PROTOTYPE DEFINITION THROUGH CONSENSUS ANALYSIS BETWEEN FUZZY C-MEANS AND ARCHETYPAL ANALYSIS

Mario Fordellone

Department of Statistical Sciences, Sapienza University of Rome (Italy)

Francesco Palumbo

Department of Political Sciences, Federico II University of Naples (Italy)

Abstract *The general aim of cluster analysis is to build prototypes, or typologies of units that present similar characteristics. In this paper we propose an alternative approach based on consensus analysis of two different clustering methods to suitably obtain prototypes. The clustering methods used are fuzzy c-means (centre approach) and archetypal analysis (extreme approach). The consensus clustering is used to assess the correspondence between the clustering solutions obtained.*

Keywords: *Archetypal analysis, Fuzzy c-means, Consensus analysis, Prototyping, Definition prototypes*

1. INTRODUCTION

Aristotelian thinking first defined *categories* as the basic entities of human knowledge. In this sense, Smith and Medin (1981) refer to the concept of category as the highest order of *genera* that cannot be defined by a mere listing of properties shared by all elements. According to Rosch (1975, 1999), prototypes are those elements that better than others represent a category (on this point see also Rocha, 1999). The degree of representativeness can be measured using a distance function to a salient entity of the category, *i.e.* a prototype. Prototypes can be observed or unobserved (abstract) entities: not necessarily as real elements of the category (Medin and Schaffer, 1978). In this paper we present a novel statistical approach to identify a set of prototypes given a training multivariate data set and an *a priori* known number of categories.

Formally, let $\mathbf{X}$ be a generic $N \times J$ data matrix, where each row represents a statistical unit described by J features, and let U be a set of descriptions $U = \{u_1, u_2, \ldots, u_K\}$ in the feature space. Prototyping consists in defining a rule that

✉ Corresponding Author: Francesco Palumbo
Università degli Studi di Napoli Federico II, Dept. of Political Sciences, fpalumbo@unina.it
Via Leopoldo Rodinò 22, 80138 Naples, IT

associates each row of $\mathbf{X}$ to the elements of U (Friedman et al., 2001). In other words, we introduce the matrix $\mathbf{Y}$ of order $N \times K$ where the general element $y_{ik} = 1$ if the generic row vector $\mathbf{x}_i$ is associated to the description u_k and 0 elsewhere, with $i = 1,\ldots,N$ and $k = 1,\ldots,K$. Under the fuzzy logic paradigm (Kaufman and Rousseeuw, 1990; Zadeh, 1994), each row of $\mathbf{X}$ is associated to one or more descriptions of U by the membership degrees y_{ik}, under the following constraints: $\sum_{k=1}^{K} y_{ik} = 1$ and $y_{ik} \geq 0$, $\forall k \in 1,2,\ldots,K$. The membership degrees represent the degree of prototypically of a concept regarding a particular category and a category can also be defined by the degrees to which its elements belong to the prototype. These descriptions represent relevant characteristics of the prototype (Rocha, 1999). Given an association criterion, the solution consists in solving an optimisation problem for $\mathbf{Y}$ and $\mathbf{P}$, where $\mathbf{P}$ is the $K \times J$ matrix of the K prototypes defined in the J-dimensional feature space. Numerical techniques that solve such an optimisation problem have been proposed in many fields and are based on several different criteria. Widely used techniques are based on non-hierarchical clustering algorithms (Diday, 1974; Gnanadesikan, 2011; Jain et al., 1999; Scheibler and Schneider, 1985); although many other approaches can be adopted (Johnson, 1967; Karypis et al., 1999; Loh and Shih, 1997).

In this paper, we propose a two-step procedure based on consensus analysis (Hubert and Arabie, 1985) to define the set of prototypes $U = \{u_1, u_2, \ldots, u_K\}$ starting from the above-defined general matrix $\mathbf{X}$. The first step aims to define two partitions of $\mathbf{X}$ 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 synthesis of the two partitions. When more than one partition can be defined in the same data, consensus analysis is proposed with a twofold aim: (*i*) to find a unique partition solution as a synthesis of all partitions; (*ii*) to measure the agreement among the different partitions and between the synthesis and all the partitions. In the specific literature such consensus analysis is also referred to as consensus clustering (Boulis and Ostendorf, 2004; Hubert and Arabie, 1985; Nguyen and Caruana, 2007; Strehl and Ghosh, 2003).

By definition, a prototype must be a description of the data irrespective of method. Thus, we believe that there should be a consensus among different methods. Such a consensus indicates sharp profiles that can be reported as prototypes. The novelty of the present proposal consists in finding a set of a given number of prototypes by the consensus analysis to pair off the partitions obtained via two different methods: fuzzy c-means (FCM) (Bezdek, 1981) and archetypal analysis (AA) (Bauckhage and Thurau, 2009; Cutler and Breiman, 1994). The former

method seeks K homogeneous groups with respect to their barycentres, whereas the latter identifies a set of extreme points, called archetypes, and creates a group around each archetype. The k-means technique (Jain, 2010; MacQueen, 1967; Steinley, 2006) represents a key reference method in non-hierarchical clustering in the class of *hard clustering* methods. The k-means algorithm minimises the sum of squared distances between the observations and the cluster mean, or centroid. Fuzzy c-means (Bezdek et al., 1984; Hathaway and Bezdek, 1994) and archetypal analysis (Cutler and Breiman, 1994; Eugster and Leisch, 2009) can be seen as a particular case (*soft clustering*) of the k-means technique, under different constraints. Archetypal analysis minimises the sum of distances between each point and a set of K archetypes, as defined by a convex combination of extreme points.

The paper is structured as follows: in Section 2 we introduce and describe fuzzy c-means and archetypal analysis, in Section 3 we present the consensus analysis methodology, in Section 4 the methods are applied to eight synthetic data sets with different characteristics, and in Section 5 we present an application to real data.

2. BACKGROUND

In this section we show that FCM and AA can be defined as two different factorisations of the data matrix $\mathbf{X}$ under different constraints that are discussed in sub-sections 2.1 and 2.2. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a generic $N \times J$ data matrix, where $\mathbf{x}_i$ ($i = 1, 2, \ldots, N$) is the generic row vector and $\mathbf{P}$ be the $K \times J$ unknown prototypes matrix. Fuzzy c-means and archetypal analysis are grounded on the solution of the following nonnegative factorisation problem (Berry et al., 2007):

$$(\mathbf{Y}, \mathbf{P}) = \arg\min_{\mathbf{Y},\mathbf{P}} \|\mathbf{X} - \mathbf{YP}\| . \tag{1}$$

Note that we define the centres matrix $\mathbf{C}$ for fuzzy c-means and the archetype matrix $\mathbf{A}$ for archetypal analysis as the $K \times J$ matrix $\mathbf{P}$. Moreover, we define the membership matrix $\mathbf{\Gamma}$ and Δ for FCM and AA, respectively, as the $N \times K$ matrix $\mathbf{Y}$. The generic elements γ_{ik} and δ_{ik} vary in $[0, 1]$ and represent the membership degree of the unit $\mathbf{x}_i'$ to the archetype $\mathbf{a}_k$ or to the centre $\mathbf{c}_k$, depending on which factorisation we are referring to.

2.1. FUZZY C-MEANS

The Fuzzy c-means clustering algorithm (Bezdek, 1981; Dembele and Kastner, 2003) is an extension of the k-means algorithm for fuzzy clustering. Fuzzy c-

means and k-means minimise the sum of the weighted squared distances between the N units from the K centres. Yet the k-means assumes that weights can vary in $\{0,1\}$ and fuzzy c-means in $[0,1]$. Formally, given the generic $N \times J$ data matrix $\mathbf{X}$, the objective function is defined as follows:

$$W = \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik}^{m} d(i,k)^2, \tag{2}$$

where $d(i,k) = \|\mathbf{x}_i - \mathbf{c}_k\|$, m is the fuzzifier parameter (commonly set to 2) and the quantity γ_{ik} represents the corresponding weight. The quantity W is minimised under the following constraints: (i) $\sum_{k=1}^{K} \gamma_{ik} = 1$; (ii) $\gamma_{ik} \geq 0$, where $k \in 1,2,\ldots,K$. Let us indicate with the notation $\|\cdot\|_2$ the quadratic norm, hence the elements $\mathbf{c}_k$ and γ_{ik} are defined according to the following formulae:

$$\mathbf{c}_k = \frac{\sum_{i=1}^{N} \gamma_{ik}^{m} \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{ik}^{m}}, \tag{3}$$

$$\gamma_{ik} = \left(\sum_{k'=1}^{K} \left(\frac{\|\mathbf{x}_i - \mathbf{c}_k\|_2}{\|\mathbf{x}_i - \mathbf{c}_{k'}\|_2} \right)^{\frac{2}{m-1}} \right)^{-1}. \tag{4}$$

The FCM algorithm runs through the following steps (Sun et al., 2004):

1. Randomly initialise the cluster centres $\mathbf{C}^{(t)}$ and set $t = 0$;
2. Calculate γ_{ik} using the Equation (4);
3. Calculate $\mathbf{C}^{(t+1)}$ using Equation (3);
4. If $| \left(\mathbf{C}^{(t)} - \mathbf{C}^{(t+1)} \right) | \leq \varepsilon$ go to Step 5; else $\mathbf{C}^{(t)} = \mathbf{C}^{(t+1)}$, set $t = t + 1$, and go to Step2;
5. Print centres matrix $\mathbf{C}$ and membership matrix $\mathbf{\Gamma}$;
6. Stop.

2.2. ARCHETYPAL ANALYSIS

Archetypal Analysis aims to represent the units of a multivariate data set as a convex combination of the most extremal K data points, here called archetypes which are linear combinations of the data points (Bauckhage and Thurau, 2009; Cutler and Breiman, 1994). In the literature, the term *archetype* is used to define *the original pattern or model of which all things of the same type are representations or copies* (http://www.merriam-webster.com/dictionary/archetype). However, in a prototyping approach the concrete problem is to find a few, not necessarily observed points (archetypes) in a set of multivariate observations such that all the

data can be well represented as convex combinations of the archetypes. Recently, the AA has been used in many fields such as philosophy, psychology and also statistics (D'Esposito et al., 2006; Eugster and Leisch, 2009).

Formally, given a $N \times J$ data matrix, AA finds a set of archetypes $\{\mathbf{a}_1, \ldots, \mathbf{a}_K\}$ that are linear combinations of the data points as shown in Formula 5:

$$\mathbf{a}_k = \sum_{i=1}^{N} \mathbf{x}_i \beta_{ik}. \tag{5}$$

Moreover, each data point must be approximated as a convex combination of the K archetypes. The coefficients $\beta_{ik} \geq 0$ such that the archetypes resemble the data and $\sum_{i=1}^{N} \beta_{ik} = 1$ are convex mixtures of the data. Then, for a given choice of archetypes, AA minimises the quantity $\left\|\mathbf{x}_i - \sum_{k=1}^{K} \mathbf{a}_k \delta_{ki}\right\|_2$, under constraints $\delta_{ki} \geq 0$ and $\sum_{k=1}^{K} \delta_{ki} = 1$, where δ_{ki} represents the associated membership level. In practice, the data points are represented as mixtures of archetypes. Furthermore, a suitable choice of K archetypes minimises the residual sum square (*RSS*):

$$RSS = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{k=1}^{K} \mathbf{a}_k \delta_{ki} \right\|_2 = \sum_{i=1}^{N} \left\| \mathbf{x}_i - \sum_{k=1}^{K} \sum_{s=1}^{N} \mathbf{x}_s \beta_{sk} \delta_{ki} \right\|_2 . \tag{6}$$

It may be convenient to write Formula 6 in linear algebra notation, as shown in Formula 7:

$$RSS = \|\mathbf{X} - \Delta\mathbf{A}\|_2 = \|\mathbf{X} - \Delta\mathbf{B}\mathbf{X}\|_2 , \tag{7}$$

where $\mathbf{X}$ is the $N \times J$ data matrix, $\mathbf{A}$ is the $K \times J$ archetype matrix, Δ is the $N \times K$ membership matrix and $\mathbf{B}$ is the $K \times N$ matrix of β's coefficients.

In agreement with Cutler and Breiman (1994) and Bauckhage and Thurau (2009), given the number of clusters K, the parameters of the AA algorithm are estimated using the following steps:

1. Randomly initialise the matrix $\mathbf{B}^{(t)}$ and set $t = 0$;
2. Find coefficient matrix $\Delta^{(t)}$ solving the problem in formula (7) under constraints $\delta_{ki} \geq 0$ and $\sum_{k=1}^{K} \delta_{ki} = 1$
3. Given the coefficients $\delta_{ki}^{(t)}$ compute intermediate archetypes solving the equation in formula (7) for $\mathbf{A}^{(t)}$;
4. Determine the coefficient matrix $\mathbf{B}^{(t+1)}$ that minimises the constrained problem $\left\|\mathbf{B}\mathbf{X}^{(t+1)} - \mathbf{A}^{(t)}\right\|_2$, under constraints $\beta_{ik} \geq 0$ and $\sum_{i=1}^{N} \beta_{ik} = 1$.

5. Set $t = t+1$, $\mathbf{B}^{(t)} = \mathbf{B}^{(t+1)}$ and calculate $\mathbf{A}^{(t)} = \mathbf{B}\mathbf{X}^{(t)}$;
6. Compute the *RSS* and, unless it falls below a threshold, continue with step 2;
7. Stop.

2.3 EXAMPLE ON A TOY DATA SET

This section presents a simple application on a simulated toy data set of fuzzy c-means and archetypal analysis. Data were generated from a three-variate Gaussian distribution with $n = 90$ units classified *a priori* in three different groups. In particular, we have homogeneous variances and mean values as follows: the first group with $\mu_1 = -5$, $\mu_2 = 5$, $\mu_3 = 0$; the second group with $\mu_1 = 5$, $\mu_2 = -5$, $\mu_3 = 5$; the third group with $\mu_1 = 10$, $\mu_2 = 10$, $\mu_3 = -5$. Figure 10 shows the pairs plot of the simulated toy data.

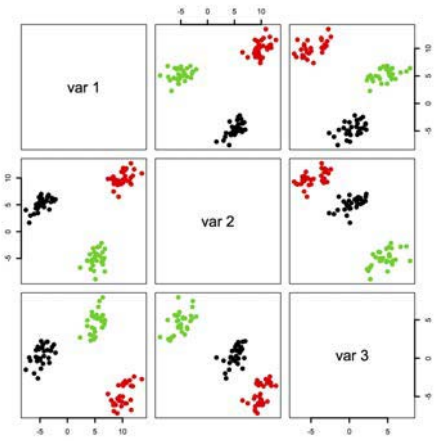

Figure 1: Pairs plot of the simulated toy data

With $K = 3$, fuzzy c-means and archetypal analysis determined, respectively, the centres and archetypes shown in Table 1, where we can appreciate that archetypal points are more extreme with respect to the corresponding fuzzy c-means centres. Figure 2 shows the membership degrees associated to each statistical unit.

Table 1: Centre matrix C (FCM) and archetype matrix A (AA)

	C			A		
Group	c_1	c_2	c_3	a_1	a_2	a_3
1	-4.674	5.173	0.380	-6.271	5.405	0.211
2	9.985	9.886	-4.783	11.050	10.690	-5.476
3	5.228	-4.995	4.669	6.120	-6.825	5.866

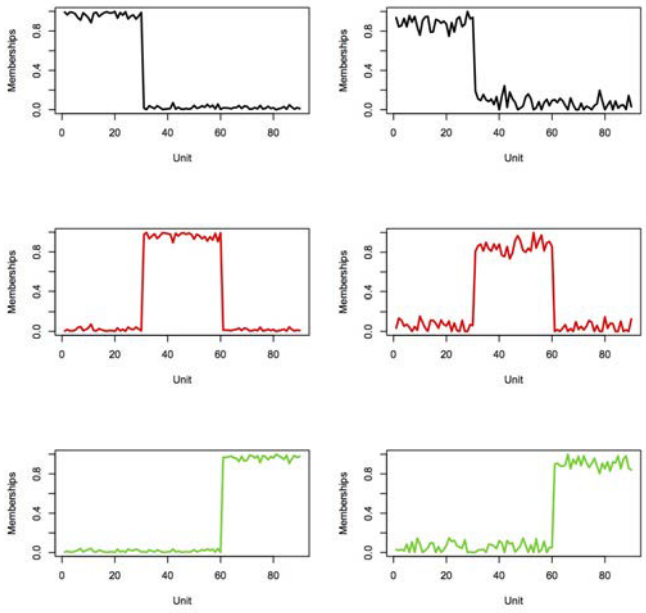

Figure 2: Membership function computed by FCM and AA

3. CONSENSUS ANALYSIS

Let $\mathbf{X}$ be a $N \times J$ data matrix, and $T = \{t_1, \ldots, t_R\}$ and $V = \{v_1, \ldots, v_C\}$ two partitions of $\mathbf{X}$. Then n_{rc} ($r = 1, \ldots, R$; $c = 1, \ldots, C$) represents the number of objects assigned to the classes t_r and v_c, with respect to the two partitioning criteria. Consensus between the partitions T and V is evaluated starting from the entries of the cross-classifying contingency table, which is shown in Table 2 and crosses the two partitions (Hubert and Arabie, 1985). Many consensus measures have been pro-

Table 2: Contingency table in comparing partitions

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

posed in the literature (Boulis and Ostendorf, 2004; Fowlkes and Mallows, 1983; Strehl and Ghosh, 2003). In the present work we consider three consensus indexes that are based on Rand's measure of agreement (1971). Such measures define the agreement between T and V through a two-step procedure: first, the groups of T and V are paired on the basis of their composition; secondly, all the possible pairs $\{\mathbf{x}_i, \mathbf{x}_{i'}\}$ (with $i \neq i'$ and $i, i' \in 1, \ldots, N$) are considered with respect to their assignment in the $R \times C$ contingency matrix $\mathbf{M}$, with general term n_{rc} ($c = 1, 2, \ldots, C$ and $r = 1, 2, \ldots, R$) (Fowlkes and Mallows, 1983). Herein, we assume $R = C = K$, where K indicates the number of groups after the cluster analysis. The square $R \times C$ table allows us to define the four quantities that are illustrated below (formulae 8 to 11). All these quantities but the first depend on K. The quantity in 8 expresses the number of ways that n units can pair

$$S(K) = \binom{n}{2} = \frac{n(n-1)}{2}. \tag{8}$$

The quantity $T(K)$ in 9 represents the total number of combinations starting from the units that have been assigned to the paired groups of T and V

$$T(K) = \sum_{r=1}^{R} \sum_{c=1}^{C} n_{rc}^2. \tag{9}$$

Analogously, the quantities $P(K)$ in 10 and quantity $Q(K)$ in 11 represent the sum, respectively by r and by t, of the number of pairs of units in Table 2, with respect to $t_1, \ldots t_r, \ldots, t_R$ and $v_1, \ldots v_c, \ldots, v_C$:

$$P(K) = \sum_{r=1}^{R} n_{r\cdot}^2 \tag{10}$$

$$Q(K) = \sum_{c=1}^{C} n_{\cdot c}^2. \tag{11}$$

Formally, the quantity in (9) represents the agreements in the classification of the objects, and the quantities in (10) and (11) represent the disagreements. Hence, the quantity

$$T(K) - \frac{1}{2}\left(P(K) + Q(K)\right),$$

represents the total number of agreements (A), whereas the quantity

$$\frac{1}{2}(P(K) + Q(K)) - T(K),$$

indicates the total number of disagreements (D) and $A + D = S(K)$. Formalisation of A and D is shown in (12) and (13).

$$A = \binom{n}{2} + \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2 - \frac{1}{2}\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] \tag{12}$$

$$D = \frac{1}{2}\left[\sum_{r=1}^{R} n_{r\cdot}^2 + \sum_{c=1}^{C} n_{\cdot c}^2\right] - \sum_{r=1}^{R}\sum_{c=1}^{C} n_{rc}^2. \tag{13}$$

It is worth noting that a rough comparison between A and D gives an idea of the consensus degree. To measure the agreement (disagreement), three main approaches have been proposed in the literature: $A/S(K)$ by Rand (1971), $D/S(K)$ by Johnson (1968) (see also Arabie and Boorman, 1973; Hubert and Arabie, 1989), and $(A-D)/S(K)$ by Hubert and Baker (1977). In all three cases, the measure varies in [0, 1] and can be interpreted as the empirical probability of agreement or disagreement (the former two) and as the difference between the probability of agreement and disagreement (the last one).

4. SIMULATION STUDY

This section presents a study on simulated data according to the scheme illustrated in subsection 4.1. We used a critical membership value to assign the units to the FCM groups and AA groups; units with the maximum membership degree can be univocally assigned to the corresponding group. By contrast, for a description of prototypes, we used the critical membership value of 0.5 for the assignment. In practice, units with a membership degree greater 0.5 can be univocally assigned to the corresponding group, while an extra group receives the units that do not exceed a membership degree of 0.5 for any group; it is termed *residual group*.

4.1 SIMULATED DATA

Data were sampled from a multivariate Gaussian distribution with eight dimensions (four are white noise) and with different sample sizes. The variables describe eight different experimental conditions derived from combining three factors, each with two levels, *i.e.*, 2^3 different levels. The three factors used, on the four modified variables, are: (*i*) sample sizes (small $N = 200$ and large $N = 1000$); (*ii*) correlation between variables (low level $0.2 - 0.4$ and high level $0.6 - 0.8$); (*iii*) kurtosis (normal level $G = 3$ and platykurtic level $G < 3$). The kurtosis level was reduced on 15% of simulated data, so that we increased the sampling probability in the tails of the distribution.

The four perturbed variables were generated with a structure of four groups of units (scheme in Table 3). FCM and AA were applied, fixing a number of groups $K = 4$. Final consensus analysis results are presented in subsection 4.2.

Table 3: Means of the simulated data groups

Units range	Var.1	Var.2	Var.3	Var.4
1 - 50	-20	-10	30	15
51 - 100	0	20	15	-5
101 - 150	15	5	-7	20
151 - 200	30	-15	15	-5
Units range	Var.1	Var.2	Var.3	Var.4
1 - 250	-20	-10	30	15
251 - 500	0	20	15	-5
501 - 750	15	5	-7	20
751 - 1000	30	-15	15	-5

4.2 CONSENSUS

The results obtained according to the eight different sampling schemes are summarised in Table 4. The first four columns describe the experimental conditions, and columns 6 to 8 the consensus measures as formally described in Section 3. Our results show that the overall consensus of prototypes is mostly influenced by the correlations between variables, where the measure of consensus by Hubert is below 0.6, the measure of agreement by Rand (1971) is below 0.8 and the measure of disagreement by Arabie and Boorman (1973) is over 0.2. However, the level of consensus rises in the presence of highly correlated variables and platykurtic

distributions, where the three indexes are about 0.5, 0.7 and 0.2, respectively.

To appreciate the geometric consensus structure, we propose the graphical visualisation of the confusion matrix that is based on correspondence analysis (CA). Correspondence analysis is an explorative computational factorial method for the study of associations between nominal variables (Fellenberg et al., 2001). The aim of the method is to embed the rows and columns of a matrix in the same space, with the first two (or three) factors that summarise the proportion of the information that it is present in the data. In its simplest form, the genetic element n_{rc} of the $R \times C$ matrix $\mathbf{M}$, represents the co-occurences of r and c and $n_{r.}$ and $n_{.c}$ respectively denote the row and column marginals. Finally, n is the grand total of $\mathbf{M}$. In other words, CA decomposes the inertia of a contingency table, which is measured by the χ^2 statistic. Higher association corresponds to higher χ^2 values. In CA, points are represented such that the sum of distances of the points to their centroid is proportional to the value of the χ^2 statistic of the data table and the farther away a point is from centroid, the higher its row contribute to the statistic. The link between rows and columns in the contingency table implies that if a column determines an outstanding entry of a row, the corresponding row and column points tend to lie on a common line through the centroid. The two points lie on the same side with respect to the common centroid, where a larger distance corresponds to a stronger association. In case of a negative association, the column-point and row-point lie on opposite sides of the centroid (Beh and Lombardo, 2014; Benzécri et al., 1992; Fellenberg et al., 2001; Greenacre, 1991). Correspondence analysis presents the results in a simple graphical display which permits a more rapid interpretation and understanding of the data (Greenacre, 2010).

In this case, CA graphically displays the distances between the pairs of groups that belong to the partitions generated by FCM and AA in each of the eight different experimental conditions. Although it is not possible to pair *a priori* the groups of the two different partitions, we re-labeled with the same number the groups of the two partitions that are closer to get a more clear interpretation, e.g. group 1 of FCM is close to group 4 of AA, then the latter is re-labeled with the number 1. Figures 3 and 4 show the plot of correspondence analysis applied on the experimental conditions from 1 to 4 and on the experimental conditions from 5 to 8, respectively. In practical terms, lower distances between the pairs represents a better association between the groups, and *vice versa*. The best results are thus presented by the first two conditions (the first two in Figure 3), while the worst results are presented by experimental conditions 5 and 6 (first two plots of Figure 4).

Table 4: Simulated data: results of consensus analysis on the definition of the prototypes

Experimental Conditions				Consensus Measures		
N	Kurt.	Corr.	N. Groups	Hubert	Rand	Arabie
1000	$G = 3$	$0.2 - 0.4$	4	0.996	0.998	0.002
200	$G = 3$	$0.2 - 0.4$	4	0.995	0.997	0.002
1000	$G < 3$	$0.2 - 0.4$	4	0.611	0.806	0.195
200	$G < 3$	$0.2 - 0.4$	4	0.627	0.814	0.187
1000	$G = 3$	$0.6 - 0.8$	4	0.374	0.687	0.313
200	$G = 3$	$0.6 - 0.8$	4	0.258	0.629	0.371
1000	$G < 3$	$0.6 - 0.8$	4	0.531	0.766	0.235
200	$G < 3$	$0.6 - 0.8$	4	0.527	0.763	0.236

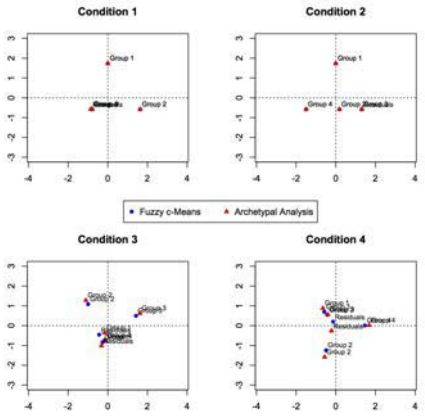

Figure 3: Simulated data: correspondence analysis for 1-4 experimental conditions

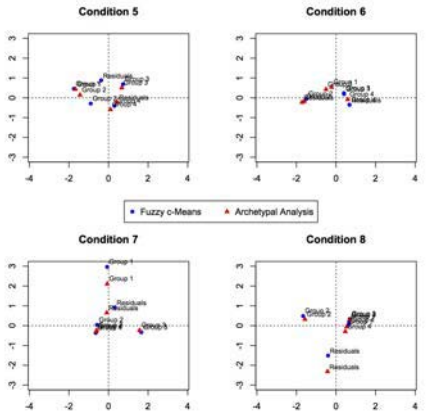

Figure 4: Simulated data: correspondence analysis for 5-8 experimental conditions

The main aim of this simulation study was to establish the degree of the reliability of consensus-prototyping under several different hypotheses. In the eight proposed cases, the lowest levels of the consensus were observed in the presence of (i) platykurtic distributions and (ii) high levels of the correlations between the variables. This happens because the AA is very sensitive to the extreme points.

5. APPLICATION ON REAL DATA

This section presents a real data application of prototyping through consensus analysis between FCM and AA on *Wine recognition data*. The data set is available at the UCI repository website (http://archive.ics.uci.edu/ml/). It is the result of the chemical analysis of wines grown in an Italian region, derived from three different cultivars. The 13 constituents were measured on 178 types of wine from the three cultivars: 59, 71 and 48 instances are in class one, two and three, respectively. The 13 chemical continuous attributes of wine data set are: 1. alcohol (Alc), 2. malic acid (Mal), 3. ash (Ash), 4. alkalinity of ash (A_Ash), 5. magnesium (Mag), 6. total phenols (Phe), 7. flavonoids (Fla), 8. nonflavanoid phenols (N_Phe), 9. proanthocyanins (Pro), 10. color intensity (Col), 11. hue (Hue), 12. OD280-OD315 of diluted wines (R_OD), and 13. proline (Pro). Figure 5 and Table 5 show the boxplots of the standardised variables and their descriptive statistics, respectively. It is worth pointing out that there are different levels of skewness and outlier values are present for some variables.

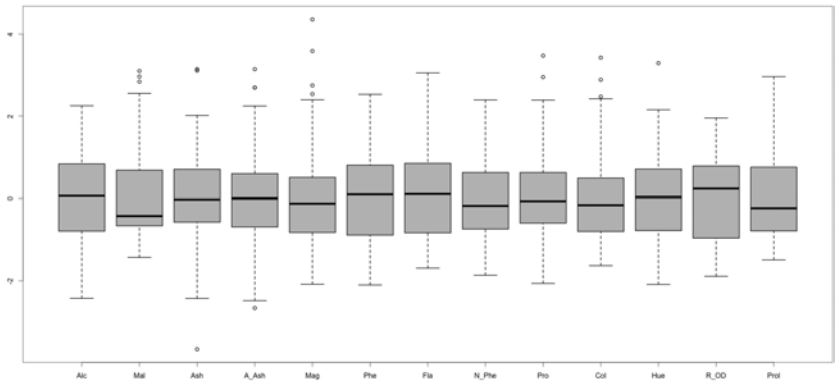

Figure 5: Boxplots of standardised variables of wine data set

5.1. FUZZY C-MEANS AND ARCHETYPAL ANALYSIS

Before applying FCM, a principal component analysis (PCA) was performed to obtain more stable results. This approach is called *tandem analysis* (Arabie et al., 1996; Vichi and Kiers, 2001) and it is commonly used in the high dimensional data clustering problems to cope with the so-called '*curse of dimensionality*' issue. Table 6 shows the first five eigenvalues: we notice that only the first three eigenvalues are greater than one and their corresponding cumulative variance is equal to 66%. Consequently, the first three principal components are kept to perform the FCM,

Table 5: Descriptive statistics on the standardised variables of wine data set

Variable	Min	1^{st}Qu.	Median	3^{rd}Qu.	Max
Alc	-2.4274	-0.7860	0.0608	0.8338	2.2534
Mal	-1.4290	-0.6569	-0.4219	0.6679	3.1004
Ash	-3.6688	-0.5705	-0.0237	0.6961	3.1474
A_Ash	-2.6635	-0.6872	0.0015	0.6004	3.1456
Mag	-2.0824	-0.8221	-0.1219	0.5082	4.3591
Phe	-2.1013	-0.883	0.0957	0.8067	2.5324
Fla	1.6912	-0.8252	0.1059	0.8467	3.0542
N_Phe	-1.8630	-0.7381	-0.1756	0.6078	2.3956
Pro	-2.0632	-0.5956	-0.0627	0.6274	3.4753
Col	-1.6297	-0.7929	-0.1588	0.4926	3.4258
Hue	-2.0888	-0.7654	0.0330	0.7112	3.2924
R_OD	-1.8897	-0.9496	0.2371	0.7864	1.9554
Prol	-1.4890	-0.7824	-0.2331	0.7561	2.9631

with the number of groups $K = 3$. Figure 6 shows the variable/factor correlations (with respect to the first two of the three chosen factors) and Figure 7 allows us to see the three true groups plotted on the first three components (different symbols represent the groups). AA works on the original data set dimensionality (13 variables) and the number of archetypes in input is fixed at three, to ensure the consistency with the FCM solution. The RSS is 0.09, which represents a good result in terms of fitting. Let us have a look at Figures 8 and 9 to appreciate the results of the two methods. Notice that points are plotted in a two dimensional subspace. To corroborate the choice of $K = 3$, we report the graphic results of the solutions achieved for $K = 3,\ldots,6$; for the sake of space, analytical results have been omitted. In Figure 8 the ellipses approximately represent the group edges, whereas in the AA results (Figure 9) the points are assigned to the closest archetype and different symbols are associated to each archetype. Also the scree plots of the two methods shown in Figure 10 and 11 suggest choosing the

Table 6: Eigenvalues associated to the first five factors

	F1	F2	F3	F4	F5
Std. deviation	4.706	2.497	1.446	0.919	0.853
Prop. of Variance	0.362	0.192	0.111	0.071	0.066
Cum. Prop.	0.362	0.554	0.665	0.736	0.802

number of groups equal to 3: the significance reduction of the squared residual is manifested for $K = 3$.

5.2. CONSENSUS ANALYSIS

Table 7 presents the confusion matrix for the partitions of two methods: high frequencies in the main diagonal cells of Table 7 denote a strong consensus between AA and FCM. The detail of the consensus analysis measurement on prototype definition are reported in Table 8. The three measures of consensus show that between the two methods there exists an high consensus level. Finally, Figure 12 jointly illustrates the membership degree behaviours with respect to AA and FCM (see Section 2.3 for the interpretation).

Table 7: Confusion matrix of the partitions FCM and AA

		Archetypal Analysis			
		a_1	a_2	a_3	
Fuzzy c-Means	c_1	51	0	0	51
	c_2	1	63	1	65
	c_3	0	3	59	62
		52	66	60	178

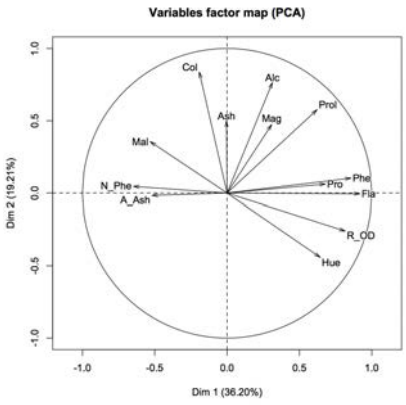

Figure 6: Eigenvectors represented on the first two principal components.

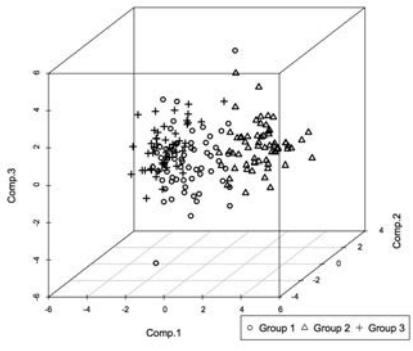

Figure 7: 3D Plot of the data represented on the first three principal components.

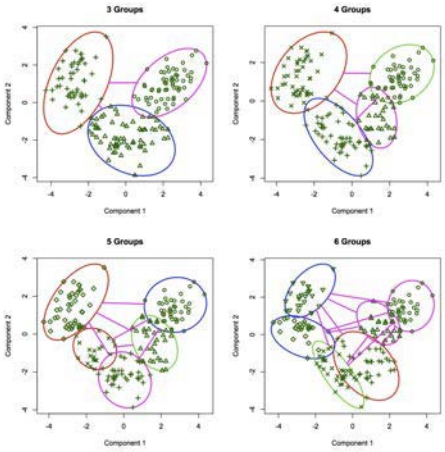

Figure 8: C-means map for K=3 to 6

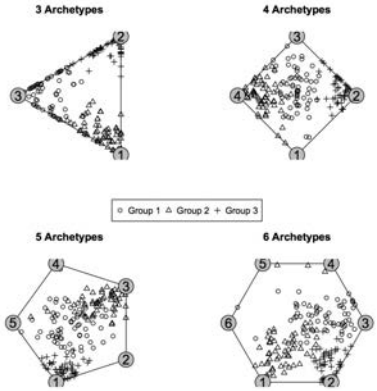

Figure 9: Archetypes map for K=3 to 6

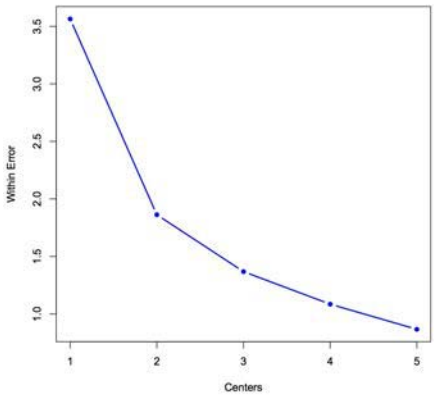

Figure 10: FCM scree plot from groups 1 to 5

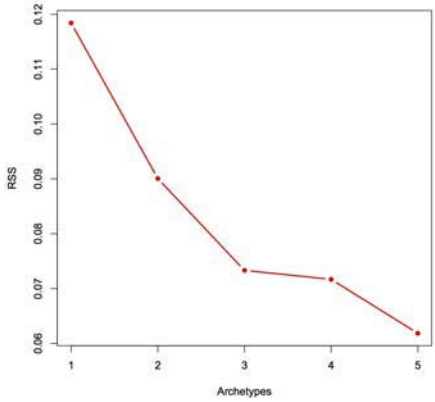

Figure 11: AA scree plot from groups 1 to 5

Table 8: Results of consensus analysis on the definition of the prototypes

Consensus Measure		
Hubert	Rand	Arabie
0.923	0.963	0.04

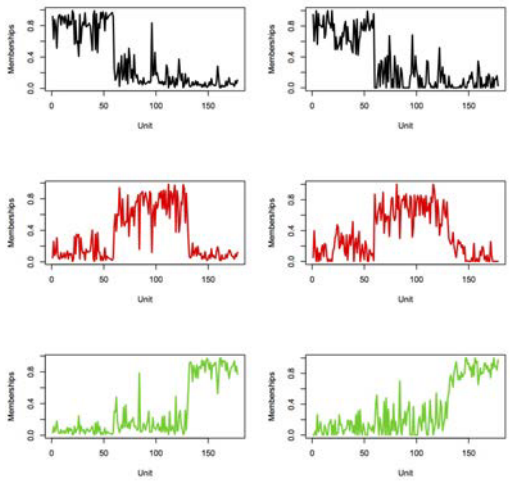

Figure 12: Membership functions computed by FCM and AA

5.3. ANALYSIS AND DESCRIPTION OF PROTOTYPES

The first representation of the prototypes is the CA applied to the partitions with $K+1$ groups. We recall that the extra group receives the units that have not reached a membership value greater than the critical value of 0.5 for any group. Figure 13 shows the results plot of correspondence analysis (details of the results in Table 9).

Table 9: Results of correspondence analysis between FCM and AA

	Principal inertia	
Dimension	1	2
Value	0.975	0.869
Prop. of Variance	0.529	0.471
Cum. Proportion	0.529	1.000

CA confirms the high consensus between the two partitions. These figures allow us to state that as many prototypes as the number of groups can be identified in the data. Hence, the last step consists in giving a description for each prototype to define the set of descriptions. In general a prototype is not described by single-valued data, but a fuzzy description is preferred. In this case we propose to describe our prototypes by interval-valued data. Such an approach has been

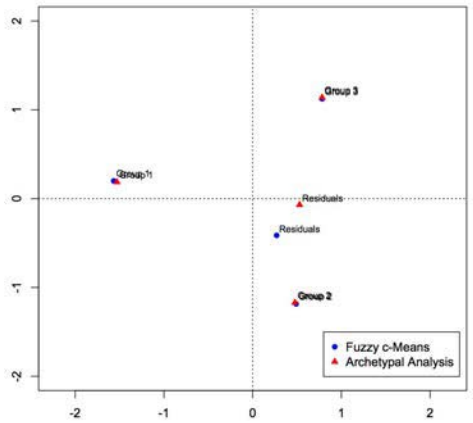

Figure 13: Visualisation of correspondence analysis between FCM and AA

proposed and widely used in the symbolic data analysis (SDA) framework, where a prototype is described by interval-valued data (Billard and Diday, 2003; Hickey et al., 2001; Kao et al., 2014; Palumbo and Irpino, 2005). In order to set the lower and upper bounds of the intervals, we consider the hypersphere having as centre the mean value as identified by the FCM and the *radius* equal to distance from the centre and its paired archetype. All points belonging to the hypersphere are associated to the prototypes and their *minimum* and *maximum* values represents the prototype description. Formally, $\mathbf{x}_i \in \mathbf{u}_k$ if $\mathbf{x}_i = x_{i1}, \ldots x_{ij}, \ldots x_{iJ}$ belongs to the (hyper)sphere *i.e.*: $\sum_{j=1}^{J}(x_{ij} - c_{jk})^2 \leq r^2(k)$, where $r^2(k) = \sum_{j=1}^{J}(a_{jk} - c_{jk})^2$ is the squared *radius* and c_{jk} and a_{jk} indicate the coordinates of the k^{th} prototype centre and archetype, respectively. Figure 14 shows the prototypes identified in the PCA reduced space $\mathbb{R}^2$. The plot shows the three prototypes represented by the three circles containing the units identified by the consensus.

The distributions of the three prototypes are represented in Figure 15. Boxplots denote that the *Prototype 1* includes the wines that have a low level of alcohol, ash, alkalinity of ash, flavonoids, magnesium, color intensity, OD280-OD315 of diluted wines and Proline whereas high levels are verified of malic acid and nonflavanoid phenols; the *Prototype 2* includes the wines that have a low level of malic acid, magnesium, flavonoids, proanthocyanins and proline whereas high levels are verified of Ash, Alkalinity of ash, Hue and OD280-OD315 of diluted wines; the *Prototype 3* includes the wines that have a low level of malic acid, ash, alkalinity of ash, Magnesium, flavonoids and proline whereas high levels are verified of alcohol, phenols, intensity and OD280-OD315 of diluted wines.

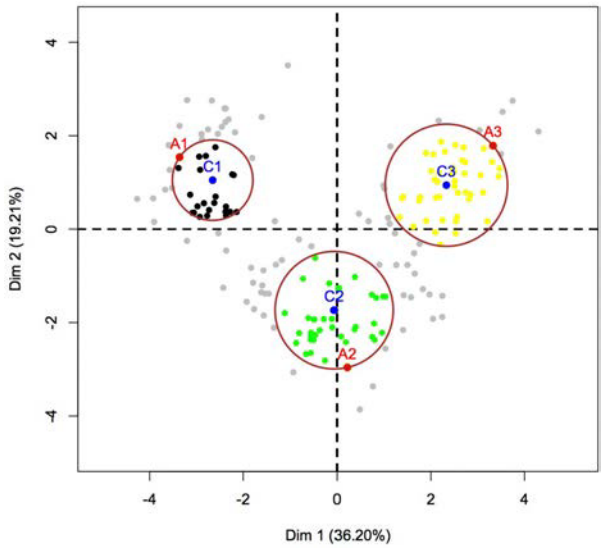

Figure 14: Representation of the prototypes in $\mathbb{R}^2$

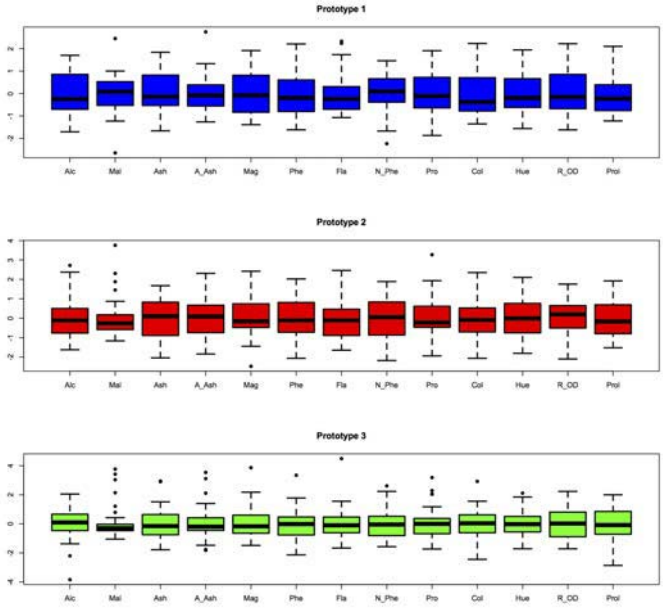

Figure 15: Boxplot of the prototypes distributions

6. CONCLUSIONS

In this paper we proposed a novel approach to identify prototypes in the data on the basis of the consensus between two different partitioning methods: archetypal analysis and fuzzy c-means. Pairs of units that were classified in the same group through the two different methods were taken into account for prototype definition. As the two methods satisfy two different criteria it was preferred to define the prototypes on the basis of fuzzy logic (Kaufman and Rousseeuw, 1990; Zadeh, 1994). In other words, each prototype summarises the properties of all those units that satisfy both criteria. It is worth pointing out that AA focuses on the extreme points and that FCM focuses on the mean point of each group.

Our simulation results confirmed our hypothesis: when groups are well defined, avoiding any overlapping, consensus analysis between the two different partitioning methods underlined the presence of the groups. Moreover, the simulation was useful to study the causes that can affect the consensus between the two approaches: first, correlation between variables, secondly presence of multivariate outliers (Figures 3 and 4).

In conclusion, the prototypes definitions through the consensus approach is more reliable in comparison to the classical approaches, i.e., finding groups in respect to the consensus-criterion, guarantees more homogeneous prototypes.

REFERENCES

Arabie, P. and Boorman, S.A. (1973). Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10(2):148–203.

Arabie, P., Hubert, L. J. and De Soete, G. (1996). *Clustering and Classification*. Word Scientific, River Edge, NJ.

Bauckhage, C. and Thurau, C. (2009). Making archetypal analysis practical. *Pattern Recognition*, pages 272–281. Springer.

Beh, E. J. and Lombardo, R. (2014). *Correspondence Analysis: Theory, Practice and New Strategies*. John Wiley & Sons.

Benzécri, J.-P. et al. (1992). *Correspondence Analysis Handbook*. Marcel Dekker New York.

Berry, W.M., Browne, M., Langville, N.A., Pauca, P.V. and Plemmons, J.R. (2007). Algorithms and applications for approximate non-negative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Bezdek, J.C., Ehrlich, R. and Full, W. (1984). Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2):191–203.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*, 98(462):470–487.

Boulis, C. and Ostendorf, M. (2004). Combining multiple clustering systems. In *Knowledge Discovery in Databases: PKDD 2004*, pages 63–74. Springer.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4):338–347.

Dembele, D. and Kastner, P. (2003). Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8):973–980.

D'Esposito, M., Palumbo, F. and Ragozini, G. (2006). Archetypal analysis for interval data in marketing research. *Statistica Applicata-Italian Journal of Applied Statistics*, 18:343–358.

Diday, E. (1974). Optimization in non-hierarchical clustering. *Pattern Recognition*, 6(1):17–33.

Eugster, M. and Leisch, F. (2009). From spider-man to hero-archetypal analysis in R. *Journal of Statistical Software*, pages 1–23.

Fellenberg, K., Hauser, N.C., Brors, B., Neutzner, A., Hoheisel, J.D. and Vin-gron, M. (2001). Correspondence analysis applied to microarray data. *Proceedings of the National Academy of Sciences*, 98(19):10781–10786.

Fowlkes, E.B. and Mallows, C.L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York.

Gnanadesikan, R. (2011). *Methods for Statistical Data Analysis of Multivariate Observations*, volume 321. John Wiley & Sons.

Greenacre, M.J. (1991). Interpreting multiple correspondence analysis. *Applied Stochastic Models and Data Analysis*, 7(2):195–210.

Greenacre, M.J. (2010). *Correspondence Analysis in Practice, second edition*. Chapman and Hall/CRC, Boca Raton, 2nd edition.

Hathaway, R.J. and Bezdek, J.C. (1994). Nerf c-means: Non-Euclidean relational fuzzy clustering. *Pattern Recognition*, 27(3):429–437.

Hickey, T., Ju, Q. and Van Emden, M.H. (2001). Interval arithmetic: From principles to implementation. *Journal of the ACM*, 48(5):1038–1068.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1):193–218.

Hubert, L. and Arabie, P. (1989). Combinatorial data analysis: Confirmatory comparisons between sets of matrices. *Applied Stochastic Models and Data Analysis*, 5(3):273–325.

Hubert, L. and Baker, F. (1977). Evaluating object set partitions free sort analysis. *Journal of Mathematical Psychology*, 16:233–253.

Jain, A.K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999). Data clustering: a review. *ACM Computing Surveys (CSUR)*, 31(3):264–323.

Johnson, C.S. (1967). Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254.

Johnson, S. (1968). A simple cluster statistic (unpublished). *Bell Labs*.

Kao, C.-H., Nakano, J., Shieh, S.-H., Tien, Y.-J., Wu, H.-M., Yang, C.-K. and Chen, C.-H. (2014). Exploratory data analysis of interval-valued symbolic data with matrix visualization. *Computational Statistics & Data Analysis*, 79:14–29.

Karypis, G., Han, E.-H. and Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. *Computer*, 32(8):68–75.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7(4):815–840.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium*, 1:281–297.

Medin, D.L. and Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207.

Nguyen, N. and Caruana, R. (2007). Consensus clusterings. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 607–612. IEEE.

Palumbo, F. and Irpino, A. (2005). Multidimensional interval-data: Metrics and factorial analysis. In Janssen, J. and Lenca, P., editors, *Applied Stochastic Models and Data Analysis*, pages 689–698. ENST Bretagne.

Rand, W.M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Rocha, L.M. (1999). Evidence sets: modeling subjective categories. *International Journal of General System*, 27(6):457–494.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192.

Rosch, E. (1999). *Concepts: Core Readings*, chapter 8. Principles of categorization, pages 189–206. The MIT Press, Cambridge, MA.

Scheibler, D. and Schneider, W. (1985). Monte carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research*, 20(3):283–304.

Smith, E.E. and Medin, D.L. (1981). *Categories and concepts*. Harvard University Press Cambridge, MA.

Steinley, D. (2006). K-means clustering: a half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34.

Strehl, A. and Ghosh, J. (2003). Cluster ensembles-A knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617.

Sun, H., Wang, S. and Jiang, Q. (2004). FCM-based model selection algorithms for determining the number of clusters. *Pattern Recognition*, 37(10):2027– 2037.

Vichi, M. and Kiers, H. (2001). Factorial k-means analysis for two way data. *Computational Statistics and Data Analysis*, 37:29–64.

Zadeh, A.L. (1994). Fuzzy logic, neural networks, and soft computing. *Commu-nications of the ACM*, 37(3):77–84.