data. Among the tested clustering methods for categorical data, the average-linkage method performed the best, followed by the k-modes method. When examining the impact of outlier removal on cluster quality, a trend of increasing cluster quality with the removal of more outliers can be observed.

Acknowledgment

The research was conducted with institutional support for the long-term conceptual development of science and research at the Faculty of Informatics and Statistics, Prague University of Economics and Business, and Project IGA F4/35/2024.

References

Barnett, V. D., & Lewis, T. (1978). *Outliers in statistical data*. 2nd edition. John Wiley & Sons

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF. *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. https://doi.org/10.1145/342009.335388

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, Part II, 179–188.

Gan, G., Ma, C., & Wu, J. (2021). *Data clustering: Theory, algorithms, and applications.* Society for Industrial and Applied Mathematics.

He, Z., Xu, X., Huang, Z., & Deng, S. (2005). FP-Outlier: Frequent pattern based outlier detection. *Computer Science and Information Systems*, *2*(1), 103–118. https://doi.org/10.2298/csis0501103h

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*. https://doi.org/10.1109/icdm.2008.17

Nowak-Brzezińska, A., & Gaibei, I. (2022). How the outliers influence the quality of clustering? *Entropy*, *24*(7), 917. https://doi.org/10.3390/e24070917

Pang, G., Cao, L., & Chen, L. (2016) Outlier detection in complex categorical data by modeling the feature value couplings. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, 2016 July 9-15. 1902-1908.

Rand, W. M. (1971). Objective criteria for the evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846. https://doi.org/10.2307/2284239

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Vu, P. N. (2024). *Problematika odlehlých pozorování v klasifikačních úlohách.* [Diploma thesis], Prague University of Economics and Business.

Contact

Jana Cibulková
Prague University of Economics and Business
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
jana.cibulkova@vse.cz

Phuong Ngoc Vu
Prague University of Economics and Business
W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic
vung03@vse.cz