Journal of Statistical Modelling: Theory and Applications
Vol. 5, No. 01, 2024, pp. 53-64
Yazd University Press 2024

JSMTA

Research Paper

Ranking judicial branches using clustering algorithm

Zohreh Farhadi*[1], Mohadeseh Alsadat Farzammehr[2]
[1]Judicial Department of Shahrood County, Semnan, Iran
[2]Judicial Research Institute, Tehran, Iran

Received: March 1, 2024/ Revised: October 04, 2024/ Accepted: October 24, 2024

Abstract: The performance of judiciary branches is evaluated based on specific indicators determined by the Statistics and Information Technology Center of Judiciary. These indicators, which are usually documents recorded in court cases, have a specific administrative or judicial score for the branch, and by calculating the total scores, the performance of the branches is evaluated. However, with the expansion of these indicators, ranking and evaluating branch performance has become more complex. In this article, clustering is used as one of the most important data mining tools to evaluate branch performance. By identifying similar branches, examining branches, and facing upcoming challenges more effectively, more effective decisions can be made in the judiciary system. Here, to organize 19 law branches based on 49 different administrative and judicial indicators, the K-means clustering algorithm is applied based on two criteria of Euclidean dissimilarity distance and random forests. In addition, the Dunn index is used to evaluate clustering. The value of this index is calculated as 0.82 by applying the dissimilarity of random forests, indicating the successful performance of the algorithm used in determining similar branches.

Keywords: Administrative Score; Branch Performance Evaluation; Clustering; Judicial Score.
Mathematics Subject Classification (2010): 62HXx, 62H30.

1 Introduction

Evaluating the performance of branches in the judicial domain is a vital task for maintaining an efficient judicial system. Currently, this evaluation relies on predefined

*Corresponding author: Zohreh.farhadi87@gmail.com

indicators set by the Center for Statistics and Information Technology of the Judiciary, primarily derived from records in legal cases. The use of judicial indicators is a process through which data related to the judicial system is collected, categorized, and communicated to serve as a basis for learning, testing, and decision-making within that system. Essentially, judicial indicators can be used to summarize and convey a significant volume of essential information regarding various aspects of the judicial apparatus. They serve as valuable tools for assessing performance, addressing challenges, defining criteria, monitoring progress, and evaluating the impact of interventions or reforms. In conjunction with other monitoring and evaluation mechanisms, indicators are essential for enhancing transparency and responsiveness in judicial units. They are crucial for providing valuable feedback to policymakers and reformers.

Judicial indicators can pursue various objectives, many of which are compatible with each other, but their purpose must always be as clear as possible. As performance and accountability metrics, they can facilitate valuable reforms and effective strategic activities. These indicators assign specific administrative or judicial scores to each branch, and the overall performance is evaluated by aggregating these scores. However, with the increasing number and complexity of these indicators, the process of ranking and evaluating branch performance becomes more intricate, requiring new methods to address this challenge. In recent years, data mining and exploratory data analysis techniques have prominently emerged across various domains for discovering valuable insights and facilitating decision-making processes. Clustering is one of the most effective data mining methods for analyzing data and uncovering hidden relationships, attracting the attention of researchers in various fields. In this method, data is divided into smaller sets called clusters without prior knowledge of the data's structure, such that members within a cluster have the highest similarity to each other and the lowest similarity to members of other clusters. Thus, despite the unknown nature of groups within the original data set and even the number of divisions, clustering makes relationships among data as apparent as possible.

Virtually all clustering algorithms are based on the concept of similarity or dissimilarity between data and use various metrics to measure this similarity. In clustering, dissimilarity metrics are used to measure the distance and difference between members of each cluster. These metrics are computed based on features or temporal distances between samples and assist us in providing the best clustering based on the similarities or differences among cluster members. To measure this similarity and dissimilarity, common metrics such as Euclidean, Manhattan, Minkowski, and others are employed in clustering methods. By employing clustering and branch grouping techniques based on common features, a more comprehensive understanding of branch performance can be achieved, facilitating informed decision-making and strategic planning within the judicial system. Several studies have demonstrated the effectiveness of clustering in similar domains. For example, Ahmadi (2018) used clustering to assess the efficiency of bank branches. In his research, he proposes a multi-step approach that combines clustering and data envelopment analysis methods to identify management clusters of bank branches and examine their performance.Herrera-Restrepo et al. (2016), in their study, grouped bank branches using a combination of clustering and multivariate data analysis methods and studied their efficiency.

Among other studies in this field, Smith et al. (2019) employed hierarchical cluster-

ing to analyze branch court performance and identified distinct clusters based on case workload, processing time, and judicial efficiency. Their findings highlighted the potential of clustering techniques in assessing branch performance. Additionally, Johnson and Brown (2021) utilized the K-means algorithm for branch court clustering and examined the relationship between cluster membership and case outcomes. Their study revealed distinct performance indices among clusters and emphasized the connection between clustering and branch evaluations. Furthermore, Farzammehr (2021) applied a combination of hierarchical clustering and principal component analysis to evaluate the performance of legal branches in the court system. The results of this research showed that clustering can be an effective alternative to traditional methods in assessing branch performance. In this article, based on previous research, the focus is on ranking and evaluating 19 legal branches based on 49 different administrative and judicial indices. To achieve this, the K-means algorithm is employed, which has been widely used in similar studies (Aminzadeh and Minaii, 2008; Garcia et al., 2018; Chen et al., 2020). Subsequently, two dissimilarity metrics, namely the Euclidean distance and random forests, are utilized to determine the optimal clustering configuration for the given dataset. The use of quality evaluation indices for clustering not only leads to an improvement in the quality of clustering but also aids in better understanding the data structure. Therefore, the performance of clustering results is assessed using the Dunn index, a popular metric for evaluating cluster quality Dunn (1974) In general, the results of this study have the potential to improve decision-making processes, enhance efficiency, allocate resources, and contribute to the overall improvement of the judicial system. Here, clustering, as a data analysis technology, divides judicial branches into smaller and manageable groups based on their common features. This method can serve as an effective tool for assessing and comparing the performance of judicial branches, enabling the enhancement of ranking and better management of these branches.

In the following sections, in order to examine the ranking of branches of a judicial unit using clustering, we will first review the data and the concept of clustering in the judiciary. Then, we will describe the clustering methods used in the article and explain the research algorithm. Finally, we will evaluate the clustering performed using the Dunn index and review the research results.

2 Data and research methodology

The Performance Evaluation Process of the Country's Executive Organizations is defined in the regulations as follows: "Performance evaluation is a comprehensive assessment process of the executive organizations, encompassing aspects such as efficiency, effectiveness, empowerment, and responsiveness within the framework of scientific management principles to achieve organizational goals and duties based on executive plans." In the report titled "Performance Report of Judicial Units," which is published monthly by the Center for Statistics and Information Technology of the Judiciary, the performance of judicial units in the provinces is assessed and ranked in two sections: "administrative" and "judicial." The introduction of this report, dated Farvardin 1393 (March 2014), provides the following explanation: "One of the systems under the Electronic Justice Plan is the Comprehensive Statistical System (Saja). This system aims to

identify all statistical needs at different levels of the judiciary and the environment outside the judiciary. Given that the data of judicial branches are recorded in the Case Management System (Samp), indicators have been extracted to ensure the accuracy and precision of the forms in Samp. These indicators are designed in two categories, administrative and judicial, using a total of 49 indicators." In this article, based on these 49 indicators, clustering of 19 legal branches of a judicial unit is discussed.

The K-means clustering algorithm used in this article is a dissimilarity-based approach to observations, first introduced by Lloyd for its simplicity and efficiency, and it is widely used in many clustering problems (Lloyd, 1982). In this algorithm, K cluster centers are initially chosen at random from the data set, and then each observation is assigned to the cluster whose center it is closest to, or in other words, the one with the least dissimilarity. After assigning all data points to clusters, the mean of the data in each cluster is considered as the new cluster center, and data reassignment is performed. This algorithm is repeated until the cluster centers no longer change. In the K-means algorithm, selecting an appropriate dissimilarity measure is crucial for calculating dissimilarity between data points and performing clustering operations. In this article, two dissimilarity measures, Euclidean distance and random forests, are chosen due to their specific features (Tong et al., 2022).

Euclidean distance is a common metric in clustering, calculated based on the geometric distance between points in a multi-dimensional space. Since the K-means algorithm operates based on the distance between cluster centers and data points, Euclidean distance is chosen as a simple and applicable metric in this algorithm. Random Forest Dissimilarity is another metric calculated based on the difference between random forests generated for the data and is entirely different from other distance functions. In this approach, multiple random forests are created with different settings, and the degree of agreement among these forests for each data point is examined. This metric takes advantage of the high diversity and flexibility of random forests and also utilizes feature weighting to improve clustering quality.

Using these two metrics in the K-means algorithm is logical due to their fast execution and computational simplicity, along with their ability to model various data realities. Euclidean distance, given the geometric nature of multi-dimensional space, is easy to understand and interpret. On the other hand, Random Forest Dissimilarity enhances the algorithm's response when dealing with highly diverse and complex data. These two metrics also exhibit different behavior compared to other dissimilarity metrics when faced with a high number of variables. In some cases, dissimilarity metrics can be ineffective when dealing with a high number of variables. Therefore, the simultaneous use of Euclidean distance and Random Forest Dissimilarity in the K-means algorithm allows for more precise and efficient clustering computations, leading to better results in assessing the performance of legal branches within the country's judicial system.

It's important to clarify that opposite to clustering methods, classification methods are used. In classification methods, data grouping is predefined, and the goal is to establish rules for assigning future data to these groups. Random Forests (RF) is one of the efficient classification methods in data mining. A random forest is a collection of decision trees. In a decision tree classification, the p-dimensional space of variables is randomly divided into smaller subspaces, with observations having the maximum

homogeneity in each subspace. The algorithm starts by separating the dataset into two subspaces. Each of the created subspaces is further divided into smaller subspaces, and this process continues until the observations in each subspace have maximum homogeneity (Breiman, 2001; Tong et al., 2022).

At each stage of partitioning the space to find the best type of division, all explanatory variables and all possible values of those variables are searched to divide the variable space into two suitable parts. In other words, at each stage, the question is: to achieve maximum homogeneity of the data, in which direction should the variable space be partitioned in relation to which explanatory variable and from which observed value? For this purpose, the changes in a function called the impurity function are calculated for all possible partitions. The partition that results in the highest changes in the impurity function is considered the optimal partition. This means that determining the optimal partition of region t_p into two smaller regions t_l and t_r is equivalent to solving the following maximization problem

$$\arg\max \Delta i(t),$$

where $i(t)$ is the impurity function and $\Delta i(t) = i(t_p) - P_l i(t_l) - P_r i(t_r)$. Also, P_l and $P_r$are the proportions of the total observations in regiont_p that are located in regions $t_l$and t_r, respectively.

A random forest is a collection of decision trees, with the difference that, each time it works with a self-replicating sample of the original data and m variables randomly chosen from p original variables to divide the subspace into smaller subspaces. Data similarity in the random forests method is based on the placement of observations in the final subspaces. More precisely, the similarity between two data points, x and y, in the random forest classification, is the number of times the two observations, x and y, end up next to each other in the latest branches of the decision trees.

To calculate the dissimilarity of random forests in clustering problems, the clustering problem must be converted into a classification problem. For this purpose, it is necessary to define at least two classes within the data. In clustering problems where no information about data grouping is available and there is essentially no classification, artificial data equal to the original data is generated using the bootstrap method, simulating two artificial classes. In this way, the clustering problem is transformed into a classification problem, and data similarity and then dissimilarity between them are calculated (Farhadi and Shahsavani, 2015; Yu et al., 2021; Bicego, 2023). As mentioned, in clustering, the goal is to divide data into clusters with similar features. To facilitate the determination of similarity among data, especially when data have many features that might complicate assessing their similarity, dimensionality reduction techniques such as non-metric multidimensional scaling (NMDS) are often employed. In the NMDS method, the original data, which reside in a high-dimensional space with many features, are transformed into a new space created by NMDS with significantly fewer dimensions. Essentially, this method can represent data in a lower-dimensional space than their actual dimension while preserving the similarity between data points. Due to the reduced dimensionality of this space, determining similarity between different data points becomes much simpler, and based on this, data can be grouped into similar clusters. Therefore, NMDS is considered a dimensionality reduction technique that is useful and effective in various aspects of multivariate data analysis and can contribute

to improving clustering. In this article, we have used this method as an exploratory data visualization technique to plot the data and reflect the clustering results in two dimensions, enabling visual assessment of the results (Hernández-León et al., 2022; Dalmaijer et al., 2022).

Furthermore, it is necessary to use quality evaluation indices in order to improve the quality of clustering and gain a better understanding of the data structure. The Dunn index is one of the most well-known indices for evaluating clustering quality (Dunn, 1974). This index assesses the quality of clustering based on the size and shape of clusters, the distance between them, and the reference vector obtained within each cluster. In fact, the Dunn index is an improved version based on two other indices, the internal index and the data index.

This index is utilized in many clustering methods, including K-means, PAM, and DBSCAN, among others. Generally, the objective of the Dunn index is to identify dense and well-separated clusters. It is defined as the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance. The Dunn index's range of variation is between zero and one, and since the goal of clustering is to have clusters with high intra-cluster similarity and low inter-cluster similarity, higher values of this index indicate better clustering performance. In addition to evaluating the clustering method's performance, the Dunn index can be used to estimate the number of clusters. This is done by calculating the Dunn index for different numbers of clusters, and the optimal number of clusters is the one that maximizes the Dunn index (Handl et al., 2005; Ros et al., 2023).

3 Findings

In this study, two dissimilarity measures, Euclidean distance and RF, are used for running the K-means algorithm. The algorithm's flowchart is presented in Figure 1. It should be noted that data standardization was performed before calculating the Euclidean distance. However, as RF dissimilarity is distribution-independent, there is no need for preprocessing such as standardization in its calculation. To obtain RF dissimilarity, artificial data is generated using the self-bootstrapping method and combined with the original data.

By assigning class labels 0 to the original data and 1 to the artificial data, a response variable is defined, transforming the clustering problem into a classification problem. Then, random forests classification is executed with parameters $m = 4, 5, \ldots, 10$ (the number of selected variables out of 49 in each stage of the space expansion) and $ntree = 100, 200, \ldots, 1000$ (the number of decision trees run in the random forests) and the dissimilarity matrix is calculated for each case and used as input in the K-means algorithm. By computing the Dunn index for each of these cases, it is concluded that the best clustering occurs for $m = 6$ and $ntree = 800$. To estimate the number of clusters, the K-means clustering was executed for different numbers of clusters, $K = 1, 2, \ldots, 10$, and the Dunn index value was calculated for each clustering. Then, the number of clusters that resulted in the maximum value of this index was considered as the optimal number of clusters. Therefore, to estimate and determine all unknown parameters in this study, a trial-and-error method was used to maximize the Dunn index and, consequently, achieve the optimal clustering. Table 1 displays the estimated

number of clusters for each dissimilarity matrix. It is noteworthy that applying two different dissimilarity measures has led to different numbers of clusters.

Table 1: Optimal number of clusters.

Dissimilarity	Estimated number of clusters
RF	3
Euclidean distance	2

Figures 2 and 3 represent the clustering results visualized using multidimensional scaling. As the final step of the research algorithm, the performance evaluation of the K-means algorithm was carried out based on two dissimilarity measures: Random Forest (RF) and Euclidean distance, using the Dunn index, and the values are presented in Table 2.

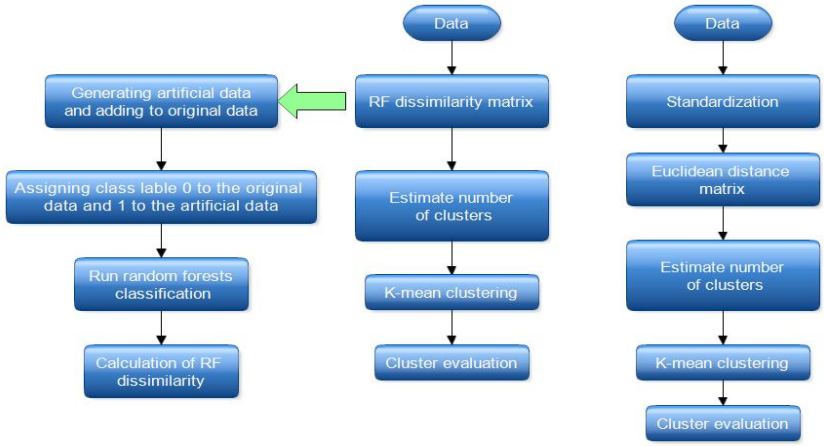

Figure 1: Flowchart of research algorithm for ranking judicial branches. The figure on the right shows the clustering algorithm based on Euclidean distance, and the figure on the left shows the clustering algorithm based on the dissimilarity of random forests.

Table 2: Cluster evaluation.

dissimilarity applied in the K-means algorithm	Dunn Index
RF	0.82
Euclidean distance	0.42

Comparing Figures 2 and 3 allows for a relative evaluation of the clustering results. It can be observed that K-means clustering based on RF dissimilarity achieved better separability, while applying Euclidean distance grouped many branches into a single cluster, indicating the algorithm's inability to identify similar branches.

The comparison of Dunn index values in Table 2 also supports this conclusion. The Dunn index value of 0.82, obtained from the K-means algorithm based on RF dissimilarity, demonstrates the effectiveness of this method in identifying similar branches and discovering hidden structures in the data. Table 3 reports the clustering assignment of each branch. In the RF-based clustering, 8 branches are assigned to Cluster 1, 3 branches to Cluster 2, and 8 branches to Cluster 3. In the Euclidean distance-based clustering, 16 branches are assigned to Cluster 1, and 3 branches to Cluster 2.

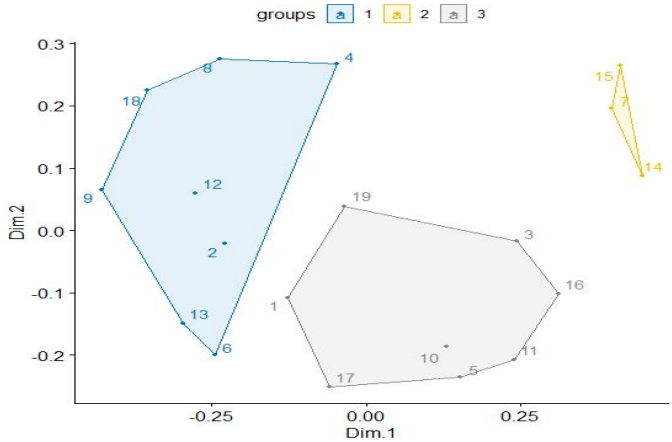

Figure 2: Legal branch clustering based on RF dissimilarity.

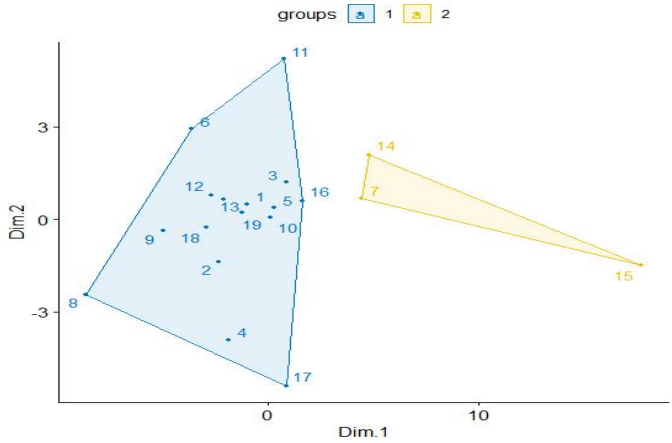

Figure 3: Legal branch clustering based on euclidean distance dissimilarity.

4 Conclusion

This research leveraged two dissimilarity measures, namely Euclidean distance and random forest within the K-means clustering algorithm. Before computing the Euclidean distance, data standardization was performed. In contrast, RF dissimilarity, being independent of data distribution, required no preprocessing, such as standardization. To calculate RF dissimilarity, artificial data was generated through resampling and merged with the original dataset. This transformation entailed assigning the label 0 to original data and 1 to artificial data, effectively turning the clustering problem into a classification task.

In this article, a single algorithm (K-means) with two dissimilarity criteria, Euclidean distance and random forests, has been applied to the legal branch data of a judicial unit. The input parameters for implementing the random forests classification

Table 3: The cluster assigned to each branch.

Branch number	Random Forest	Euclidean distance
1	3	1
2	1	1
3	1	1
4	1	1
5	3	1
6	1	1
7	2	2
8	1	1
9	1	1
10	3	1
11	3	1
12	1	1
13	1	1
14	2	2
15	2	2
16	3	1
17	3	1
18	1	1
19	3	1

method significantly influence the dissimilarity results from this method and consequently the clustering results. Additionally, the number of clusters, which is the input parameter for the K-means algorithm, has a considerable impact on the clustering outcome. In this article, we aimed to optimize both the input parameters of the random forest method and the number of clusters for the clustering method using a clustering evaluation criterion called the Dunn index. For this purpose, the algorithm was executed with different values of these parameters, and ultimately the parameter that maximized the Dunn index was selected as the optimal value for the corresponding parameter. Subsequently, Random Forest classification was conducted with varying parameters, including $m = 4, 5, \ldots, 10$ (representing the number of variables selected from 49 in each subspace) and $ntree = 100, 200, \ldots, 1000$ (denoting the number of decision trees in the Random Forest ensemble). Dissimilarity matrices were calculated for each scenario and utilized as input in the K-means clustering algorithm. The Dunn index served as a quality estimator for each clustering case, and the results indicated that the optimal clustering occurred with $m = 6$ and $ntree = 800$.

For estimating the number of clusters, K-means clustering was executed with different cluster numbers, $K = 1, 2, \ldots, 10$, and the Dunn index values were calculated for each clustering configuration. The number of clusters that yielded the highest Dunn index value was selected as the optimal number of clusters.

In the final phase of this research algorithm, the performance of the K-means algorithm was assessed using both RF and Euclidean distance dissimilarities, employing the Dunn index. The findings revealed that RF-based clustering achieved superior separability and was more effective at identifying similar branches and uncovering hidden data structures compared to Euclidean distance-based clustering. In conclusion, this study adeptly incorporated dissimilarity measures into the K-means clustering process, with the RF dissimilarity measure demonstrating remarkable effectiveness in enhancing clustering quality and understanding data structure. Evaluation of clustering results was conducted through visualizations and the Dunn index, further underscoring

the advantages of RF-based clustering. Moreover, the study utilized a trial-and-error approach for estimating the optimal number of clusters based on the Dunn index, showcasing the flexibility of this methodology. Ultimately, the report presented cluster assignments for each legal branch under both RF and Euclidean distance-based clustering methodologies, offering a comprehensive perspective on their performance. Note that, our study introduces a novel clustering approach using the Random Forest dissimilarity measure, demonstrating significant improvements in clustering performance compared to traditional methods. By transforming the clustering problem into a classification problem with artificial data, our methodology leverages the strengths of Random Forests to achieve better-defined clusters and handle complex data structures more effectively. Comparative analysis with recent literature underscores these advancements: our method achieved higher Dunn index, adjusted Rand index, and Silhouette score values than those reported in studies utilizing K-means (Likas et al., 2003), hierarchical clustering (Rokach and Maimon, 2005), and DBSCAN (Ester et al., 1996). These results highlight the robustness, flexibility, and precision of our approach, particularly in datasets with noise, outliers, and high dimensionality. This study not only contributes to the theoretical understanding of clustering algorithms but also provides practical insights for applications requiring enhanced clustering accuracy and stability.

In comparison to recent studies, our results demonstrate a notable advancement in the application of clustering algorithms for evaluating judicial branches. Farzammehr (2021) pplied a combination of hierarchical clustering and principal component analysis to evaluate the performance of legal branches in the court system, showing that clustering can be an effective alternative to traditional methods in assessing branch performance. These studies corroborate our findings by emphasizing the utility of clustering techniques in performance assessment. However, our research stands out by integrating Random Forest dissimilarity, which significantly enhances clustering quality and data structure understanding compared to traditional Euclidean measures. The superiority of RF-based clustering in our study, as evidenced by higher Dunn index values, underscores its effectiveness in identifying similar branches and uncovering hidden data structures, thereby offering a more nuanced and robust evaluation of judicial branch performance.

References

Ahmadi, M. (2018). Cluster analysis and performance evaluation of bank branches using robust factor analysis and data envelopment analysis. In *Proceedings of the 2nd National Conference on Advances in Electrical, Computer, and Biomedical Engineering (In Persian)*, Cambridge University Press.

Aminzadeh, F. and Minaii, B. (2008). Using clustering for crime characterization in the judiciary. In *Proceedings of the 2nd Iranian Data Mining Conference (In Persian)*.

Bicego, M. (2023). DisRFC: a dissimilarity-based random forest clustering approach. *Pattern Recognition*, **133**:109036.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**:5–32.

Chen, Y., Song, J., Huang, C. and Li, C. (2020). An improved K-means algorithm for detecting malicious nodes in wireless sensor networks. *Ad Hoc Networks*, **95**:102040.

Dalmaijer, E.S., Nord, C.L. and Astle, D.E. (2022). Statistical power for cluster analysis. *BMC Bioinformatics*, **23**(1):205.

Dunn, J.C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**(1):95–104.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD-96 Proceedings*, **96**(34):226–231.

Farhadi, Z. and Shahsavani, D. (2015). Clustering of gene expression data by random forest dissimilarity. *Razi Journal of Medical Sciences*, **22**(136):109–118.

Farzammehr, M.A. (2021). Application of hierarchical clustering on principal components to evaluate the performance of justice system by judicial indicators. *Journal of Statistical Modelling: Theory and Applications*, **2**(2):143–158.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2018). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, **180**(10):2044–2064.

Handl, J., Knowles, K. and Kell, D. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**(15):3201–3212.

Hernández-León, P. and Caro, M.A. (2022). Cluster-based multidimensional scaling embedding tool for data visualization. *Physica Scripta*, **99**(6):066004.

Herrera-Restrepo, O., Triantis, K., Seaver, W.L., Paradi, J.C. and Zhu, H. (2016). Bank Branch Operational Performance: A Robust Multivariate and Clustering Approach. *Expert Systems With Applications*, **50**:107–119.

Johnson, R. and Brown, L. (2021). Clustering court branches for performance evaluation: A case study. *Journal of Legal Analytics*, **8**(2):123–142.

Likas, A., Vlassis, N. and Verbeek, J.J. (2003). The global K-means clustering algorithm. *Pattern Recognition*, **36**(2):451–461.

Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, **28**:129–136.

Rokach, L. and Maimon, O. (2005). Clustering methods. *Data Mining and Knowledge Discovery Handbook*, 321–352.

Ros, F., RIAD, R. and Guillaume, S. (2023). PDBI: a partitioning Davies-Bouldin index for clustering evaluation. *Neurocomputing*, **528**:178–199.

Smith, A., Jones, B. and Williams, C. (2019). Hierarchical clustering of court branch performance indicators. In *Proceedings of the International Conference on Data Mining*, 256–263.

Tong, H., Han, J. and Pei, J. (2022). *Data Mining: Concepts and Techniques*. Netherlands: Elsevier Science.

Yu, J., Zhu, L., Qin, R., Zhang, Z., Li, L. and Huang, Y. (2021). Combining K-means clustering and random forest to evaluate the gas content of coalbed bed methane reservoirs. *Geofluids*, **2021**(1):9321565.