Permutation tests for cluster analysis

Rosa Arboretti Giancristofaro * Riccardo Ceccato † Marta Disegna ‡
Luca Pegoraro § Luigi Salmaso ¶

Abstract
Cluster analysis is a powerful, versatile tool used in several fields to classify objects according to a set of observed characteristics. This analysis commonly involves some critical decisions, such as choosing the number of clusters to consider. To help practitioners undertake this specific challenge, we propose a permutation-based approach which makes it possible to compute a ranking of C different partitions into K_c, $c = 1, \ldots, C$ clusters. In particular, this procedure avoids choosing a single clustering quality index for the choice of optimal number of clusters, and bases the decision on multiple indices. A case study presenting an example of successful application of the aforementioned approach is considered.

Key Words: cluster analysis; permutation;

1. Introduction

Cluster analysis refers to the use of a large set of statistical techniques to appropriately group units in an available sample. In particular, Tuena et al. (2020) [1] define cluster analysis as "the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)". The key is therefore to assign units in a data set to different clusters so that the similarity within groups and the dissimilarity between groups are maximized.

Cluster analysis has gained popularity in many different research fields, such as medicine, social sciences and engineering (see Figure 1), and many different clustering techniques have been introduced in the literature [2]. These are traditionally divided into hierarchical (e.g. single link [3]) and partitional algorithms (e.g. k-means[4,5]).

Jain (2010) [2] stresses that when performing a cluster analysis, the practitioner is faced with several different challenges, such as choice of clustering algorithm and number of clusters. For the latter, several different clustering quality indices have been introduced, such as the Dunn index [6], the Davies-Bouldin index [7] and the Calinski-Harabasz index [8], but this requires a further choice to be made, i.e. which quality index to use.

In this paper, we propose an approach which avoids having to make this decision by considering multiple indices at the same time and combining indications provided by each of them using the nonparametric combination (NPC) methodology [9]. Section 2 is devoted to the description of this approach. In section 3 we present a case study in which the procedure is applied, and in section 4 we provide our conclusions.

2. Method

Let us suppose we need to choose between two possible numbers of clusters, namely K_1 and K_2, using the k-means algorithm. The first part of the proposed approach requires us

*Civil, Environmental and Architectural Engineering, University of Padova, Padova, Italy
†Department of Management and Engineering, University of Padova, Vicenza, Italy
‡Department of Management and Engineering, University of Padova, Vicenza, Italy
§Department of Management and Engineering, University of Padova, Vicenza, Italy
¶Department of Management and Engineering, University of Padova, Vicenza, Italy

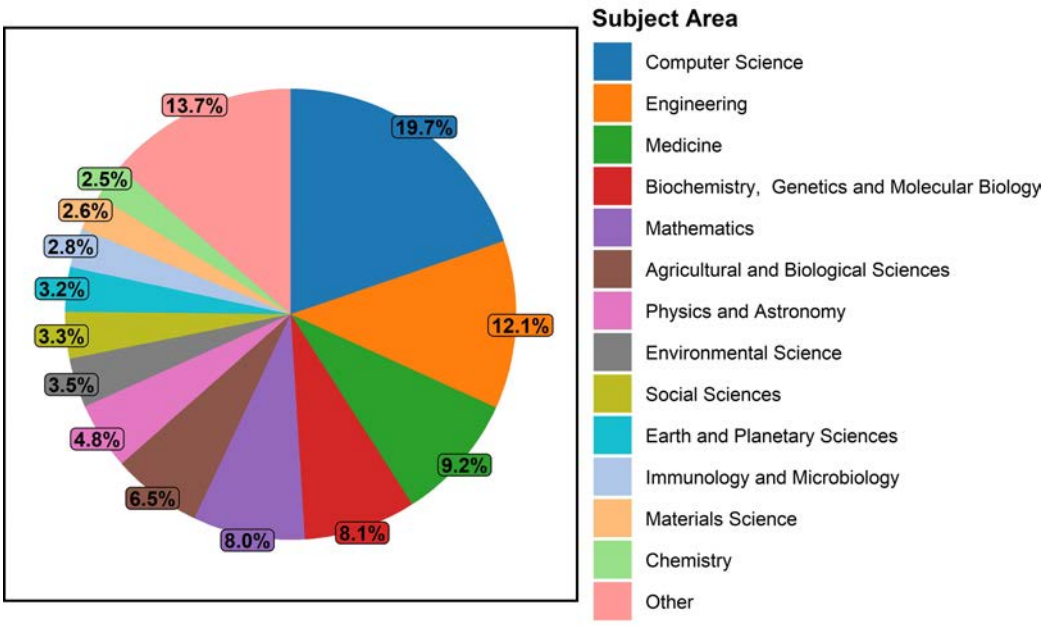

Figure 1: Documents distribution by subject area. Source: Scopus. Query: ”cluster analysis” AND ”clustering”

to:

1. For $k = K_1$,K_2:
   (a) For $l = 1, \ldots, N$:
      i. Randomly choose the k initial cluster centers.
      ii. Apply the k-means algorithm and identify the optimal partition P_k into k clusters.
      iii. Compute J different clustering quality indices to evaluate the partition.

Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be the data sets, made up of N rows and J columns, containing the obtained values of the clustering quality indices considering K_1 and K_2 clusters, respectively. For each $j = 1, \ldots, J$ we now need to test if $\mathbf{X}_{1j} \overset{d}{>} (\overset{d}{<}) \mathbf{X}_{2j}$, where the direction of the comparison is determined by the nature of the index (i.e. if it needs to be maximized or minimized), and combine the indications provided by each individual test. To this end, the nonparametric combination methodology [9] can be applied.

Given that $\mathbf{X}_1$ and $\mathbf{X}_2$ are paired samples, according to this methodology we need to:

1. Compute the matrix of differences $\mathbf{Y} = \mathbf{X}_1 - \mathbf{X}_2$.
2. Compute the test statistics $T_j = \sum_{i=1}^{N} Y_{ij}$, $j = 1, \ldots, J$.
3. For $b = 1, \ldots, B$:
   (a) Consider a random permutation of the N rows of $\mathbf{Y}$ by randomly generating N signs S_i. The permuted data set $\mathbf{Y}^*$ is achieved where $Y^*_{ij} = Y_{ij} \times S_i$.
   (b) Compute the test statistics $T^*_j = \sum_{i=1}^{N} Y^*_{ij}$, $j = 1, \ldots, J$.
4. Compute a partial p-value and its simulated permutation distribution for each test statistic T_j, $j = 1, \ldots, J$.
5. Combine p-values by means of an appropriate combining function (e.g. Fisher's omnibus combining function [10]). A second order combined test is therefore achieved.
6. Compute the global p-value for the second order combined test to combine indications provided by each clustering quality index.

Further details about the algorithm used in the nonparametric combination methodology and the calculation of the p-values and their permutation distributions are provided by Pesarin and Salmaso (2010) [9].

The global p-value achieved using this procedure can be used to choose between the partition P_{K_1} and P_{K_2} into K_1 and K_2 clusters, respectively.

However, we commonly wish to consider more than two possible numbers of clusters, let us say for example C. In this case, it is possible to integrate the nomparametric combination methodology with the ranking procedure by Arboretti et al. (2014) [11]. NPC is initially used to perform all possible pairwise comparisons between partitions P_{K_c} into K_c clusters in terms of J indices. Then, appropriate computations are performed to achieve a ranking of the partitions according to the adjusted (for multiplicity) p-values obtained through the $\frac{C \times (C-1)}{2}$ comparisons. The partition P_{K^*} into K^* clusters to which the first ranking position is assigned should therefore be selected.

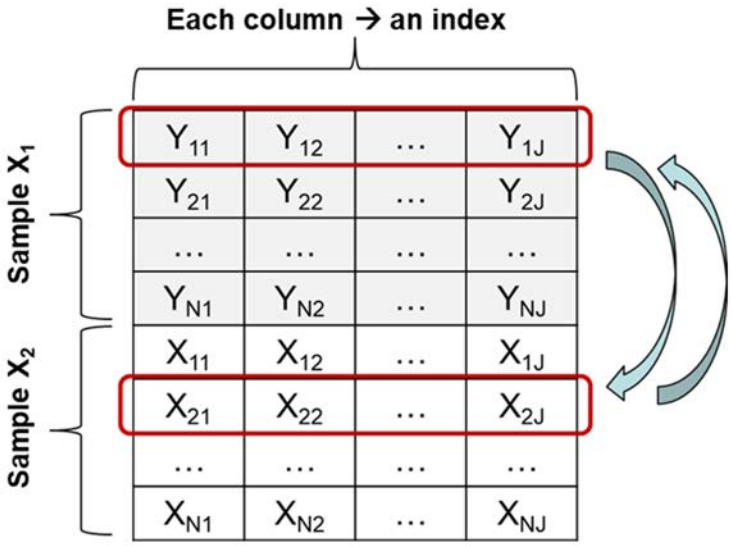

Figure 2: Illustration of a permutation of the pooled sample.

3. Case study

This study considers an industrial problem involving 30 detergents which needed to be grouped in terms of their performances measured by 8 different variables.

After standardization, we decided to use the k-means algorithm and apply the aforementioned procedure to choose the optimal number of clusters from the sequence $(2, \ldots, 10)$ while considering three different clustering quality indices, namely the Dunn index [6], the Davies-Bouldin index [7] and the Calinski-Harabasz index [8]. The number of permutations was set at 2000, Fisher's combining function was adopted to combine p-values, and the Bonferroni-Holm-Shaffer multiplicity correction [12] was applied to p-values before computing rankings.

The results of this analysis are presented in Tables 1 and 2. As can be seen, a number of clusters equal to 2 appears to be the best choice given that it provided the highest values for the Dunn and Calinski-Harabasz indices and the third lowest value for the Davies-Bouldin index. Indeed, the NPC-based approach proposed in this paper showed that overall, partition into 2 clusters appears to significantly outperform the others, and assigned a ranking position of 1 to said partition.

To sum up, adoption of the proposed approach allowed us to successfully select the optimal number of clusters by combining indications provided by multiple clustering quality indices.

4. Conclusions

Carrying out a cluster analysis means practitioners have to address several different challenges, such as the choice of number of clusters. This paper proposed a permutation-based procedure that takes advantage of the nonparametric combination methodology [9] and the

Table 1: Clustering quality indices - average values.

No. of clusters	Dunn	Davies-Bouldin	Calinski-Harabasz
2	0.23	0.78	38.78
3	0.16	0.88	32.20
4	0.14	0.93	30.04
5	0.15	0.90	28.75
6	0.15	0.87	28.70
7	0.17	0.83	29.94
8	0.18	0.79	31.41
9	0.18	0.75	32.04
10	0.18	0.71	32.68

Table 2: Ranking.

No. of clusters	2	3	4	5	6	7	8	9	10
Ranking	1	5	7	9	7	5	4	3	2

ranking procedure by Arboretti et al. (2014) [11] to help make this specific choice. In particular, this approach allows practitioners to base their decision on multiple clustering quality indices and combine indications provided by each of them. Ranking is provided and the top rank is assigned to the optimal number of clusters (i.e. to the partition into K^* which significantly outperforms the other partitions in terms of the considered clustering quality indices).

A case study was considered which allowed us to show successful application of such a procedure, and it appears to be a promising method to support practitioners in overcoming one of the main challenges posed by cluster analysis.

References

[1] C. Tuena, M. Chiappini, C. Repetto, and G. Riva, "Artificial intelligence in clinical psychology," in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2020.

[2] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern recognition letters*, vol. 31, no. 8, pp. 651–666, 2010.

[3] C. F. Olson, "Parallel algorithms for hierarchical clustering," *Parallel computing*, vol. 21, no. 8, pp. 1313–1325, 1995.

[4] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[5] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[6] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[7] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[8] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[9] F. Pesarin and L. Salmaso, *Permutation tests for complex data: theory, applications and software*. Wiley, 2010.

[10] R. A. Fisher, "Statistical methods for research workers," in *Breakthroughs in statistics*, pp. 66–70, Springer, 1992.

[11] R. Arboretti, S. Bonnini, L. Corain, and L. Salmaso, "A permutation approach for ranking of multivariate populations," *Journal of Multivariate Analysis*, vol. 132, pp. 39–57, 2014.

[12] J. P. Shaffer, "Modified sequentially rejective multiple test procedures," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 826–831, 1986.