Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Bootstrap approach for clustering method with applications

Sulafah M. Saleh Binhimd, Zakiah I. Kalantan *

Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 28 June 2022
Received in revised form
24 November 2022
Accepted 26 December 2022

Keywords:
Bootstrap method
K-means method
Cluster method
Parametric and semi-parametric methods

ABSTRACT

Discovering patterns of big data is an important step to actionable insights data. The clustering method is used to identify the data pattern by splitting the data set into clusters with associated variables. Various research works proposed a bootstrap method for clustering the array data but there is a weak view of statistical or theoretical results and measures of the model consistency or stability. The purpose of this paper is to assess model stability and cluster consistency of the K-number of clusters by using bootstrap sampling patterns with replacement. In addition, we present a reasonable number of clusters via bootstrap methods and study the significance of the K-number of clusters for the original data set by looking at the value of the K-number that provides the most stable clusters. Practically, bootstrap is used to measure the accuracy of estimation and analyze the stability of the outcomes of cluster methods. We discuss the performance of suggestion clusters through running examples. We measure the stability of clusters through bootstrap. A simulation study is presented in order to illustrate the methods of inference discussed and examine the satisfactory performance of the proposed distributions.

© 2022 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Analyzing big data poses a great challenge for numerous researchers to explore the data structure (Yaqoob et al., 2016). In machine learning, clustering is the most used method to identify complex data patterns. It has been employed to summarize and describe the data set through cluster prototypes. Clustering methods determine the optimal grouping of observations within each cluster is similar. The similarity between pairs is based on some measurement of distance (Naganathan et al., 2016). The identification of clustering is made without detecting the dimension of clusters. This would be useful in order to estimate the intrinsic dimensionality of the data. Usually, the criteria to select the clusters are based on the sum of squared distances that is computed within and/or between clusters, illustrated in Hartigan (1975), Krzanowski and Lai (1988), and Kaufman and Rousseeuw (2009). In the early twenty century, Tibshirani et al. (2001) estimated the number of clusters using the gap statistic while Sugar and James (2003) proposed a jump statistic to find it. Steinley (2008) discussed the stability procedure and developed it in a way to select the cluster by performing k-means several times with a different initializing number. Clustering methods could be classified as hierarchical such as single linkage and non-hierarchical clustering as k-means (Jhun, 1990).

Various fields apply clustering in their applications such as Chemometrics, marketing, Geosciences, political science, and medical research, among others. Many clustering methods are used to classify the data set to cluster, without the possibility of evaluating the stability of cluster results.

The bootstrap method is a resampling method introduced by Efron (1992), it is used for estimating the distribution of statistics based on independent observations, then developed to work with other statistical inferences. In this paper, the bootstrap is used to analyze the stability of cluster results and to estimate the number of clusters in an efficient way. The method is used to estimate the number of clusters that are needed to classify the data set. The studying of a set of related variables in a large data set is usually designed to deal with sub-dimension and moderate size. This step is important to allow a fast analysis of big data. Various methods are proposed to examine the existence of clusters in a data structure. Jhun (1990) used bootstrap in the case of k-means clustering. Some pieces of literature

* Corresponding Author.
Email Address: zkalanten@kau.edu.sa (Z. I. Kalantan)
https://doi.org/10.21833/ijaas.2023.03.023
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7040-5623
2313-626X/© 2022 The Authors. Published by IASE.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

discussed the work of bootstrapping clustered and methods effectiveness which appear the methods worked quite well, although the limitation of theoretical results, such as McCullagh (2000) and Field and Welsh (2007).

In addition, the researchers presented the criteria for suitable and successful bootstrap based on the distribution of the observations but sometimes they didn't measure the model consistency or stability. The aim of this paper is to assess model stability and cluster consistency of the K-number of clusters by using bootstrap sampling patterns with replacement. The measure over bootstrap samples is used to study the significance of the K-number of clusters for the original data set. Practically, the ideal number of clusters is estimated by looking at the value of the K-number that provides the most stable clusters. For each number of bootstrap algorithms.

The rest of the paper is organized as follows: Section 2 introduces the clustering algorithm and explores its computational behavior. In Section 3 the nonparametric bootstrap method is presented, and the procedure of bootstrap is illustrated to measure the stability and estimate the efficient number of clusters. In Section 4, the experimental methodology and results are presented. Finally, conclusions are drawn in Section 5.

2. Clustering methods

Cluster analysis is a tool to find an optimal grouping for data points, where the observations in each cluster are similar and dissimilar to other clusters. In cluster analysis, the number of grouping is not known in advance. We can seek graphically for sub-regions (clusters) by plotting the data points with various graphs depending on the data dimension such as scatterplots or biplots among others. Clustering methods could be classified as hierarchical such as single linkage and non-hierarchical clustering as k-means (Kalantan, 2019). Many methods are used to identify the observation vectors that are similar and group them into clusters. One of the methods is the distance between two data points. A common distance is a Euclidean distance, defined as follows:

$$d(x,y) = \sqrt{(x-y)'(x-y)} = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}.$$

2.1. Determining the number of clusters in data set

Usually, the user needs to determine the number of clusters. In hierarchical clustering, the K clusters are selected from the dendrogram by cutting across the branches at a given level of the distance measure used by one of the axes Tibshirani and Efron (1993). In this case, the number of clusters doesn't need to be handled before the clustering process. On other hand, non-hierarchical (partition) clustering requires the number of clusters before starting the clustering process. In practice, it is difficult to specify the accurate K clusters of the data set. One practical approach is to use the Elbow method which examines the within-cluster dissimilarity as a function of the number of clusters. It uses the turning point in the curve of the sum of within-cluster variance with respect to the number of clusters. Another method for selecting the number of clusters is based on clustering stability. Ben-Hur et al. (2001) tried to obtain the clustering stability by using the distribution of pairwise similarity between clustering of sub-samples of a dataset. Lange et al. (2004) proposed a new measure of clustering stability to assess the validity of a cluster model. Wang (2010) proposed some novel criteria for estimating the number of clusters by measuring the quality of clustering through their instability from sample to sample.

3. Bootstrap approach

The bootstrap method was introduced by Efron (1992). It is a resampling technique for estimating the distribution of statistics based on independent observations and then developed to work with other statistical inferences. It is used for assigning the measures of accuracy of the sample estimate, especially the standard error. The bootstrap sample is obtained by randomly sampling n times with replacement from the original sample. For a good introduction to the bootstrap method see Tibshirani and Efron (1993). The bootstrap method can be applied with a lot of applications, such as tests, regression, and confidence intervals. And used as a useful statistical tool with clustering and mixture models, (Jaki et al., 2018; Fang and Wang, 2012).

3.1. Bootstrap method for selecting the number of clustering

There are many methods proposed for selecting the number of clusters in a clustering algorithm.The main goal is maximizing the cluster stability or minimizing the clustering instability. Fang and Wang (2012) illustrated the algorithm as follows:

- Generate B independent bootstrap sample pairs $(X_i^{n*}, \tilde{X}_i^{n*})$, i=1, …, B, from the original data set, each of size n
- Construct pairs of clusters based on bootstrap sample pairs
- For each pair of clusters calculate their empirical clustering distance
- Estimate the clustering instability by averaging the empirical clustering distance over B bootstrap samples
- Estimate the optimal number of clusters that minimize the estimated clustering instability

This algorithm can be applied to distance base or non-distance-based clustering algorithms and will apply here to estimate the optimal number of

clusters with Olive data. Fang and Wang (2012) used the bootstrap method for it is many advantages. It is more efficient because the bootstrap samples have the same size as the original sample, and can use for any number of clusters. In addition to that, estimating the clustering instability using the bootstrap method is the nonparametric maximum likelihood estimate.

3.2. Testing cluster stability via bootstrap method

The bootstrap method was applied to measure the stability of clusters in Hennig (2007), which is one of the objectives of this paper. A resampling method such as bootstrap provides a general framework within which one can analyze the stability of cluster results. The bootstrap approach proceeds as follows:

- Draw B bootstrap samples of n points $x_n^i$with replacement from the original data set. Let C is a cluster from the original clustering $E_n(x)$
- Compute the clustering $E_n(x_n^i)$
- Let $x_*^i = x_n \cap x_n^i$ be the points of the original data set that are also in the bootstrap sample. Let $C_*^i = C \cap x_*^i$, $\Delta = E_n(x_n^i) \cap x_*^i$
- If $C_*^i \neq 0$, compute the maximum Jaccard similarity between the $C_*^i$and Δ on x_*^i: $\gamma_{C,i} = \max \gamma\ (C_*^i, D)$, D is the maximizer of $\gamma(C_*^i, D)$
- Now there is a sequence $\gamma_{C,i}, i = 1, \dots, B$, then the mean is computed $\bar{\gamma}_C = \frac{1}{B^*}\sum_{i=1}^{B}\gamma_{C,i}$which is the stability measure ($B^*$is the number of bootstrap replications for which $C_*^i \neq 0$).

The general method for drawing statistical conclusions from clustering tools used on gene expression microarray data is presented in Kerr and Churchill (2001). The method makes use of an analysis of the variance model to normalize and evaluate the differential expression of genes under various situations. The bootstrapping method supports statistical reasoning. For estimates of differential expression for specific genes, bootstrapping has previously been used to derive confidence intervals. The authors in this instance use bootstrapping to evaluate the consistency of cluster analysis results.

The use of the cluster bootstrap for extended estimating equation conclusions for clustered/longitudinal data is theoretically justified by Cheng et al. (2013). They demonstrate that the cluster bootstrap produces a consistent approximation of the distribution of the regression estimate as well as a consistent approximation of the confidence sets under the general exchangeable bootstrap weights. They further show that an asymptotically identical inference is provided by a computationally more effective one-step version of the cluster bootstrap.

4. Experimental results

We deal with different types of datasets; artificial and real data. The data variables are scaled before the implementation. The computation results for each data example are presented in the following sections.

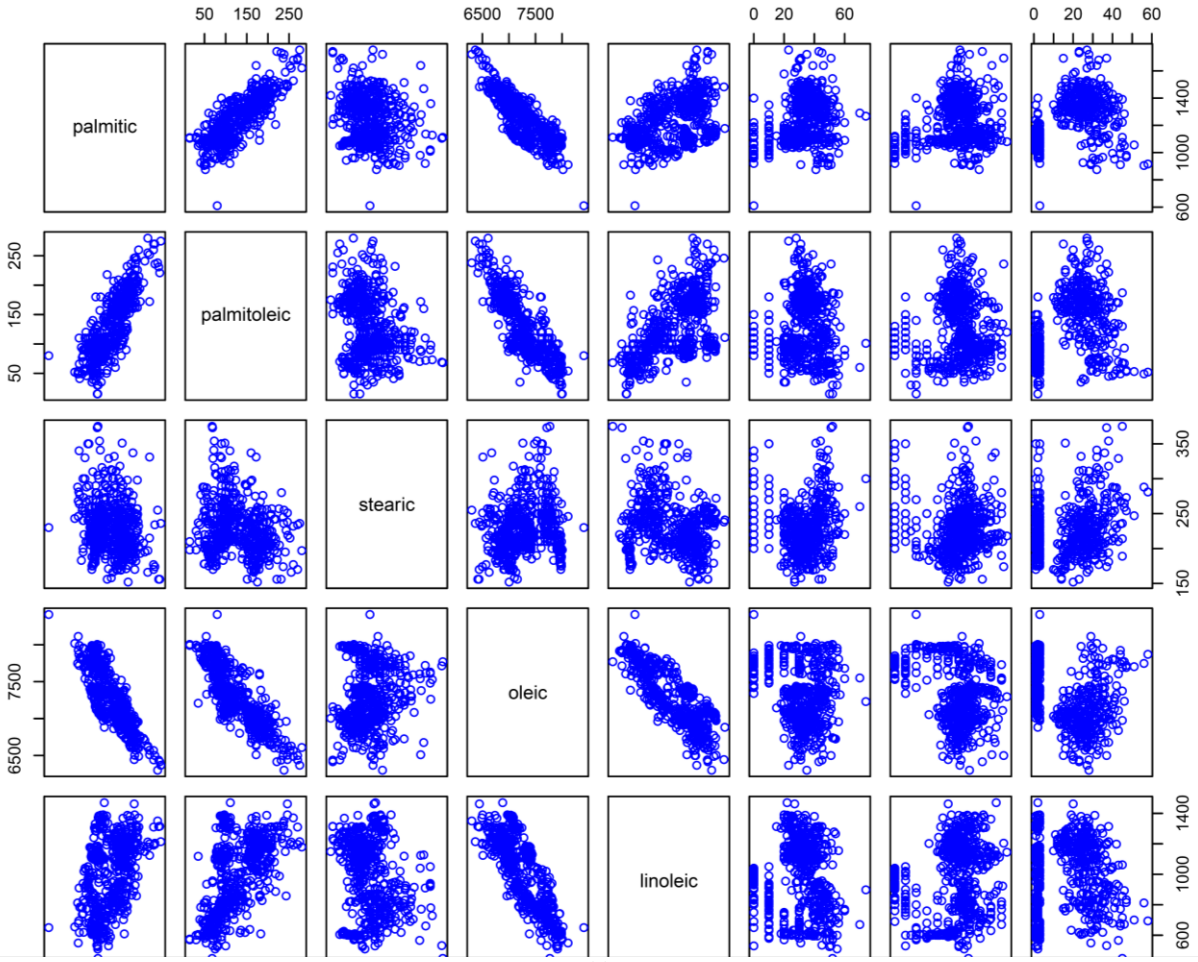
4.1. Olive data

Olive data is available in the pdfCluster package. The data consists of 572 rows of observations. The first and the second column correspond to the area (Centre-North, South, Sardinia) and the geographical region of origin of the olive oils (northern Apulia, southern Apulia, Calabria, Sicily, inland Sardinia, and coast Sardinia, eastern and western Liguria, Umbria), respectively. The remaining columns represent the chemical measurements (on the acid components for the oil specimens) palmitic, palmitoleic, stearic, oleic, linoleic, linolenic, arachidic, and eicosanoid.

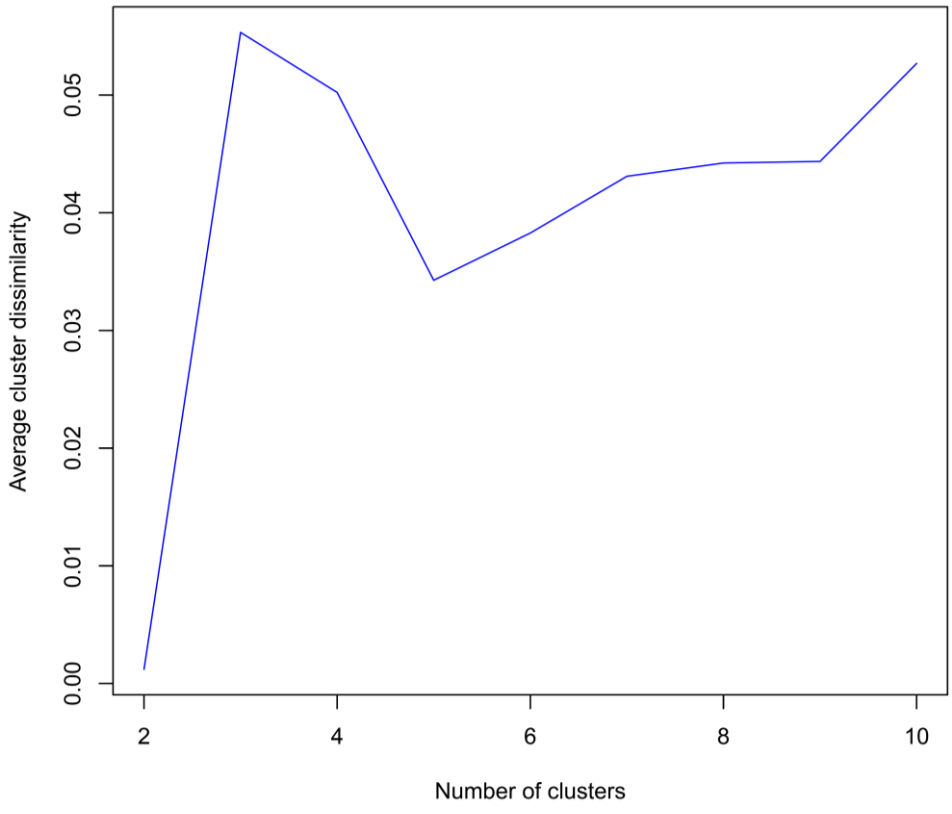
In this paper the bootstrap method is applied for two missions: select the number of clusters and measure the stability of clusters. For the first mission, 50 bootstrap sample pairs are generated and the noise MCLUST CBI cluster algorithm is applied for the number of clusters k, k=2, …, 10. This method selects the optimal number of clusters as $\hat{k} = 2$, which estimates the true optimal number of clustering, see the result in Fig. 1. For the second mission, to measure the stability of clusters, 100 bootstrap samples are generated and used the same cluster method to measure the stability of $\hat{k} = 2$ clusters. The results are 0.9997315 for cluster 1 and 0.9998065 for cluster 2, which is considered as a high stable.

4.2. Air quality data

Air Quality data has 111 observations recorded in New York in 1973, which is a daily measurement of air quality. The data consists of six numerical variables; (Ozone) mean ozone, (Solar.R) solar radiation, (Wind) average of wind speed, (Temp)maximum daily temperature, month, and day. The first four variables are considered in our implementation. When following the same method used with Olive data we found that the optimal number of clusters is $\hat{k} = 8$, see Fig. 2. By studying the stability of these clusters, as presented in Fig. 3, we find that they take the following values: 0.7478524, 0.6964059, 0.8838197, 0.6014733, 0.5591663, 0.6065413, 0.6883570 and 0.8809348, respectively, which considered as good stability.

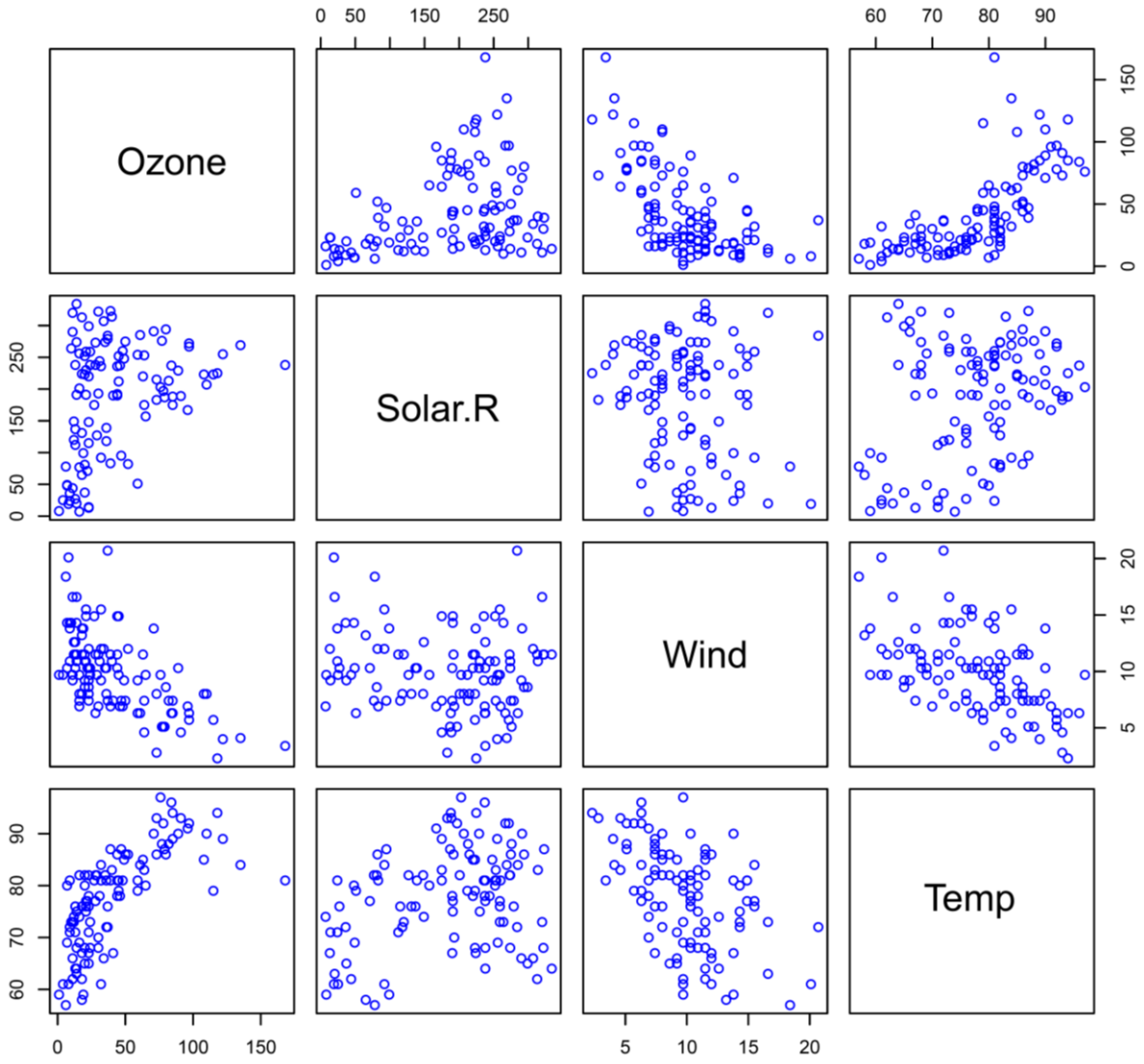


(a)

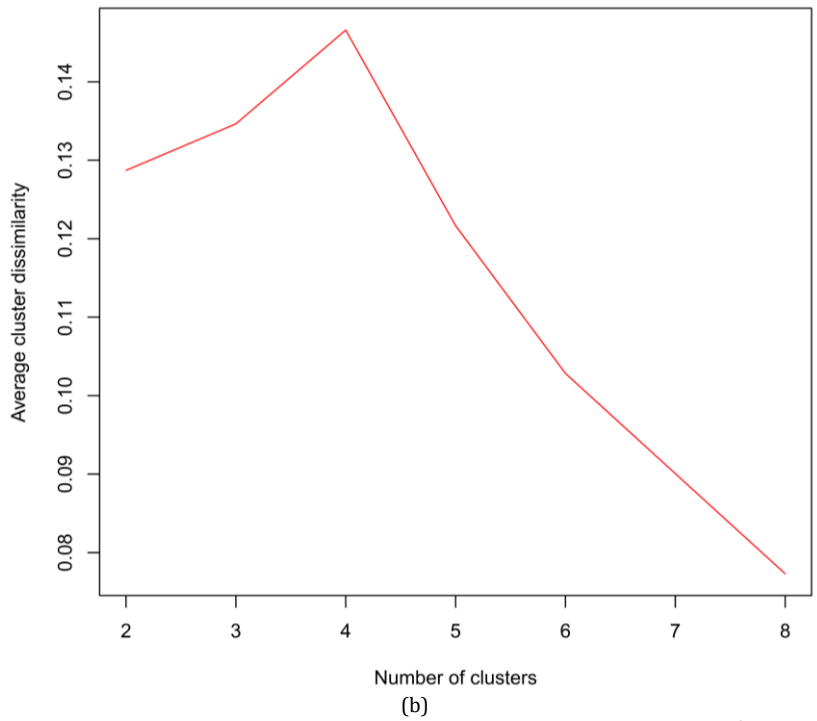


(b)

Fig. 1: Olive data, (a) data pairs, (b) optimal cluster number at $\hat{k} = 2$



(a)



(b)

Fig. 2: Air quality data, (a) data pairs, (b) optimal cluster number at $\hat{k} = 8$

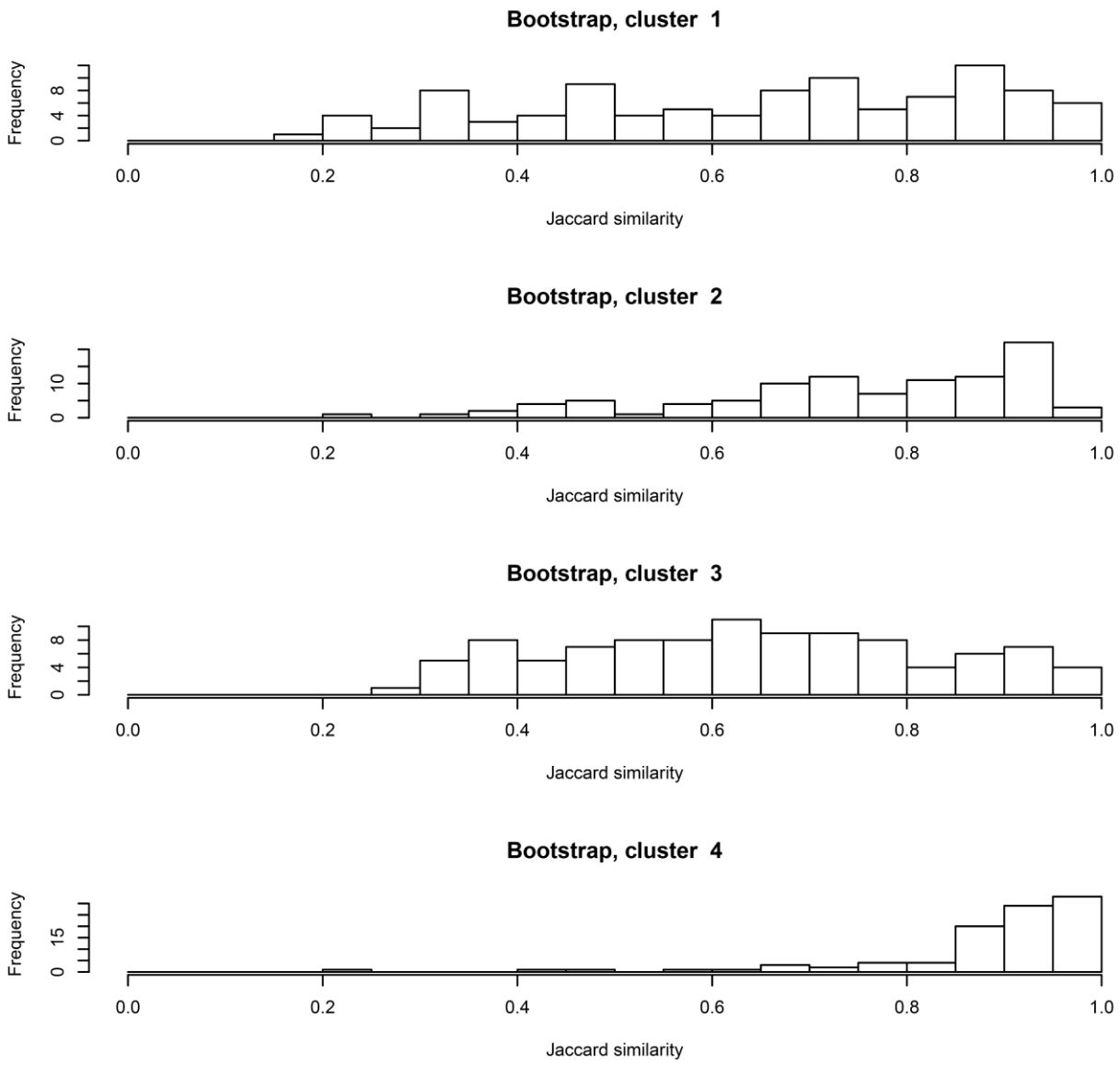


Fig. 3: Jaccard similarity of each cluster in air quality data

5. Conclusion and discussion

The clustering method is a statistical method for data classifications that identify the complex data structure, the clustering techniques could not provide the stability of cluster results. Although the availability of clustering and bootstrap methods, there is a limitation in theoretical results, such as evaluating the consistency of cluster analysis by estimates of differential expression for specific genes (Kerr and Churchill, 2001) or via the distribution of the regression using a one-step version of the cluster bootstrap a presented in Cheng et al. (2013). In this paper, we used the bootstrap method for selecting the optimal number of clusters, which is an alternative to other methods. We study the stability of the clusters by computing the average maximum Jaccard coefficient, the results indicate suitable visualization and properties. In addition, the number of bootstrap samples B does not have to be very large, this algorithm can be used for any data dimensions.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Ben-Hur A, Elisseeff A, and Guyon I (2001). A stability based method for discovering structure in clustered data. In: Russ B Altman, A Keith Dunker, Lawrence Hunter, Kevin Lauderdale, and Teri E Klein (Eds.), Biocomputing 2002: 6-17. World Scientific, Kauai, USA. https://doi.org/10.1142/9789812799623_0002

Cheng G, Yu Z, and Huang JZ (2013). The cluster bootstrap consistency in generalized estimating equations. Journal of Multivariate Analysis, 115: 33-47. https://doi.org/10.1016/j.jmva.2012.09.003

Efron B (1992). Bootstrap methods: Another look at the jackknife. In: Kotz S and Johnson NL (Eds.), Breakthroughs in statistics: 569-593. Springer, New York, USA. https://doi.org/10.1007/978-1-4612-4380-9_41

Fang Y and Wang J (2012). Selection of the number of clusters via the bootstrap method. Computational Statistics and Data Analysis, 56(3): 468-477. https://doi.org/10.1016/j.csda.2011.09.003

Field CA and Welsh AH (2007). Bootstrapping clustered data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(3): 369-390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Hartigan J (1975). Clustering algorithms. Wiley, New York, USA.

Hennig C (2007). Cluster-wise assessment of cluster stability. Computational Statistics and Data Analysis, 52(1): 258-271. https://doi.org/10.1016/j.csda.2006.11.025

Jaki T, Su TL, Kim M, and Van Horn ML (2018). An evaluation of the bootstrap for model validation in mixture models. Communications in Statistics-Simulation and Computation, 47(4): 1028-1038. https://doi.org/10.1080/03610918.2017.1303726 **PMid:30533972 PMCid:PMC6284826**

Jhun M (1990). Bootstrapping k-means clustering. Journal of the Japanese Society of Computational Statistics, 3(1): 1-14. https://doi.org/10.5183/jjscs1988.3.1

Kalantan ZI (2019). Implementing correlation dimension: K-means clustering via correlation dimension. In the 3rd International Conference on Computing, Mathematics and Statistics, Springer, Reims, France: 359-366. https://doi.org/10.1007/978-981-13-7279-7_44

Kaufman L and Rousseeuw PJ (2009). Finding groups in data: An introduction to cluster analysis. John Wiley and Sons, New York, USA.

Kerr MK and Churchill GA (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. Proceedings of the National Academy of Sciences, 98(16): 8961-8965. https://doi.org/10.1073/pnas.161273698 **PMid:11470909 PMCid:PMC55356**

Krzanowski WJ and Lai YT (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics, 44(1): 23-34. https://doi.org/10.2307/2531893

Lange T, Roth V, Braun ML, and Buhmann JM (2004). Stability-based validation of clustering solutions. Neural Computation, 16(6): 1299-1323. https://doi.org/10.1162/089976604773717621 **PMid:15130251**

McCullagh P (2000). Resampling and exchangeable arrays. Bernoulli, 6(2): 285-301. https://doi.org/10.2307/3318577

Naganathan H, Chong WO, and Chen X (2016). Building energy modeling (BEM) using clustering algorithms and semi-supervised machine learning approaches. Automation in Construction, 72: 187-194. https://doi.org/10.1016/j.autcon.2016.08.002

Steinley D (2008). Stability analysis in k-means clustering. British Journal of Mathematical and Statistical Psychology, 61(2): 255-273. https://doi.org/10.1348/000711007X184849 **PMid:17535479**

Sugar CA and James GM (2003). Finding the number of clusters in a dataset: An information-theoretic approach. Journal of the American Statistical Association, 98(463): 750-763. https://doi.org/10.1198/016214503000000666

Tibshirani R, Walther G, and Hastie T (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2): 411-423. https://doi.org/10.1111/1467-9868.00293

Tibshirani RJ and Efron B (1993). An introduction to the bootstrap. Chapman and Hall, New York, USA.

Wang J (2010). Consistent selection of the number of clusters via crossvalidation. Biometrika, 97(4): 893-904. https://doi.org/10.1093/biomet/asq061

Yaqoob I, Hashem IA, Gani A, Mokhtar S, Ahmed E, Anuar NB, and Vasilakos AV (2016). Big data: From beginning to future. International Journal of Information Management, 36(6): 1231-1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009