Journal of Radioanalytical and Nuclear Chemistry (2020) 325:719–724
https://doi.org/10.1007/s10967-020-07186-6

Quantitative methods of standardization in cluster analysis: finding groups in data

André Luiz Nogueira[1] · **Casimiro S. Munita**[2]

Received: 16 December 2019 / Published online: 19 May 2020

Abstract
The aim of this paper is to evaluate the impact of three standardization methods (z-score, $\log_{10}$ and improved min–max) in determining the number of clusters for a dataset of 146 archaeological ceramic fragments in which mass fractions of chemical elements were determined by INAA. The results showed a tendency towards clustering, which did not occur to the non-standardized data. The standardization methods indicated the presence of three groups within the database. Quality evaluation of these clusters, by means of internal validation indexes, showed that the best performance was obtained with the $\log_{10}$ transformation. This transformation also performed well in the calculation of compactness, while the improved min–max showed better performance in terms of separability.

Keywords Archaeometry · Instrumental neutron activation analysis · Clustering analysis · Validation indexes

Introduction

Analytical techniques used to characterize archaeological artifacts by their physical and chemical properties play a major role in archeology [1, 2]. This set of techniques belongs to a multidisciplinary research area called Archaeometry. [3].

Among the many available analytical methods, we can cite the most commonly used by archaeometrists: instrumental neutron activation analysis, INAA [4, 5], X-ray fluorescence, XRF [6], inductively coupled plasma spectrometry, ICP [7, 8], X-ray diffraction, XRD, thermoluminescence dating, TL, and radiocarbon dating [9].

INAA, ICP and XRF, are mainly used to determine the source and origin of raw materials such as rocks and clays used to make archaeological artifacts, information which can inform about several aspects of the history of the people that made them. [10]. It is noteworthy that both INAA and XRF are non-destructive methods as opposed, for example, to ICP. Subsequently, INAA remains the most used and preferred method among archaeometrists [11].

In view of the increasing advance of physicochemical techniques in archaeometric studies, the amount of data generated (results) has increased significantly. For the interpretation of these results, it is necessary to use increasingly sophisticated statistical methods, such as multivariate techniques. In general, multivariate statistical methods allow the evaluation of a set of samples by taking into account the correlations between the variables. These techniques generally consider that each sample analyzed can be represented as a point in multidimensional space, where each dimension (variable) of hyper-space corresponds to axes determined by the physicochemical composition of the samples. The variables can be in different scales. The principal component analysis (PCA) and clustering analysis related to Euclidean distance are scale dependent [12]. If some variables have a large scale or a great variability these variables greatly affect the performance of the cluster analysis algorithms [13]. Standardization is used to ensure that variables make an equal contribution to the calculation of distances [14]. In order to group the samples according to their similarity/dissimilarity, sample groups should be formed according to some statistical criteria. Results can be organized into a $X_{n\times p}$ data matrix where n is the number of samples and p is the number of variables (chemical elements) [15]. This paper

✉ Casimiro S. Munita
camunita@ipen.br

[1] Federal Institute of Sergipe, IFS-SE, Rod. Lourival Batista, s/n, Povoado Carro Quebrado, Lagarto, SE CEP 49400-000, Brazil

[2] Nuclear and Energy Research Institute, IPEN-CNEN/SP, Av. Professor Lineu Prestes, São Paulo, SP CEP 05508-000, Brazil

evaluates the impact of three standardization techniques: z-score, improved min–max, and logarithm, in determining the number of clusters in a data set.

Theoretical aspects

Standardization techniques

Clustering analysis classifies data (samples, elements, observations or characteristic vectors). Data are represented by multidimensional vectors, where each component of the vector corresponds to a characteristic (or variable) of the sample.

In cluster analysis, the variables are not used in a direct way, the performance of the clustering algorithms being sensitive to the magnitude in which the samples are used and uses the calculation of distances between them [16]. Thus, it is of importance to choose among different standardization methods prior to analysis [17]. In this work, three standardization methods were studied: z-score, $\log_{10}$ and improved min–max.

After the application of the z-score transformation, the variables exhibit 0 as the mean value and 1 as standard deviation [18]. In this transformation, information about localisation and scale of the samples is altered. It is defined by the expression:

$$z_1(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \tag{1}$$

where $\bar{x}_j$ and σ_j are mean and standard deviation of the jth column of matrix $X_{n\times p}$. For further details about this transformation see [18].

In the improved min–max transformation [19], the concentrations of the variables remain contained within an [a, b] interval. Its calculation consists of the following steps:

1. Selection of the values of the matrix columns $X_{n\times d}$ that occur more than once (R_{ki}), to build the set $R_k = \{R_{k1}, R_{k2}, \ldots,\}$;
2. Calculation of the mean $m(R_k)$ and standard deviation $s(R_k)$;
3. Calculation of $R_{kA} = m(R_k) + s(R_k)$
4. Application of the transformation:

$$z_2(x_{ij}) = \begin{cases} \frac{x_{ij} - \min(x_j)}{2R_{kA} - 2\min(x_j)}, & x_{ij} \le R_{kA} \\ 0.5 + \frac{x_{ij} - R_{kA}}{\max(x_j) - R_{kA}}, & x_{ij} > R_{kA}. \end{cases} \tag{2}$$

Finally, the log transformation compensates for magnitude differences between variables [20], by using the Eq. (3):

$$z_3 = \log_{10} x_{ij}. \tag{3}$$

Clustering analysis

Determining multidimensional data structure is one of the main challenges in data analysis and pattern recognition. Clustering analysis algorithms have been used for this purpose. However, cluster analysis find and specify groups in data even if groups are not present. For example, Fig. 1 presents a set of two-dimensional data randomly distributed, i.e. there are no clusters present in this data set. So, it makes no sense to apply a clustering algorithm.

Thus, it is appropriate to measure the tendency of clustering or randomness of the data before applying a clustering algorithm [21]. A widely used tendency measure is the Hopkins statistic [22].

The Hopkins statistic calculation basically involves two distances, W_i and U_i. Distances W_i are obtained by calculating the distance between actual data and its nearest neighbors [22]. The U_i distance is calculated in two steps. Firstly, pseudo-data are generated by selecting an actual variable value at random, this being repeated for each of the variables. Secondly, the distance between the pseudo-data and its nearest neighbor (i.e., real data) is calculated. The previously calculated distances are used to determine the value of the Hopkins statistic (H), using the expression (4):

$$H = \frac{\left(\sum_i U_i\right)}{\left(\sum_i U_i + \sum_i W_i\right)} \tag{4}$$

If data are arranged in clusters, W_i distances will be relatively much smaller than U_i, therefore H will be approximately 1, according to [22], and if $H > 0.75$ the data have a tendency to group.

Once the clustering tendency of the data is verified via the Hopkins statistics, the next step is to determine the number of clusters, which is done by using the Ward and Silhouette methods [23]. Only then can the analysis of clusters be performed, with the aim of verifying the existence of similar behaviors between samples, regarding certain variables. The

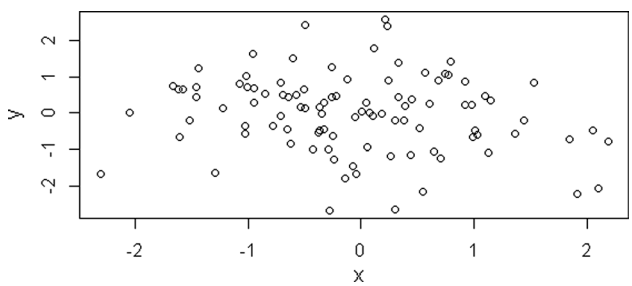

Fig. 1 Randomly distributed data

purpose is to create clusters in which similarity between samples within the same group increases (intracluster similarity), while similarity between groups diminishes (intergroup similarity) [24].

In this work, in order to determine the number of clusters, Ward's hierarchical clustering algorithm was used. Ward's hierarchical clustering method seeks to form partitions $P_n, P_{n-1}, \ldots, P_1$ to minimize information loss caused by cluster mergers [25]. The validation of the partitions obtained by the Ward method was performed using the silhouette method [23].The silhouette method (S) calculates an index, by using the expression (5):

$$S = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i - a_i}{\max\left(a_i, b_i\right)}, \tag{5}$$

where n is the total number of samples, a_i is the average distance between sample i and all other samples in its own cluster, and b_i is the minimum of the average distance between sample i and samples of the other clusters. The range of silhouette index values is $-1 < S < 1$. The maximum value of S determines the number of clusters. Once the number of clusters is determined, the next step is to find the clusters. This is performed with the k-means clustering algorithm. The idea behind the k-means algorithm is to define clusters so that the internal variation of the clusters is minimized [26].

Validation indices

The procedure of evaluating the results of a clustering algorithm is known as cluster validation [27] and can be performed by one of three different procedures: internal, external, and relative [28].

The internal validation technique is based on clustering properties such as compactness and separability and does not require additional database information. While compactness is a measure of proximity between cluster elements, separability (SE) measures the distance between two clusters, according to the expression (6) [28]:

$$\mathrm{SE} = \frac{D_{\max}}{D_{\min}}\sum_{i=1}^{c}\left(\sum_{j=1}^{c}\bar{X}_i - \bar{X}_j\right)^{-1} \tag{6}$$

where c number of clusters; $\bar{X}_i, \bar{X}_j$ centroids of clusters i e j; $||x|| = \sqrt{xx^T}$.

Compactness (CP) is obtained by the expression (7) [29]:

$$\mathrm{CP} = \frac{1}{N}\sum_{i=1}^{c}\left(\sum_{x_k, x_j \in X_i} d\left(x_k, x_j\right)\frac{2}{n_i\left(n_i - 1\right)}\right) \tag{7}$$

where n total number of samples; X_i ith cluster; $d\left(x_k, x_j\right)$: distance between the elements x_k e x_j; n_i : number of elements of the i-th cluster.

Three internal validation indices were used in this study: (1) Dunn, (2) Davies–Bouldin index and (3) Calinski–Harabasz.

The main purpose of the Dunn index is to maximize the distance between clusters while minimizing the distance between samples within the same cluster. The Dunn index [30] identifies sets of clusters that are compact and well separated [31]. The index between samples from the same group is obtained by means of a distance ratio (D) between the samples and their nearest neighbor, and by the separation of the more distant samples. It is calculated by the Eq. (8) [32]:

$$D(U) = \frac{\min\limits_{X_K \in X}\left(\min\limits_{X_l \in X - \{X_k\}}\delta\left(X_k, X_l\right)\right)}{\max\limits_{X_k \in X}\Delta\left(X_k\right)} \tag{8}$$

where $\delta\left(X_k, X_l\right)$ is the distance between clusters X_k and X_l, $\Delta\left(X_k\right)$ represents the distance between the elements (or samples) of cluster X_k and c is the number of clusters in the U partition. The maximum value of D indicates the best performance.

The Davies–Bouldin index, DB, [33] aims to identify clusters that are compact and well separated. The intracluster cohesion is determined by the distance between cluster samples and their centroid. The separation between clusters is based on the distance between the centroids of different clusters [31], this being calculated with the expression (9) [32]:

$$\mathrm{DB}(U) = \frac{1}{c}\sum_{i=1}^{c}\frac{\max\left(\Delta\left(X_i\right) + \Delta\left(X_j\right)\right)}{\left(\delta(\bar{X}_i, \bar{X}_j\right)} \tag{9}$$

where $\bar{X}_i$ and $\bar{X}_j$ represent the centroids of clusters i and j respectively. The lowest the obtained DB value, the best the performance.

The Calinski–Harabasz index, CH, [34], is obtained through the ratio of separation between clusters (as calculated from the distance of each group centroid to the global centroid) and the cohesion of samples (as calculated from the distance of cluster samples to their centroid). The Eq. (10) is used [31]:

$$\mathrm{CH} = \left(\frac{N - c}{c - 1}\right)\frac{\left(\sum_{X_k \in X}|X_k| d\left(\bar{X}_k, \bar{X}\right)\right)}{\left(\sum_{X_k \in X}\sum_{x_i \in X} D(x_i, \bar{X}_k)\right)} \tag{10}$$

where N is the total number of samples, c is the number of clusters, x_i represents the ith sample and $\bar{X}$ is the global

Table 1 Clustering tendency of the Hopkins statistic

	Hopkins statistic value
z-score	0.79
Improved min–max	0.96
Logarithm	0.78
Non-standard data	0.69

centroid. The number that maximizes CH is what determines the best performance.

Database

The study was carried out using a database composed of 146 ceramic samples from 3 archaeological sites, called Água Limpa, Prado and Rezende, the former located in Monte Castelo city, São Paulo state, and the latter located in Predizes and Centralina city, both in Minas Gerais state. The powder samples were obtained drilling to a depth of 2–3 mm using a tungsten carbide rotary file attached to a flexible shaft with variable speed drill. Details of the sampling and archaeological studies carried out were published elsewhere [35, 36].

The INAA technique was used to determine of Na, K, La, Sm, Yb, Lu, Sc, Cr, Fe, Co, Zn, Rb, Cs, Ce, Eu, Hf, Ta, and Th. To determine the mass fractions two reference materials were used as standards: Buffalo River Sediment, NIST-SRM-2704, and Coal Fly Ash, ICHTJ-CTA-FFA-1. Brick Clay, NIST-SRM-679, and Ohio Red Clay were used to check the analytical procedure. Groups of 6 samples and one of each reference material were irradiated in the research reactor IEA-R1m at a thermal neutron flux of about 5×10^{12} cm^{-2} s^{-1} for 8 h. Two measurements series were carried out using a Germanium (hyperpure) detector. Details of the analytical procedure can be found elsewhere [35].

Results and discussion

In order to evaluate the impact of standardization techniques to the database of 146 archaeological ceramic samples, the Hopkins statistic was calculated in the first place to determine the clustering tendency. The Ward and Silhouette methods were then used to establish the number of groups. Groups generated by the k-means clustering algorithm were evaluated by internal validation indices, separability and compactness.

Table 1 presents the results for the Hopkins statistic, using the *z*-score, improved min–max, $\log_{10}$ and non-standard data.

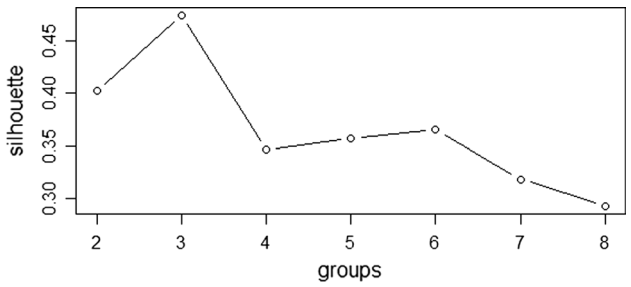

Fig. 2 Silhouette index/*z*-score

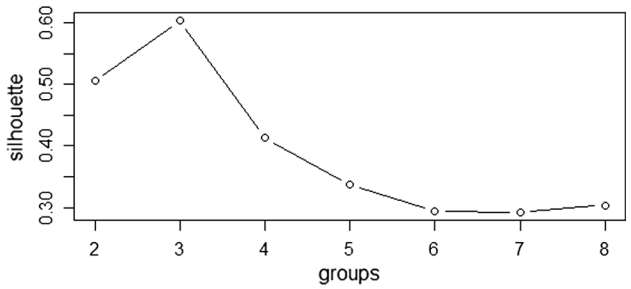

Fig. 3 Silhouette index/logarithm

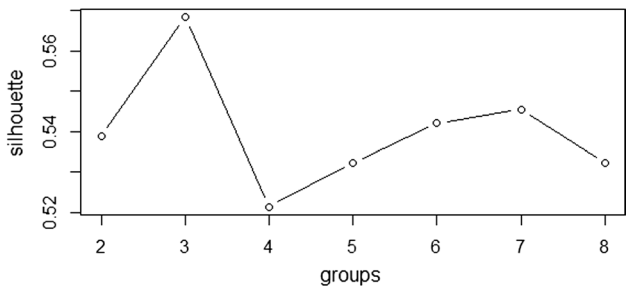

Fig. 4 Silhouette index/improved minimum–maximum

As can be seen in Table 1, the encountered value for the Hopkins statistic was 0.69, when data were not subjected to standardization. By applying the standardization procedures (*z*-score, improved min–max and $\log_{10}$) the Hopkins statistic rendered values higher than 0.75, which, according to current literature, indicates the existence of clusters within the database [22].

After the standardization methods were applied, the number of groups was determined by using the Ward's algorithm, while the validation of those groups was performed through the Silhouette method.

Figures 2, 3 and 4 show the number of groups versus the Silhouette index for the *z*-score, $\log_{10}$ and improved min–max standardization methods.

As derived from Figs. 2, 3 and 4 the maximum value of the Silhouette index occurs when the number of clusters

Table 2 Validation indexes with standardized and non-standardized data

	Validation indices		
	Dunn	Davies–Bouldin	Calinski–Harabasz
z-score	0.29	0.77	116.13
Improved-min–max	0.01	0.58	291.68
Logarithm	0.54	0.54	303.83
Non-standard data	0.07	1.49	137.33

Table 3 Compactness and separability of the standardized and non-standardized samples data

	Compactness	Separability
z-score	2.84	0.36
Improved-min–max	2.03	0.49
Logarithm	1.97	0.40
Non-standard data	7021.05	0.00046

is 3. This shows the existence of three groups within the analyzed database.

The evaluation of the groups encountered with the standardization methods was then performed by means of internal validation, separability and compactness. With these, the objective was to explore both the internal homogeneity of each group and the distance between groups. The k-means clustering algorithm was used that minimize the internal variation between samples in each group.

Table 2 shows the performance of standardization techniques in cluster analysis using the k-means algorithm. The evaluation was performed using the Dunn, Davies–Bouldin and Calinski–Harabasz indices. As can be observed, the standardization technique that presented the best result was the logarithmic one, since the values obtained from the Dunn, Davies–Bouldin and Calinski–Harabasz indexes were 0.54, 0.54 and 303.83, respectively. The values obtained with *z*-score standardization were 0.29, 0.77 and 116.13 and, finally, the values corresponding to the improved minimum–maximum standardization were 0.01, 0.58 and 291.68.

When comparing the results of the standardization techniques with the results of the non-standard data, in Table 2, it is noted that the logarithmic transform with validation indexes 0.54 (Dunn), 0.54 (Davies–Boldin) and 303.83 (Calinski–Harabasz), had a superior performance when compared to the indexes obtained from the non-standard data, 0.07 (Dunn), 1.49 (Davies–Bouldin) and 137.33 (Calinski–Harabasz). These results showed a better validation index after the $\log_{10}$ standardization was executed.

Clustering separability and compactness studies performed for the 146 archaeological ceramic samples are presented in Table 3.

As it can be observed in Table 3, $\log_{10}$ standardization performed better for the calculation of compactness. Thus, a lesser internal distance between group samples was detected. In terms of separability, the best standardization result was the improved min–max, which exhibited the greater distance between groups.

Also, when comparing standardized and non-standardized samples data in Table 3, it is possible to observe a greater separability between the groups making up the former than the groups of the latter. Compactness was less in the standardized samples.

Conclusion

A study of the effects of three standardization methods (*z*-score, $\log_{10}$ and improved min–max) on cluster analysis was performed, using a database of 146 archaeological ceramic samples analyzed by INAA. To reach that goal, the number of clusters, clusterization tendency and clusterization quality were all determined.

On the one hand, the Hopkins statistic [22] applied to the non-standardized data does not allow to conclude that the data tend to cluster. On the other hand, the data transformed by the standardization techniques make it possible to conclude that there is a tendency towards clustering, something that was expected since the data were obtained from three different archaeological sites.

For the determination of the number of clusters within the database, the Ward and Sillhouette methods were studied. Standardization methods showed the presence of three groups.

The quality validation of the obtained clusters was performed using validation indexes, compactness and separability. The validation indexes calculated from data obtained from the logarithmic transform, 0.54 (Dunn), 0.54 (Davies–Boldin) and 303.83, were greater to the indexes calculated using the non-standardized data, 0.07 (Dunn), 1.49 (Davies–Boldin) and 137.33. In terms of separability, the best performance was obtained with the improved minimum–maximum transformation. However, compactness showed better results when the logarithmic transformation was applied.

Finally, it was observed that standardized data performed better than the non-standardized ones, both for determining the degree of separability as well as for the calculation of compactness.

References

1. Hazenfratz R, Munita CS, Glascock MD, Neves EG (2016) Study of exchange networks between two Amazon archaeological sites

by INAA. J Radioanal Nucl Chem 309:195–205. https://doi.org/10.1007/s10967-016-4758-9
2. Martin FF, Di Piazza A, D'oriano C, Carapeza ML, Paonita A, Rotolo SG, Sagnotti L (2017) New insights in the provenance of the obsidian fragments of the Island of Ustica (Palermo, Sicily). Archaeometry 59:435–454. https://doi.org/10.1111/arcm.12270
3. Antonelli F, Ermeti AL, Lazzarini L, Verità M, Raffaelli G (2014) An archaeometric contribution to the characterization of renaissance maiolica from Urbino and a comparison with coeval maiolica from Pesaro (the M arches, central Italy). Archaeometry 56:784–804. https://doi.org/10.1111/arcm.12045
4. Santos JO, Reis MS, Munita CS, Silva JE (2017) Box-Cox transformation on dataset from compositional studies of archaeological potteries. J Radioanal Nucl Chem 311:1427–1433. https://doi.org/10.1007/s10967-016-4987-y
5. Munita CS, Paiva RP, Alves MA, Momose EF, Saiki M (2000) Chemical characterization by INAA of Brazilian ceramics and cultural implications. J Radioanal Nucl Chem 244(3):575–578. https://doi.org/10.1023/A:100675031293
6. Yu KN, Miao JM (1998) Multivariate analysis of the energy dispersive X-ray fluorescence results from blue and white Chinese porcelains. Archaeometry 40:331–339. https://doi.org/10.1111/j.1475-4754.1998.tb00841.x
7. Shackley MS (2008) Archaeological petrology and the archaeometry of lithic materials. Archaeometry 50:194–215. https://doi.org/10.1111/j.1475-4754-2008.00390.x
8. Funtua II, Oladipo MOA, Njinga RL, Jonah SA, Yusuf I, Ahmed YA (2012) Evaluation for the accuracy and tapplicability of instrumental neutron activation analysis of geological materials on Nigeria Nuclear Research Reactor -1 (NIRR-1). Int J Appl Sci Technol 2:286–292
9. Tudela DR, Tatumi SH, Yee M, Brito SL, Morais JL, Morais DD, Piedade SC, Munita CS, Hazenfratz R (2012) TL, OSL and C-14 dating results of the sediments and bricks from mummified nuns' grave. An Acad Bras Cienc 84(2):237–244. https://doi.org/10.1590/S0001-37652012005000031
10. Nyarko BJB, Bredwa-Mensah Y, Serfor-Armah Y, Dampare SB, Akaho EHK, Osae S, Chatt A (2007) Investigation of trace elements in ancient pottery from Jenini, Brong Ahafo region, Ghana by INAA and Compton suppression spectrometry. Nucl Instr Methods Phys Res Sect B 263:196–203. https://doi.org/10.1016/j.nimb.2007.04.086
11. Speakman RJ, Glascock MD (2007) Acknowledging fifty years of neutron activation analysis in archaeology. Archaeometry 49:179–183. https://doi.org/10.1111/j.1475-4754.2007.00294.x
12. Mucha HJ, Bartel HG, Dolata J (2008) Effects of data transformation on cluster analysis of archaeometric data. In: Preisach C, Burkhardt H, Smidt-Thieme L, Decker R (eds) Data analysis, machine learning and applications. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, Heidelberg, pp 681–688. https://doi.org/10.1007/978-3-540-78246-9_80
13. Tanioka K, Yadohisa H (2012) Effect of data standardization on the result of k-means clustering. In: Gaul W, Geyer-Schulz A, Schmidt-Thieme L, Kunze J (eds) Challenges at the interface of data analysis, computer science, and optimization. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, Heidelberg, pp 59–67. https://doi.org/10.1007/978-3-642-24466-7_7
14. Chu CW, Holliday JD, Willett P (2009) Effect of data standardization on chemical clustering and similarity searching. J Chem Inf Model 49(2):155–161. https://doi.org/10.1021/ci800224h
15. Mucha HJ, Bartel HG (2015) Resampling techniques in cluster analysis: is subsampling better than bootstrapping? In: Lausen B, Krolak-Schwerdt S, Böhmer M (eds) Data science, learning by latent structures, and knowledge discovery. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, Heidelberg, pp 113–122. https://doi.org/10.1007/978-3-662-44983-7_10
16. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice Hall, Englewood Cliffs
17. Milligan GW, Cooper MC (1988) A study of standardization of variables in cluster analysis. J Classif 5:181–204. https://doi.org/10.1007/BF01897163
18. Williams WT, Lambert JM (1966) Multivariate methods in plant ecology: V. Similarity analyses and information-analysis. J Ecol. https://doi.org/10.2307/2257960
19. Kabir W, Ahmad MO, Swamy MNS (2016) A new anchored normalization technique for score-level fusion in multimodal biometric systems. In: 2016 IEEE international symposium on circuits and systems (ISCAS), pp 93–96
20. Bartlett MS (1947) The use of transformations. Biometrics 3:39–52. https://doi.org/10.2307/3001536
21. Cross GR, Jain AK (1982) Measurement of clustering tendency. Proc. IFAC Symp Digit Contr. https://doi.org/10.1016/S1474-6670(17)63365-2
22. Lawson RG, Jurs PC (1990) New index for clustering tendency and its application to chemical problems. J Chem Inf Comput Sci 30:36–41. https://doi.org/10.1021/ci00065a010
23. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7
24. Fávero LP, Fávero P (2017) Análise de Dados: Técnicas multivariadas exploratórias com SPSS e Stata. Elsevier, Amsterdam
25. Gan G, Ma C, Wu J (2007) Data clustering: theory, algorithms, and applications, vol 20. Siam, Philadelphia
26. Hartigan JA, Wong MA (1979) Algorithm AS 136: a k-means clustering algorithm. J R Stat Soc Ser C Appl 28:100–108. https://doi.org/10.2307/2346830
27. Halkidi M, Batistakis Y, Vazirgiannis M (2001) On clustering validation techniques. J Intell Inf Syst 17:107–145. https://doi.org/10.1023/A:1012801612483
28. Brun M, Sima C, Hua J, Lowey J, Carroll B, Suh E, Dougherty ER (2007) Model-based evaluation of clustering validation measures. Pattern Recogn 40:807–824. https://doi.org/10.1016/j.patcog.2006.06.026
29. Iam-on N, Garrett S (2010) Linkclue: a matlab package for link-based cluster ensembles. J Stat Softw 36(9):1–36
30. Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybern 3(3):32–57. https://doi.org/10.1080/01969727308546046
31. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recogn 46:243–256. https://doi.org/10.1016/j.patcog.2012.07.021
32. Bolshakova N, Azuaje F (2003) Cluster validation techniques for genome expression data. Signal Process 83:825–833. https://doi.org/10.1016/s0165-1684(02)00475-9
33. Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Patt Anal Machine Intel 2:224–227. https://doi.org/10.1109/TPAM1.1979.4766909
34. Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27. https://doi.org/10.1080/03610927408827101
35. Munita CS, Paiva RP, Alves MA, Oliveira PMS, Momose EF (2001) Major and trace elemento characterization of prehistoric ceramic from Rezende archaeological site. J R N Ch 248(1):93–96. https://doi.org/10.1023/A:1010682209370
36. Munita CS, Paiva RP, Alves MA, Oliveira PMS, Momose EF (2003) Provenance study of archaeolgical ceramic. J Trace Microprobe Tech 21(4):697–706. https://doi.org/10.1081/TMA-120025819