Finding Prototypes Through a Two-Step Fuzzy Approach

Mario Fordellone and Francesco Palumbo

Abstract Usually, the aim of cluster analysis is to build prototypes, i.e., typologies of units that present similar characteristics. In this paper, an alternative approach based on consensus clustering between two different clustering methods is proposed to obtain homogeneous prototypes. The clustering methods used are fuzzy c-means (that minimizes the objective function with respect to centers of the groups) and archetypal analysis (that minimizes the objective function with respect to extremes of the groups). The consensus clustering is used to assess the correspondence between the clustering solutions obtained and to find the prototypes as a compromise between the two clustering methods.

1   Introduction

According to Rosch [24, 25], prototypes are those elements that represent a category better than others. The degree of representativeness can be measured using a distance function to a salient entity of the category, i.e., a *prototype* [13]. Prototypes can be observed or unobserved (abstract), and they can be represented by a single value or by interval-valued variables. Several numerical techniques to find prototypes in a given multivariate dataset have been proposed in the statistical literature based on different criteria. The most widely used techniques are generally based on non-hierarchical clustering algorithms [11, 18, 27], although many other approaches can be adopted [19, 20, 22].

M. Fordellone (✉)
Sapienza University of Rome, Rome, Italy
e-mail: mario.fordellone@uniroma1.it

F. Palumbo
Federico II University of Naples, Napoli, Italy
e-mail: fpalumbo@unina.it

This paper proposes a two-step procedure based on consensus clustering (CC) to find prototypes within multidimensional data. The first step aims to define two partitions of the $N \times J$ data matrix $\mathbf{X}$ in K groups, where K is assumed to be known; the second step aims to find the correspondence between these two partitions and define the partition solution as the compromise between the two partitions [1, 7, 17, 29].

This work aims to represent a new approach to finding a set of K prototypes through CC to pair the partitions obtained via two different partitioning methods: fuzzy c-means (FCM) [5] and archetypal analysis (AA) [2, 8]. The former seeks K homogeneous groups vis-à-vis their barycenters, while the latter identifies a set of K extreme points, called archetypes, and creates a group around each archetype. Formally, AA minimizes the sum of distances between each point and a set of K archetypes, as defined by a convex combination of extreme points. K is given or can be selected by running the algorithm for different values of K and choosing the desired value according to the most commonly used methods. Such methods are generally based on graphic displays.

The paper is structured as follows: Sect. 2 presents a brief background on CC, FCM, and AA; Sect. 3 then demonstrates a simulation study in order to examine the reliability of the method. Finally, Sect. 4 reports on the results of an application using real data.

2 Methodology

Given a multivariate dataset and two or more partitioning criteria, consensus analysis aims to find a compromise in the set of the partitions [17]. Consensus clustering aims for the same goal among two or more partitions obtained via cluster analysis approaches. Taking into account the final aim of the analysis, the researcher chooses a consensus measure. The present proposal considers two fuzzy approaches as partitioning methods that optimize two different criteria, given the number of groups K: fuzzy c-means (FCM) and archetypal analysis (AA). Proposed by Bezdek et al. [6], the former aims to maximize the homogeneity within the K groups, while the latter, proposed by Cutler and Breiman [8], identifies the K groups with respect to a set of K extreme points, called archetypes, and aims to maximize the heterogeneity among the K groups.

FCM and AA can be defined in terms of a factorization problem of the data matrix $\mathbf{X}$ under different constraints. Formally, let $\mathbf{X}$ be a generic $N \times J$ data matrix, and let $\mathbf{P}$ be an unknown $K \times J$ prototypes matrix; FCM and AA are based on the solution of the following non-negative factorization problem [4]:

$$f(\mathbf{Y}, \mathbf{P}) = \arg \min_{\mathbf{Y}, \mathbf{P}} \|\mathbf{X} - \mathbf{Y}\mathbf{P}\|_2^2, \tag{1}$$

where the notation $\|.\|_2^2$ denotes the quadratic norm, $\mathbf{Y}$ is the generic $N \times K$ memberships matrix, and $\mathbf{P}$ refers to the matrix of the centers in the FCM context

and to the archetypes matrix in the AA context. In order to avoid any confusion in the remainder of this paper, the $\mathbf{Y}$ matrix will be referred to as $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$, and the $\mathbf{P}$ matrix as $\mathbf{C}$ and $\mathbf{A}$ when we refer to the FCM or to the AA, respectively. The generic elements y_{ik} vary in [0,1] and represent the membership degree of the generic unit x_i' to the generic element p_k.

2.1 Fuzzy c-Means

Both the fuzzy c-means clustering method [5, 6] and the traditional k-means method minimize the sum of the weighted squared distances between the N units and the K centers. Formally, given an $N \times J$ data matrix $\mathbf{X}$ FCM minimizes the objective function shown in Eq. (2).

$$f(\Gamma, \mathbf{C}) = \|\mathbf{X} - \Gamma\mathbf{C}\|_2^2, \tag{2}$$

where Γ represents the memberships matrix with elements γ_{ik}. The function in Eq. (2) is minimized under the constraints $\sum_{k=1}^{K} \gamma_{ik} = 1$ and $\gamma_{ik} \geq 0$. The elements γ_{ik} of the Γ matrix are defined according to Eq. (3), while the $\mathbf{C}$ matrix is defined according to Eq. (4).

$$\gamma_{ik} = \left(\sum_{k'=1}^{K} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_k\|_2}{\|\mathbf{x}_i - \mathbf{c}_{k'}\|_2} \right)^{\frac{2}{m-1}} \right)^{-1}, \tag{3}$$

$$\mathbf{C} = (\Gamma^T \Gamma)^{-1} \Gamma^T \mathbf{X} \tag{4}$$

Note that m is the fuzzifier parameter, commonly set to 2 [6]. Including Eq. (4) in Eq. (2), the objective function becomes:

$$f(\Gamma) = \left\| \mathbf{X} - \Gamma(\Gamma^T \Gamma)^{-1} \Gamma^T \mathbf{X} \right\|_2^2 \tag{5}$$

Then, once the number of groups K is fixed, the FCM algorithm runs through the following steps [6, 9, 30]:

1. Randomly initialize the cluster centers $\mathbf{C}^{(t)}$ and set $t = 0$;
2. Calculate γ_{ik} using Eq. (3);
3. Calculate $\mathbf{C}^{(t+1)}$ using Eq. (4);
4. If $\left\| \mathbf{C}^{(t)} - \mathbf{C}^{(t+1)} \right\|_2^2 \leq \epsilon$, go to Step 5; otherwise, $\mathbf{C}^{(t)} = \mathbf{C}^{(t+1)}$; set $t = t + 1$ and go to Step 2;
5. Print centers matrix $\mathbf{C}$ and membership matrix Γ;
6. Stop.

2.2 Archetypal Analysis

The term *archetype* is used in the literature to define different meanings. In the prototyping approach, the challenge is to find a few points (archetypes), not necessarily observed, in a set of multivariate observations such that all the data can be well represented as convex combinations of the archetypes.

Formally, given an $N \times J$ data matrix $\mathbf{X}$, archetypal analysis [8, 10, 12] finds a set of archetypes $\{a_1, \ldots, a_K\}$ that are linear combinations of the data points, as shown in Eq. (6).

$$\mathbf{A} = \mathbf{BX}, \tag{6}$$

where $\mathbf{B}$ is the $K \times N$ coefficients matrix with $\sum_{k=1}^{K} \beta_{ki} = 1$ and $\beta_{ki} \geq 0$, such that the archetypes resemble the data as a convex mixture. For a given choice of archetypes, AA minimizes the objective function shown in Eq. (7).

$$f(\Delta, \mathbf{A}) = \|\mathbf{X} - \Delta \mathbf{A}\|_2^2, \tag{7}$$

under the constraints $\sum_{k=1}^{K} \delta_{ki} = 1$ and $\delta_{ki} \geq 0$. Including Eq. (6) in Eq. (7), the objective function becomes:

$$f(\Delta, \mathbf{B}) = \|\mathbf{X} - \Delta \mathbf{BX}\|_2^2. \tag{8}$$

Once the number of groups K is fixed, the AA algorithm then runs through the following steps [2, 8, 12]:

1. Randomly initialize the matrix $\mathbf{B}^{(t)}$ and set $t = 0$;
2. Find coefficient matrix $\Delta^{(t)}$, solving the problem in Eq. (8) under constraints $\delta_{ki} \geq 0$ and $\sum_{k=1}^{K} \delta_{ki} = 1$;
3. Given the coefficients $\delta_{ki}^{(t)}$, compute the intermediate archetypes, solving the equation in (8) for $\mathbf{A}^{(t)}$;
4. Update Eq. (8) over $\mathbf{B}$ under constraints $\sum_{k=1}^{K} \beta_{ki} = 1$ and $\beta_{ki} \geq 0$;
5. Set $t = t + 1$, $\mathbf{B}^{(t)} = \mathbf{B}^{(t+1)}$ and calculate $\mathbf{A}^{(t)} = \mathbf{B}^{(t)}\mathbf{X}$;
6. Compute the objective function and, unless it falls below a threshold, continue with Step 2;
7. Stop.

Note that the matrices $(\boldsymbol{\Gamma}^T \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T$ and $\mathbf{B}$ play the same role, i.e., to project the single points in a K-dimensional space. In fact, defining the $(\boldsymbol{\Gamma}^T \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T = \mathbf{B}^*$, it is possible to shown that the objective functions of FCM (9) and AA (10) optimize

an equivalent criterion.

$$f(\Delta, \mathbf{B}) = \|\mathbf{X} - \Gamma \mathbf{B}^* \mathbf{X}\|_2^2. \tag{9}$$

$$f(\Delta, \mathbf{B}) = \|\mathbf{X} - \Delta \mathbf{B} \mathbf{X}\|_2^2. \tag{10}$$

2.3 Consensus Clustering

In this paper, the consensus clustering procedure is structured in two fundamental steps: (1) to find and represent the consensus between the partitions of fuzzy c-means and archetypal analysis through correspondence analysis [3, 15, 16] and (2) to measure the consensus through the principal indices of CC.

Let $\mathbf{X}$ be an $N \times J$ data matrix with $T = \{T_1, \ldots, T_R\}$ and $V = \{V_1, \ldots, V_C\}$ two partitions of $\mathbf{X}$: the consensus between partitions T and V is found by starting from the entries shown in the cross-classifying contingency table (shown in Table 1) and crossing the two partitions [17]. Many proposals have been put forth in the literature for the consensus measurement, including Boulis and Ostendorf [7], Fowlkes and Mallows [14], Hubert and Arabie [17], Steinley [28], Strehl and Ghosh [29], and Yeung and Ruzzo [31]. This paper has used the Adjusted Rand Index (ARI) among these different options [23, 26]. ARI was first proposed by Hubert and Arabie [17] in such a context; the index assumes a generalized hypergeometric distribution as a null hypothesis. The two clusterings are drawn randomly, with a fixed number of clusters and a fixed number of elements in each cluster. ARI is then the normalized difference between the Rand Index and its expected value under the null hypothesis. The ARI is defined as shown in Eq. (11).

$$ARI = \frac{\sum_{r=1}^{R} \sum_{c=1}^{C} \binom{n_{rc}}{2} - \binom{n}{2}^{-1} \sum_{r=1}^{R} \binom{n_{r.}}{2} \sum_{c=1}^{C} \binom{n_{.c}}{2}}{\frac{1}{2} \left[\sum_{r=1}^{R} \binom{n_{r.}}{2} + \sum_{c=1}^{C} \binom{n_{.c}}{2} \right] - \binom{n}{2}^{-1} \sum_{r=1}^{R} \binom{n_{r.}}{2} \sum_{c=1}^{C} \binom{n_{.c}}{2}}. \tag{11}$$

ARI has an expected value of 0 for independent clusterings and a maximum value of 1 for identical clusterings.

Table 1 Cross-table between partition T and partition V

		Partition V				
		v_1	v_2	$\cdots$	v_C	
Partition T	t_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
	t_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	t_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	n

3   Simulation Study

This section demonstrates an application of simulated data; in particular, it evaluates the consensus between fuzzy c-means and archetypal analysis in different experimental conditions.

According to Fordellone and Palumbo [13], data were generated from four multivariate Gaussian distributions, each with four dimensions. The first is a multivariate Gaussian distribution with $\mu = [0, 0, 0, 0]^T$ and $\mathbf{\Sigma} = \mathbf{I}$ (i.e., noise); the last three are multivariate Gaussian distributions that simulate three groups of units according to the experimental conditions shown in Table 2; the groups follow the scheme shown below:

Group 1: $\mathbf{X} \sim \mathcal{N}(\mu = [\text{-}20 \quad 10 \quad 30 \quad 15]^T, \mathbf{\Sigma})$
Group 2: $\mathbf{X} \sim \mathcal{N}(\mu = [\quad 0 \quad 20 \quad 15 \quad \text{-}5]^T, \mathbf{\Sigma})$
Group 3: $\mathbf{X} \sim \mathcal{N}(\mu = [\quad 15 \quad 5 \quad \text{-}7 \quad 20]^T, \mathbf{\Sigma})$

Table 2 also shows the consensus results obtained from the eight experimental conditions; in particular, the Rand and *adjusted* Rand indices are reported, which are measurements of agreement/disagreement (i.e., consensus) between two different partitions (FCM and AA in this case). It is worth noting that the maximum consensus was achieved in the first two experimental conditions, where there was a low correlation between the variables and a normal kurtosis level, whereas the minimum level of consensus was shown in the experimental conditions, which demonstrated platykurtic kurtosis levels (rows 3, 4, 7, and 8 in the table).

The aim of the simulation study was to establish the degree of reliability of consensus prototyping under several different hypotheses. In the eight proposed cases, the lowest levels of the consensus occurred most notably when platykurtic distributions were present; this occurred because a low level of kurtosis stimulated the presence of outlier points, and AA is very sensitive to extreme points.

Table 2 Consensus results from simulated data

Experimental conditions			Prototyping results		Consensus measurement	
N	Cor	Kurt.	Groups	Size	Rand	$Rand_{adj}$
900	0.2–0.4	$G = 3$	3	900 (100.0%)	1.000	1.000
300	0.2–0.4	$G = 3$	3	300 (100.0%)	1.000	1.000
900	0.2–0.4	$G < 3$	3	625 (69.4%)	0.725	0.382
300	0.2–0.4	$G < 3$	3	185 (61.7%)	0.683	0.296
900	0.6–0.8	$G = 3$	3	599 (66.6%)	0.753	0.444
300	0.6 − 0.8	$G = 3$	3	202 (67.3%)	0.758	0.460
900	0.6 − 0.8	$G < 3$	3	533 (59.2%)	0.698	0.329
300	0.6 − 0.8	$G < 3$	3	189 (63.0%)	0.720	0.373

4   Personality Traits Finding

In psychology, trait theory (also called dispositional theory) is an approach that aims to study human personality. Trait theorists are primarily interested in the measurement of different traits that can be defined as habitual patterns of behavior, thought, and emotion [21, 22].

Consensus clustering between fuzzy c-means and archetypal analysis is used in this section to delineate the personality traits (http://personality-testing.info/) from a sample of 898 adults (493 males and 405 females). The dataset has 40 ordinal items (from "Strongly disagree" to "Strongly agree"), with ten items for each measurement scale (numeric). Four different scales were used as part of an experimental "DISC" personality test from the International Personality Item Pool (http://ipip.ori.org/newCPIKey.htm):

- *Assertiveness* (AS): the quality of being self-assured and confident without being aggressive;
- *Social confidence* (SC): generally described as a state of being certain about something;
- *Adventurousness* (AD): represented by a desire to engage in activities with some potential for physical danger;
- *Dominance* (DO): conceptualized as a measure of individual differences in levels of group-based discrimination.

The following subsection shows the results that were obtained by the application of the consensus-prototyping approach on the four quantitative measurement scales. The scales were computed using principal component analysis (PCA) applied to the fifty available items. The number of groups ($K = 3$) was chosen according to the FCM and AA scree-plots shown in Fig. 1. The scree-plots show the reductions in the objective functions at different values of K.
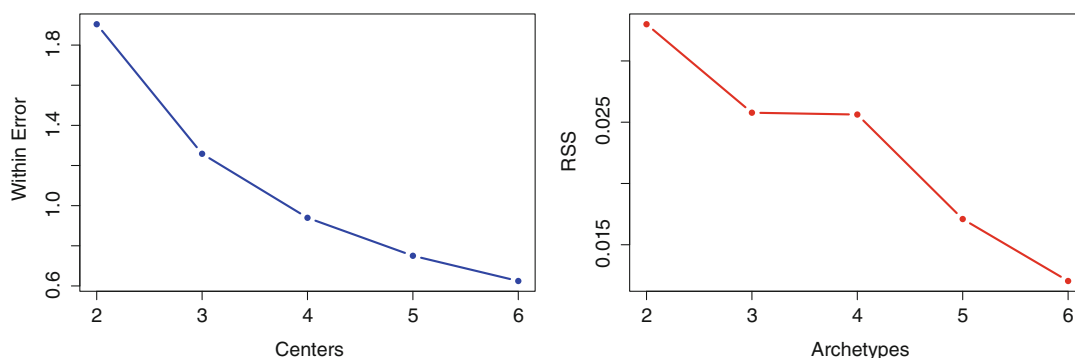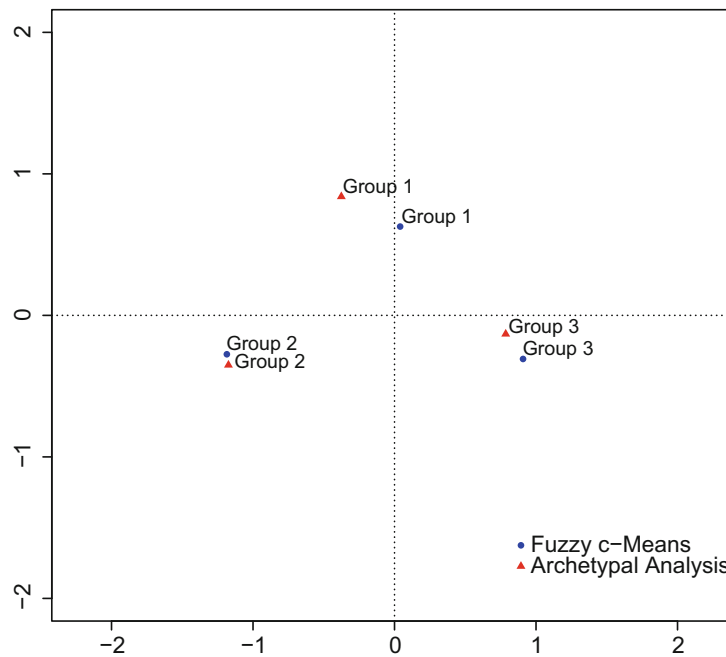

Fig. 1 Scree-plots of the fuzzy c-means and archetypal analysis approaches for different numbers of groups

Table 3 Consensus table between the fuzzy c-means and archetypal analysis approaches

| | | Archetypal analysis | | | |
		Group 1	Group 2	Group 3	Total
Fuzzy c-means	Group 1	126	35	125	286
	Group 2	53	218	0	271
	Group 3	0	1	340	341
	Total	179	254	465	898

Fig. 2 Correspondence analysis applied to contingency Table 3

The results of CC are shown in consensus Table 3. A correspondence analysis is also applied on the consensus table to graphically illustrate the consensus level. The results are represented in Fig. 2, where we may see that the three groups are very close when they have a maximum level of inertia. In Table 3, in contrast, the consensus prototypes represent roughly 76% of the sample; the Adjusted Rand Index equals 0.499, while the Rand Index equals 0.768.

Figure 3 shows the defined groups and the single distributions of the scales. The different colors (purple, blue, and red) and the different symbols ($\bullet$, $\triangle$, and +) of the points represent the three prototypes. The white points in Fig. 3 represent the observation without consensus.

Due to space considerations, we cannot discuss in depth the profiles of the prototypes that were found. Looking at the scatterplot matrix shown in Fig. 3, however, we can state that prototype 1 (the red $\bullet$ symbol) is characterized by a high level of scale *AD*; prototype 2 (the blue $\triangle$ symbol) presents low values for scales *SC, DO, AS*; and the last prototype (the purple + symbol) is characterized by high levels of scales *SC, AS* and a low level of *AD*.

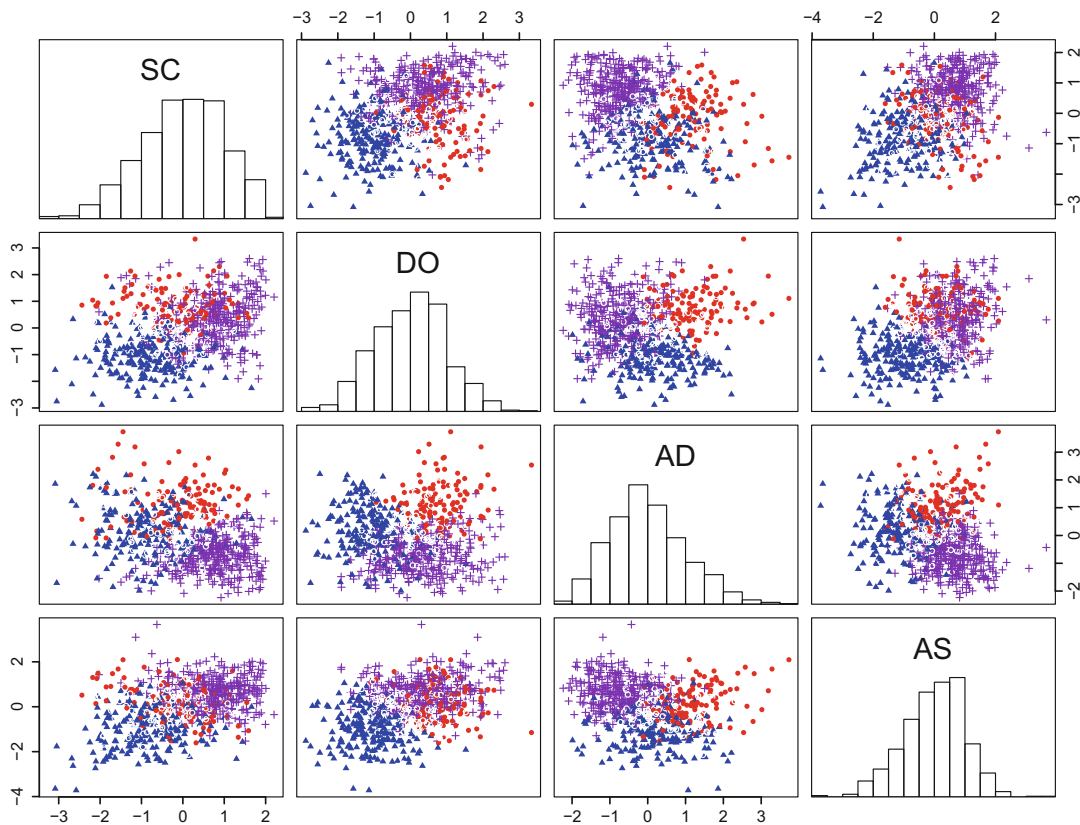**Fig. 3** Scatterplots of the three prototypes computed via consensus clustering

5 Concluding Remarks

The empirical results shown in the present paper lead us to argue that when the groups are well defined (thus avoiding any overlap), the consensus clustering between the two different partitioning methods (i.e., fuzzy c-means and archetypal analysis) clarifies the presence of well-definable prototypes. The simulation study was helpful for appreciating which causes can deeply affect the consensus among the two approaches: the platykurtic level and the presence of multivariate outliers in the data have greatly affected performance of the consensus analysis; the high correlation among variables has worsened performance, but not much in comparison with the previous perturbater effects.

References

1. Arabie, P., Boorman, S.A.: Multidimensional scaling of measures of distance between partitions. J. Math. Psychol. **10**(2), 148–203 (1973)
2. Bauckhage, C., Thurau, C.: Making archetypal analysis practical. In: Pattern Recognition, pp. 272–281. Springer, Heidelberg (2009)
3. Benzécri, J.P.: Correspondence Analysis Handbook. Marcel Dekker, New York (1992)
4. Berry, W.M., Browne, M., Langville, N.A., Pauca, P.V., Plemmons, J.R.: Algorithms and applications for approximate non-negative matrix factorization. Comput. Stat. Data Anal. **52**(1), 155–173 (2007)
5. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA (1981)
6. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy c-means clustering algorithm. Comput. Geosci. **10**(2), 191–203 (1984)
7. Boulis, C., Ostendorf, M.: Combining multiple clustering systems. In: Knowledge Discovery in Databases: PKDD, pp. 63–74. Springer, Berlin (2004)
8. Cutler, A., Breiman, L.: Archetypal analysis. Technometrics **36**(4), 338–347 (1994)
9. Dembele, D., Kastner, P.: Fuzzy c-means method for clustering microarray data. Bioinformatics **19**(8), 973–980 (2003)
10. D'Esposito, M., Palumbo, F., Ragozini, G.: Archetypal analysis for interval data in marketing research. Ital. J. Appl. Stat. **18**, 343–358 (2006)
11. Diday, E.: Optimization in non-hierarchical clustering. Pattern Recogn. **6**(1), 17–33 (1974)
12. Eugster, M., Leisch, F.: From spider-man to hero-archetypal analysis in R. J. Stat. Softw. **1**, 1–23 (2009)
13. Fordellone, M., Palumbo, F.: Prototypes definition through consensus analysis between Fuzzy c-Means and Archetypal Analysis. Ital. J. Appl. Stat. **26**(2), 141–162 (2014)
14. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. Am. Stat. Assoc. **78**(383), 553–569 (1983)
15. Greenacre, M.J.: Interpreting multiple correspondence analysis. Appl. Stoch. Model Data Anal. **7**(2), 195–210 (1991)
16. Greenacre, M.J.: Correspondence Analysis in Practice, 2nd edn. Chapman and Hall/CRC, Boca Raton (2010)
17. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**(1), 193–218 (1985)
18. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: a review. ACM Comput. Surv. (CSUR) **31**(3), 264–323 (1999)
19. Johnson, C.S.: Hierarchical clustering schemes. Psychometrika **32**(3), 241–254 (1967)
20. Karypis, G., Han, E.H., Kumar, V.: Chameleon: hierarchical clustering using dynamic modeling. Computer **32**(8), 68–75 (1999)
21. Kassin, S.: Psychology. Prentice-Hall, Upper Saddle River, NJ (2003)
22. Loh, W.Y., Shih, Y.S.: Split selection methods for classification trees. Stat. Sin. **74**(4), 815–840 (1997)
23. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971). doi:10.2307/2284239
24. Rosch, E.: Cognitive representations of semantic categories. J. Exp. Psychol. **104**(3), 192 (1975)
25. Rosch, E.: Principles of categorization. In: Concepts: Core Readings, pp. 189–206. The MIT Press, Cambridge, MA (1999)
26. Santos, J.M., Embrechts, M.: On the use of the adjusted Rand index as a metric for evaluating supervised classification. In: Artificial Neural Networks CANN 2009, pp. 175–184. Springer, New York (2009)
27. Scheibler, D., Schneider, W.: Montecarlo tests of the accuracy of cluster analysis algorithms: a comparison of hierarchical and nonhierarchical methods. Multivar. Behav. Res. **20**(3), 283–304 (1985)

28. Steinley, D.: Properties of the Hubert-Arable Adjusted Rand Index. Psychol. Methods **9**(3), 386 (2004)
29. Strehl, A., Ghosh, J.: Cluster ensembles-a knowledge reuse framework for combining multiple partitions. J. Mach. Learn. Res. **3**, 583–617 (2003)
30. Sun, H., Wang, S., Jiang, Q.: FCM-based model selection algorithms for determining the number of clusters. Pattern Recogn. **37**(10), 2027–2037 (1985)
31. Yeung, K.Y., Ruzzo, W.L.: Details of the adjusted Rand index and clustering algorithms, supplement to the paper "An empirical study on principal component analysis for clustering gene expression data". Bioinformatics **17**(9), 763–774 (2001)