Agência Brasileira do ISBN
ISBN 978-85-99141-08-3

THE EFFECT OF DATA STANDARDIZATION IN CLUSTER ANALYSIS

André L. Nogueira[1] and Casimiro S. Munita[2]

[1] Coordenação de Licenciatura em Física – Instituto Federal de Sergipe
Rod. Lourival Batista, s/n
49400-000
andre.nogueira@ifs.edu.br

[2] Instituto de Pesquisas Energéticas e Nucleares (IPEN / CNEN - SP)
Av. Professor Lineu Prestes 2242
05508-000 São Paulo, SP
camunita@ipen.br

ABSTRACT

The application of multivariate techniques to experimental results requires a responsibility on behalf of the researcher to understand, evaluate and interpret their results, especially the ones that are more complex. In this work, the impact of three standardization techniques on the formation of clusters by the SOM (self-organizing map) neural network were studied. The techniques studied were logarithm ($\log_{10}$), generalized-log and improved min-max. The studies were performed using two databases consisting of 298 and 146 samples and containing the mass fractions of As, Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tn, Hf and Th, determined by neutron activation analysis. The results were evaluated using validation indices.

1. INTRODUCTION

The future advancement of physicochemical techniques means that the quantity of results generated will increase significantly. For results analysis, it is necessary to use more sophisticated methods, such as multivariate techniques. In general, multivariate statistical methods allow one to evaluate a set of samples in terms of the correlations between variables. These techniques consider that each sample can be represented as a point in multidimensional space, where each dimension of hyperspace corresponds to an axis determined by the physicochemical composition of the samples. One of the ways to verify the existence of similar behaviors between the samples in relation to the different variables is by carrying out a clustering analysis. A problem that arises during cluster analysis involves the decision to standardize the samples before calculating the distance measurements, while the existence of several standardization techniques complicates this decision further. The present article proposes to study the effect of three standardization methods on cluster analysis: $\log_{10}$, generalized-log [1], and improved min-max [2]. After applying data standardization, they are submitted to a SOM neural network which aims to gather samples to create clusters, so that there is internal homogeneity in the clusters and external heterogeneity among them [3]. The SOM network is a self-organizing map of unsupervised training: the central idea of the SOM network is competitive learning, since when presenting the sample to the network, the neurons compete with each other and the winner has their weights adjusted to better answer to network stimuli [4]. In addition, there is a process of cooperation between neurons and their neighbors, who also receive adjustments. The characteristics contained in the sample will stimulate a special region of the network associated with a particular group.

The purpose of this paper was to compare, using experimental results, three standardization methods: $\log_{10}$, generalized-log and improved min-max prior clustering. The study was performed using two databases. Named B1 and B2 with 298 and 146 samples, respectively. In both, the mass fractions of As, Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tn, Hf and Th were determined by neutron activation analysis. To evaluate the results obtained in normalization, were used three validation indices Jaccard [5], Rand [6] and Fowlkes-Mallows [7].

2. THEORETICAL ASPECTS

2.1 Standardization Techniques

In many cluster analysis applications, raw data, or actual measurements, are not used directly unless a probabilistic model for pattern generation is available [8]. Thus, there is a need to prepare the data for cluster analysis through a transformation aimed at standardizing the data.

2.1.1. Log_{10}

Several authors studied whether the chemical elements for geological samples are distributed normally or lognormally [9]. The results showed that composition data are distributed lognormally by two reasons. First, was observed that more often that for trace elements, the data appear to be more normally distributed when treated as logarithms of the measured concentrations. The second reason is that transformation of concentration data into logarithms compensates for the differences in the magnitudes between the major elements, such as K, Fe, and the trace elements, such as rare earth elements. The transformation to logarithms effects a quasi-standardization that is convenient and facilitates the use of cluster analysis and other multivariate methods. The $\log_{10}$ standardization to improve symmetry of the datasets. A common strategy in the analysis of datasets chemical compositions, with measurements strictly positive, is the $\log_{10}$ standardization, which emphasizes clusters without introducing spurious effects [10].

2.1.2. Generalized-log

Generalized-log standardization is based on the q-logarithm function which is a generalization of logarithmic function and is used as the intermediate domain between the log and linear domain [1], the generalized-log function is defined as:

$$\bar{x}^k = \exp_q\left(\frac{\log_q(x^k)-\frac{1}{N}\sum_{\mathrm{i}=1}^{N-1}\log_q x^i}{1+\mathrm{q}\frac{1}{N}\sum_{\mathrm{i}=1}^{N-1}\log_q x^i}\right) \tag{1}$$

where $\bar{x}^k$is the standardized data, N is the number of samples and x^k is the raw data and the q-logarithm is defined as follows:

$$\log_q(x) = \begin{cases}\dfrac{x^q-1}{q}, q \neq 0\\ \log(x), q = 0\end{cases} \tag{2}$$

and the inverse of the q-logarithm, the q-exponential is defined as the following:

$$\exp_q(x) = \begin{cases}(1+qx)^{\frac{1}{q}}, q \neq 0\\ \exp(x), q = 0.\end{cases} \tag{3}$$

2.1.3. Improved Min-Max

In the improved min-max standardization [2] a set of R$_k$ is constructed for each column that is composed of values that occur more than once. The mean and standard deviation of R$_k$ are summed to obtain R$_{KA}$. At the end, the improved min-max standardization is applied using the expression:

$$F(x) = \begin{cases}\dfrac{\frac{x_k-\min(x_k)}{2(R_{KA}-\min(x_k))},x_k\leq R_{KA}}{0.5+\frac{x_k-R_{KA}}{\max(x_k)-R_{KA}},x_k>R_{KA}}\end{cases} \tag{4}$$

where: $R_{KA} = R_{Kavg} + R_{Kstd}$, $R_{Kavg} = mean(R_{KA}), R_{Kstd} = std(R_{Kstd})$ (standard deviation).

2.2 Validation Indexes

The indices described below are used to evaluate the effect of standardization techniques in cluster analysis, comparing the results obtained from the SOM neural network with predefined information. The existence of two partitions is assumed, one obtained by the SOM neural network and the other with additional information about the base [11].

2.2.1 Jaccard

The Jaccard index also named coefficient of similarity, is a well known measure of similarity between clusters described by the presence or absence of samples, used in cluster analysis. It counts the number of pairs of samples belonging to the same group in partitions A and B, and the number of pairs of samples that belong to the same group on at least one of the partitions. The Jaccard index is give by [5]:

$$R = \frac{a}{a+b+c}. \tag{5}$$

where a is the number of pairs of samples belonging to the same cluster, in A and B; b is the number of pairs of samples belonging to different groupings in A, but same group in B; c is the number of samples belonging to the same group in A, but different clusters in B.

2.2.2 Rand

The Rand index is a statistical measure of the proportion of pairs of samples belonging to the same or different clusters in both partitions and is defined by [6]:

$$R = \frac{a+d}{a+b+c+d}. \tag{6}$$

where a, b, and c are the same as the previous index, and d is the number of samples belonging to different clusters in A and B.

2.2.3 Fowlkes-Mallows

The Fowlkes-Mallows index [7] is a geometric mean of the proportion of pairs of samples belonging to the same group in both partitions. Let A and B be two partitions, with the same number of samples. Let m = [m_{ij}], i, j = 1, ..., k, where m_{ij} is the number of samples in common with the ith cluster of A and the jth cluster of B. The similarity measure proposed by [7]:

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}} \tag{7}$$

where

$$P_k = \sum_{i=1}^{k} m_i^2 - n, Q_k = \sum_{j=1}^{k} m_j^2 - n\,, T_k = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}^2 - n, m_i = \sum_{j=1}^{k} m_{ij},$$

$$m_j = \sum_{i=1}^{k} m_{ij}, n = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{ij}.$$

3. RESULTS

The tests were performed using two databases, one containing 298 samples and the other with 146 samples, named B1 and B2, respectively, in which the mass fractions of As, Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tn, Hf and Th, were obtained using the method of neutron activation analysis. Figures 1 and 2 show the scatter plot of bases B1 and B2, respectively. The scatter plot was obtained after applying the principal component analysis (PCA) in the raw data. The PCA is a transformation of correlated variables to pairwise uncorrelated variables in the lower dimensional space [12]. It is often used to display structure in the data. This article was used the first two transformed variable PC1 and PC2 to generate the scatter plot.The scatter plot was obtained after applying the PCA to the raw data.

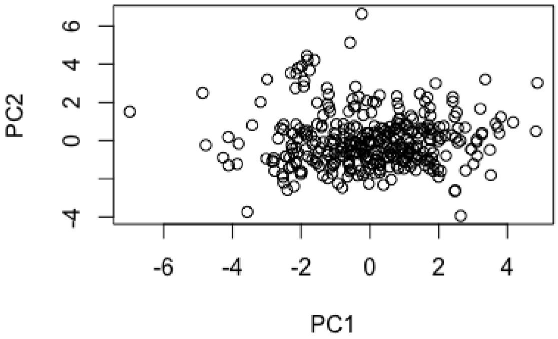

Figure 1: Scatter plot of the principal components, base B1

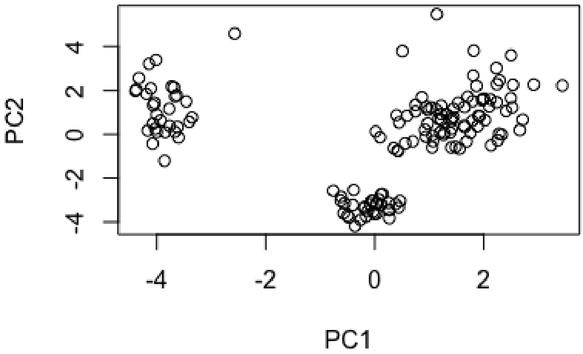

Figure 2: Scatter plot of the principal components, base B2

3.1 Neural Networks

Neural networks are made up of basic processing units called neurons. The figure below shows a neuron of an artificial neural network [4].

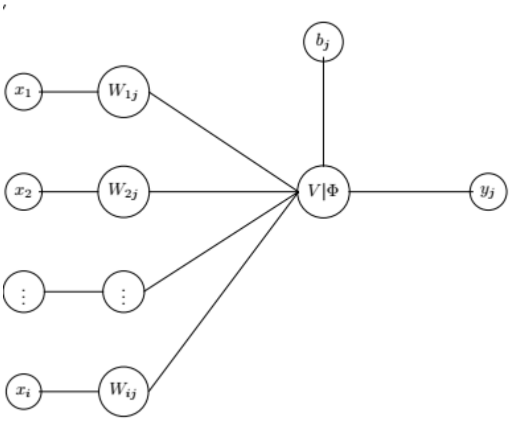

Figure 3: Artificial neuron

The artificial neuron, similar to the natural neuron, receives input signals and returns a single output signal, which may be the output of the network or the input signal to one or more neurons in the back layer.

The inputs of a neural network x_1, x_2, ..., x_i are multiplied by the corresponding synaptic weights W_{1j}, W_{2j}, ..., W_{ij} generating the following weighted sum:

$$V = \sum_{k=1}^{i} W_{kj} x_k \qquad (8)$$

This function is called the activation function. The weighted sum is presented to a transfer function whose purpose is to avoid the progressive addition of output values [4].

The artificial neural model can also include an input bias (b_j) in order to increase the degree of freedom of the activation function [4].

An artificial neural network is a combination of neurons, their connections, and the algorithm used in training. The neural network has two stages of processing: learning and network application. In learning, the adjustment of weights occurs in response to data presented to the network. In the application of the network, one has the way in which the network responds to the data without there being changes in the weights.

3.1.1. SOM Network

The SOM network is a self-organizing map model of unsupervised training [13,4]. In this structure, the neurons are arranged in a normally two-dimensional grid, which can be square, rectangular, triangular, and so on. What characterizes the SOM network is the formation of a topological map of input data patterns in which the locations of the neurons indicate the characteristics of the input data.

The central idea of this model is competitive learning, because when presenting an input sample to the network, the neurons compete with each other and the winner has their weights adjusted to better respond to the stimulus presented to the network. In addition, there is a process of cooperation between neurons and their neighbors, who also receive adjustments. The characteristics contained in the input sample will stimulate some special region of the network and the sample is assigned to the corresponding group.

The evaluation of the impact that standardization causes in clustering analysis using self-organizing maps was performed through the validation indexes of Jaccard [5], Rand [6] and Fowlkes-Mallows [7] the higher the index, the better result obtained by the SOM neural network.

The results obtained after cluster analysis of the transformed data corresponding to databases B1 and B2 are shown in Table 1.

Table 1: Validation indices obtained after the application of the standardization techniques in the B1 and B2 databases

	Jaccard		**Rand**		**Fowlkes-Mallows**	
Standardization	**B1**	**B2**	**B1**	**B2**	**B1**	**B2**
$\log_{10}$	0.54	1	0.57	1	0.71	1
Generalized-log	0.33	1	0.62	1	0.50	1
Improved Min-max	0.64	0.57	0.77	0.71	0.79	0.75

Table 1 shows that in the tests performed with B1, the standardization technique that presented better performance was the improved min-max, since the values obtained from the Jaccard, Rand and Fowlkes-Mallows indices were, respectively, 0.64 , 0.77 and 0.79. The values obtained with $\log_{10}$ standardization were 0.54, 0.57, and 0.71 and finally the values corresponding to generalized-log standardization were 0.33, 0.62 and 0.50.

In B2 database, the improved min-max standardization presented the worst performance, since the values of the validation indexes of Jaccard, Rand and Fowlkes-Mallows were 0.57, 0.71 and 0.75. On the other hand, both $\log_{10}$ and generalized-log standardization presented all values of validation indices equal to 1.

This tests indicates that, when there is overlap between the clusters, as in the case of base B1 (Figure 1), the standardization technique with the best performance is improved min-max. On the other hand, if the clusters do not present overlap, as in database B2 (Figure 2), $\log_{10}$ and generalized-log techniques perform better.

4. CONCLUSION

This work presented the study of three standardization methods in cluster analysis: $\log_{10}$, generalized-log and improved min-max. The study was made using two databases of 298 (B1) and 146 (B2) samples, in which were determined mass fractions of As, Na, K, La, Yb, Lu, U, Sc, Cr, Fe, Cs, Eu, Tn, Hf and Th, by neutron activation analysis. After applying data standardization, they are submitted to a SOM neural network which aims to gather samples to create clusters. To evaluate the results we used the validation indices Jaccard, Rand and Fowlkes-Mallows. The preliminary study using the two datasets showed that when there is overlap between clusters (Figure 1), the standardization that presented better performance was the improved minimum maximum. on the other hand when the dataset has a well-defined clusters structure (Figure 2), the standardizations that presented the best performance were $\log_{10}$ and generalized-log. In the future to validate the previous hypothesis, an extensive comparative study with the standardization methods will be performed on artificial bases, varying the size of the data, the number of clusters and the distance between the clusters.

REFERENCES

1. H. F. Pardede; K. Shinoda, "Generalized-log spectral mean normalization for speech recognition", *Twelfth Annual Conference of the International Speech Communication Association* (2011).
2. W. Kabir; M. O. Ahamad; M. N. S. Swamy, "A new anchored normalization technique for score-level fusion in multimodal biometrie systems", *2016 IEEE International Symposium on Circuits and Systems (ISCAS) IEEE* pp. 93-96 (2016).
3. L. P. Fávero; P. Fávelo,"*Análise de Dados: Técnicas multivariadas exploratórias com SPSS e STATA*", Elsevier Brasil (2017).
4. S. Haykin, "*Neural network: A comprehensive foundation"* Prentice Hall PTR (2004).
5. P. Jaccard, "Nouvelles recherches sur la distribution florale," *Bull. Soc. Vaud. Sci. Nat.*, Vol. 44, pp. 223-270 (1908).
6. W. M. Rand, "Objective criteria for the evaluation of clustering methods". Journal of the American Statistical association, **Vol. 66**, n. 336, pp. 846-850 (1971).
7. E. B. Fowlkes; C. L. Mallows, "A method for comparing two hierarchical clusterings,"*Journal of the American statistical association*, Vol. 78, n. 383, pp. 553-569 (1983).
8. A. K. Jain; R. C. Dubes. "*Algorithms for clustering data* ", Englewood Clffs: Prentice hall (1988)
9. M. D. GLASCOCK, "Characterization of archaeological ceramics at MURR by neutron activation analysis and multivariate statistics". *Chemical characterization of ceramic pastes in archaeology,* **v. 7**, pp. 11e26, (1992).

10. M. J. BAXTER, “Notes on Quantitative archaeology and R”. Unpublished work available through http://www.mikemetrics.com/download/i/mark_dl/u/4011023365/4623239688/Book.pdf , (2015).
11. M. Brun, “Model-based evaluation of clustering validation measures”, Pattern recognition, **Vol. 40**, n. 3, pp. 807-824 (2007).
12. G. M. ARNOLD; A. J. COLLINS, “Interpretation of transformed axes in multivariate analysis”, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **v. 42**, n. 2, pp. 381-400, (1993).
13. J. Vesanto; Alhoniemi, “Clustering of the self-organizing map,” *IEEE Transactions on neural networks*, **Vol 11,** n. 3, pp. 586-600 (2000).