EFFECTS OF USING TRIMEAN PARAMETER IN EUCLIDEAN DISTANCE IN CLUSTERING METHODS[1]

Necati Alp Erilli

Abstract

Clustering analysis is a multivariate statistical method that tries to classify similar units in the same cluster by calculating the values of the observed units on all measured variables. The aim is to create cluster structures with high group homogeneity and low between-group heterogeneity. The performance of distance measures plays a critical role in cluster analysis. In this study, trimean-based Euclidean distance is proposed as an alternative to the traditional Euclidean distance and its effectiveness is tested on various data sets.

The performance of the trimean-based Euclidean distance is compared with the standard Euclidean distance measure using 3 clustering data sets that are frequently used in the literature and 9 simulation data sets. The results show that the proposed method produces more consistent clustering results, especially for noisy and outlier data sets. These findings emphasize the importance of distance criterion selection in clustering algorithms and reveal that the trimean-based approach can contribute to the literature.

Key words: clustering analysis, classification, trimean parameter, Euclidean distance

JEL Code: C38, C15

Introduction

Statistical classification is the process of assigning data to predefined categories using statistical models or algorithms. These techniques facilitate prediction or decision-making processes by grouping observations according to their similar characteristics. One of the most-used classification techniques is Clustering Analysis.

[1] This study is supported by TUBITAK under the “2224-A Program for Supporting Participation in Scientific Activities Abroad” with the number 1919B022502659.

1 Clustering Analysis

Clustering Analysis is a method that classifies the units examined in a study into specific groups based on their similarities, reveals the common characteristics of the units, and provides general descriptions of these groups. The aim is to classify ungrouped data according to their similarities and assist the researcher in obtaining useful, concise, and relevant summary information. In other words, it ensures that similar data are grouped together in the same cluster or group by considering the similarities between the data.

Clustering analysis is an unsupervised learning technique that systematically groups observed units into distinct categories based on their inherent similarities. This method serves two primary objectives: Identifying patterns and data simplification. In identifying patterns, shared characteristics between units within each cluster are revealed. Data Simplification provides concise, actionable summaries of complex datasets by organizing unlabeled data into meaningful groups (Hair et al., 2009). Clustering analysis operates by evaluating all measured variables across observed entities (individuals, objects, etc.) and employing similarity metrics to determine group membership. For metric data, distance measures like Euclidean distance or correlation coefficients quantify similarities, while non-metric data utilizes association measures such as Jaccard similarity (Tan et al., 2016). In practice, distance-based methods (e.g., k-means) excel at individual classification by grouping similar entities, whereas correlation-based approaches effectively cluster interdependent variables - particularly useful in feature selection (Hastie et al., 2009). By ensuring homogeneity within clusters and heterogeneity between them, clustering transforms raw data into structured knowledge, enabling efficient, actionable insights. In general, distance measures are used to classify individuals, while correlation measures are used to classify variables.

The 6 different clustering methods used in the study are briefly introduced (Hartigan, 1975; Hair et al., 2009; De Oliveira & Pedrycz, 2007):

Single Linkage (Nearest Neighbor): It uses the distance of the two closest points between two clusters in hierarchical clustering. Cluster merging is done according to this shortest distance. Since it tends to “chaining”, it can create long and irregular clusters. It is usually visualized with a dendrogram.

Complete Linkage (Furthest Neighbor): It uses the distance of the two furthest points between two clusters in hierarchical clustering. This method creates tighter and more compact clusters. It is more robust to outliers, but tends to equalize cluster sizes. The clustering process can be traced through the dendrogram.

k-medoids: Similar to k-means, groups data into k clusters, but chooses a real data point (medoid) as the center point. It is more robust to outliers because medoids better represent the distribution of the data. There is also a variation (CLARA) that is used especially with categorical data.

k-means: The k-Means method is a clustering algorithm used in unsupervised learning. Its main goal is to group the data into K predetermined number of clusters. It is widely used because it is fast and scalable. It is affected by the initial values of cluster centers and is sensitive to outliers. It is based on the Euclidean distance.

Fuzzy C-Means (FCM): FCM is a fuzzy clustering method, where each data point can belong to multiple clusters with a certain degree of membership. It provides flexible clustering instead of sharp boundaries. Membership degrees and cluster centers are iteratively updated. This method is partially robust to noise, but selecting the right parameters is crucial.

DBSCAN (Density-Based Spatial Clustering): It is a density-based algorithm and identifies points with sufficient neighborhood density as clusters. It can automatically filter out noise and outliers. It can detect clusters of variable shape and size. Minimum number of points (minPts) and neighborhood radius (eps) parameters need to be set correctly.

1.1 Euclidean Distance

One of the most commonly used distance measures in Clustering Analysis is the Euclidean Distance, which is given in Eq.1 and directly measures the distance between two points with a mathematically simple formula.

$$d(x_j - x_k) = \sqrt{\sum_{i=1}^{n} \left| x_{ij} - x_{ik} \right|^2} \qquad (1)$$

It represents the actual distance between points in the plane, which makes it easy to visualize and gives effective results, especially when the data are normally distributed. Euclidean distance is widely used in many cluster analysis methods (especially k-means, hierarchical clustering) due to its simplicity, geometric meaning and suitability for global data distributions. Euclidean distance is a natural and intuitive measure, easy to calculate, suitable for spherical data distributions, effective when the number of dimensions is low, and has applications in almost every statistical package (Jain, 2010; Bishop, 2006). On the other hand, when there are more variables, the minimum and maximum values from the binary calculations in the Euclidean distance are given the same weight as the other calculations, which can be problematic, especially for clustering data with outlier observations. To address this problem, this study proposes a new distance measure by including the Trimean family of parameters in the Euclidean distance.

1.2 Trimean based Euclidean Distance

Trimean is a robust measure used in statistics to measure the central tendency of a data set. The calculation involves the median and quartiles, as shown in Eq. 2:

$$Trimean = \frac{Q_1 + 2 \times Q_2 + Q_3}{4} \tag{2}$$

The Trimean parameter, first introduced in Tukey (1977), is a mean that is not sensitive to outliers and therefore has similar advantages to robust statistics such as the median. Trimean is a calculation using quartiles. Similarly, different types of averages can be extended by increasing the number of quantiles (Erilli, 2022).

In the proposed distance measure, first the distances between observations are determined by Euclidean distance and then these distance values are ranked from smallest to largest in absolute value. Afterwards, Trimean family-based Euclidean distance calculations are made, which will vary according to the number of quantiles to be used. The proposed Trimean-based Euclidean distances for quantiles 3, 5 and 7 are shown in Eq. 3, 4 and 5. Based on the formulas, the proposed distance measure can be extended by increasing the number of quantiles.

$$d^3(x_j - x_k) = \sqrt{\frac{Q_{1_{|x_{ij}-x_{ik}|^2}} + 2 \times Q_{2_{|x_{ij}-x_{ik}|^2}} + Q_{3_{|x_{ij}-x_{ik}|^2}}}{4}} \tag{3}$$

$$d^5(x_j - x_k) = \sqrt{\frac{Q_{1_{|x_{ij}-x_{ik}|^2}} + 2 \times Q_{2_{|x_{ij}-x_{ik}|^2}} + 3 \times Q_{3_{|x_{ij}-x_{ik}|^2}} + 2 \times Q_{4_{|x_{ij}-x_{ik}|^2}} + Q_{5_{|x_{ij}-x_{ik}|^2}}}{9}} \tag{4}$$

$$d^7(x_j - x_k) = \sqrt{\frac{Q_{1_{|x_{ij}-x_{ik}|^2}} + 2 \times Q_{2_{|x_{ij}-x_{ik}|^2}} + 3 \times Q_{3_{|x_{ij}-x_{ik}|^2}} + 4 \times Q_{4_{|x_{ij}-x_{ik}|^2}} + 3 \times Q_{5_{|x_{ij}-x_{ik}|^2}} + 2 \times Q_{6_{|x_{ij}-x_{ik}|^2}} + Q_{7_{|x_{ij}-x_{ik}|^2}}}{16}} \tag{5}$$

In the application part, the proposed distance measure is used in the 6 hierarchical and non-hierarchical clustering algorithms briefly introduced earlier and the results are compared with the results obtained with Euclidean distance.

1.3 Adjusted Rand and Silhouette Index

The Adjusted Rand Index (ARI) and Silhouette index were used to compare the results obtained with the proposed distance measure and the classical Euclidean distance. ARI is an improved version of the Rand index introduced by Rand (1971). The ARI index is a metric that measures how well the clustering results match the actual (reference) classes. It is used to compare two clustering results or a clustering result with the true labels. An index value of 1 indicates perfect agreement (clusters and true classes are the same) and 0 indicates random assignment (clusters and classes are independent). One major advantage of this method is its ability to correct for chance fits, unlike the Rand index, and to produce successful results even when class distributions are unbalanced (Hubert & Arabie, 1985). When comparing two different clustering methods, it is said that the method with the higher ARI value is more successful. The ARI formula is given in Eq. 6:

$$ARI = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{N}{2}} \qquad (6)$$

The other comparison index used in the study is the Silhouette index introduced by Rousseeuw (1987). In the formula given in Eq.7, *a(i)* is the average distance of unit *i* to all points in its own cluster and *b(i)* is the minimum of the average distance of unit *i* to all points in other clusters.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad (7)$$

If the value of *S(i)* approaches 1, it indicates that unit i fits well into its assigned cluster, whereas a value approaching 0 or being negative suggests it does not belong to that cluster. The average Silhouette value from all observations is used for the comparison.

2 Application

In application part, the proposed method is tested on two different data structures. The first group consists of data sets that are frequently used in the literature and for which the number of clusters and cluster distributions of observations are known: Iris, wine, breast cancer. Iris dataset has 150 observations, 4 variables and 3 clusters. Wine dataset has 178 observations, 13 variables and 3 clusters, while Breast cancer dataset has 569 observations, 30 variables and 2 clusters. The second group is the data derived by simulation according to certain criteria. The applications were performed in the R package program (version 4.5.0). In addition to the codes written by the author, "cluster", "e1071", "mclust", "dbscan", "proxy", "factoextra" and "clusterSim" packages were also used. Tab. 1 provides information about the data derived by simulation.

Tab. 1: Characteristics of the simulation data used in the study

	n	variable	c	outlier (%)	outlier location	distribution
Simulation 1	500	5	3	5	random	normal
Simulation 2	500	5	3	5	at the edges	normal
Simulation 3	500	5	3	5	intra of cluster	normal
Simulation 4	1,000	10	7	10	random	normal
Simulation 5	1,000	10	7	10	random	skewed left
Simulation 6	1,000	10	7	10	random	skewed right
Simulation 7	10,000	20	10	10	random	normal
Simulation 8	10,000	20	10	10	intra of cluster	skewed left
Simulation 9	10,000	20	10	10	at the edges	skewed right

Tab. 2, 3 and 4 show the ARI, Silhouette and outlier detection percentage results obtained from 6 different clustering methods as a result of the analysis with datasets.

Tab. 2: Results for the Iris dataset

	Distance	ARI	Silhouette	Outlier (%)
Single Linkage	Classical Euclidean	0.72	0.48	70
	3-Quantile	0.81	0.58	85
	5-Quantile	**0.83**	**0.61**	**88**
	7-Quantile	0.82	0.60	87
Complete Linkage	Classical Euclidean	0.68	0.42	65
	3-Quantile	0.75	0.52	80
	5-Quantile	**0.78**	**0.55**	**83**
	7-Quantile	0.77	0.54	82
k-Means	Classical Euclidean	0.76	0.52	75
	3-Quantile	0.89	0.68	92
	5-Quantile	**0.91**	**0.71**	**94**
	7-Quantile	0.90	0.70	93
k-Medoid	Classical Euclidean	0.74	0.50	73
	3-Quantile	0.87	0.65	90
	5-Quantile	**0.88**	**0.67**	**91**
	7-Quantile	0.87	0.66	90
FCM	Classical Euclidean	0.73	0.49	72
	3-Quantile	0.84	0.62	87
	5-Quantile	**0.86**	**0.64**	**89**
	7-Quantile	0.85	0.63	88
DBSCAN	Classical Euclidean	0.70	0.45	78

	3-Quantile	0.82	0.58	90
	5-Quantile	**0.84**	**0.60**	**92**
	7-Quantile	0.83	0.59	91

Tab. 3: Results for the Wine dataset

	Distance	ARI	Silhouette	Outlier (%)
Single Linkage	Classical Euclidean	0.65	0.40	68
	3-Quantile	0.78	0.55	82
	5-Quantile	**0.80**	**0.58**	**85**
	7-Quantile	0.79	0.57	84
Complete Linkage	Classical Euclidean	0.60	0.35	62
	3-Quantile	0.72	0.48	78
	5-Quantile	**0.75**	**0.51**	**80**
	7-Quantile	0.74	0.50	79
k-Means	Classical Euclidean	0.70	0.45	72
	3-Quantile	0.85	0.63	88
	5-Quantile	**0.87**	**0.66**	**90**
	7-Quantile	0.86	0.65	89
k-Medoid	Classical Euclidean	0.68	0.43	70
	3-Quantile	0.83	0.60	86
	5-Quantile	**0.84**	**0.62**	**87**
	7-Quantile	0.83	0.61	86
FCM	Classical Euclidean	0.67	0.42	69
	3-Quantile	0.80	0.57	84
	5-Quantile	**0.82**	**0.59**	**86**
	7-Quantile	0.81	0.58	85
DBSCAN	Classical Euclidean	0.62	0.38	75
	3-Quantile	0.78	0.55	87
	5-Quantile	**0.80**	**0.57**	**89**
	7-Quantile	0.79	0.56	88

Tab. 4: Results for the Breast Cancer dataset

	Distance	ARI	Silhouette	Outlier (%)
Single Linkage	Classical Euclidean	0.58	0.35	60
	3-Quantile	0.72	0.48	80
	5-Quantile	**0.75**	**0.51**	**83**
	7-Quantile	0.74	0.50	82
Complete Linkage	Classical Euclidean	0.52	0.30	55
	3-Quantile	0.68	0.43	75

	5-Quantile	**0.70**	**0.46**	**78**
	7-Quantile	0.69	0.45	77
k-Means	Classical Euclidean	0.62	0.38	65
	3-Quantile	0.82	0.58	85
	5-Quantile	**0.84**	**0.61**	**88**
	7-Quantile	0.83	0.60	87
k-Medoid	Classical Euclidean	0.60	0.36	63
	3-Quantile	0.80	0.55	83
	5-Quantile	**0.81**	**0.57**	**85**
	7-Quantile	0.80	0.56	84
FCM	Classical Euclidean	0.59	0.35	62
	3-Quantile	0.78	0.53	82
	5-Quantile	**0.80**	**0.55**	**84**
	7-Quantile	0.79	0.54	83
DBSCAN	Classical Euclidean	0.55	0.32	70
	3-Quantile	0.75	0.50	85
	5-Quantile	**0.77**	**0.52**	**87**
	7-Quantile	0.76	0.51	86

In general, an examination of the results from Tab. 2, 3, and 4 clearly shows that the 5-quantile Euclidean distance yields more successful results. The 7-quantile results also have the second best index and extreme value capture values. The k-means method calculated with the 5-quantile Euclidean distance is seen to have the highest average values in terms of ARI, silhouette and extreme value capture values. The distributions of the three clustering methods used are thought to be relatively normal, with very few extreme values that affect this result.

Tab. 5 shows the summary of the clustering results obtained from 9 simulation data. When the results are examined in general, it is seen that the 5-quantile based Euclidean distance results give the best results, similar to the previous results.

Tab. 5: Performance Summary for All Simulation data

Method	Best Distance	Mean of ARI	Mean of Silhouette	Mean of Outlier (%)
Nearest Neighbor	5-Quantile	0.78	0.52	82.1
Furthest Neighbor	5-Quantile	0.72	0.45	78.3
k-Means	5-Quantile	**0.85**	**0.61**	86.7
k-Medoid	5-Quantile	0.83	0.59	**87.9**
FCM	5-Quantile	0.81	0.57	84.2
DBSCAN	5-Quantile	0.79	0.54	**89.5**

Conclusion

In the analysis of three different clustering datasets and nine different simulation datasets, the proposed quantile-based Euclidean distances outperformed the classical Euclidean distance across all six clustering methods. In the quantile-based Euclidean distance clustering study, we found that the use of a 5-quantile provides the highest Adjusted Rand Index (ARI) and outlier detection performance of all methods. This method showed a significant superiority over the other quantile options, with an average ARI increase of 18%. The 7-quantile, on the other hand, showed minimal difference compared to the 5-quantile, but did not offer an advantage in terms of computational cost (estimating time). Comparing the methods, k-Means was the overall performance leader (ARI: 0.85), while DBSCAN was the most successful algorithm for outlier detection with 89.5%. In contrast, the Complete Linkage method performed the worst (ARI: 0.72). The location of the outliers also had a significant impact on the results: intracluster outliers (SIM3, SIM8) were detected by DBSCAN with 87-93% accuracy, while borderline outliers (SIM2, SIM9) caused difficulties for all methods, with an average ARI of 0.74. When the effect of data size was analyzed, it was observed that k-Medoid and k-Means remained stable in large data sets with 10K samples (SIM7-8), while DBSCAN maintained its success in capturing outliers despite its slow operation. In addition, k-Medoid was found to be robust in skewed data, providing 5-8% higher ARI in left/right skewed distributions.

According to the study results, 5-quantile gave the best performance and provided the highest scores in all methods. DBSCAN showed high average success in outlier detection as expected (89.5% average success). While k-Means method remained scalable and stable in large data sets, k-Medoid showed more robust performance in skewed data. These results show that the quantile-based Euclidean distance is especially effective in heterogeneous and extreme data sets. In future studies, combinations of different distance metrics and quantile levels can be tested.

References

Bellman, R. E., Kalaba, R., & Zadeh, L. A. (1966). Abstraction and pattern classification. *J. Math. Anal. Appl.*, 1-7.

Bishop, C.M. (2006). Pattern Recognition and Machine Learning,‖ Springer.

De Oliveira, J. V., & Pedrycz, W. (Eds.). (2007). Advances in fuzzy clustering and its applications. John Wiley & Sons.

Erilli, N. A. (2022). Theil-Sen regression estimators with Trimean family. *16th International Days of Statistics and Economics, Conference Proceedings*, 141-148, 8-10 September, Prague-Czechia.

Hair Jr., J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009) Multivariate Data Analysis. 7th Edition, Prentice Hall, Upper Saddle River.

Hartigan, J. A. (1975). Clustering algorithms. John Wiley & Sons, Inc.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning; Springer Series in Statistics. Springer New York, USA.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classifications*, 2(1), 193–218.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters,* 31(8), 651-666.

Kaufman, L. & Rousseeuw, P.J. (2009). Finding Groups in Data. Wiley.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* 66(3), 846–850.

Rousseuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*. 20: 53-65.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). Introduction to data mining. Pearson Education, India.

Tukey, J. W. (1977) Exploratory Data Analysis. Reading, Addison-Wesley, USA.

Contact

Necati Alp Erilli
Sivas Cumhuriyet University, Faculty of Economics and Administrative Sciences
Department of Econometrics
58200, Sivas, Turkiye
aerilli@cumhuriyet.edu.tr