Ranking and selection of unsupervised learning marketing segmentation

Germán Sánchez-Hernández[a,b], Francisco Chiclana[c,d], Núria Agell[a], Juan Carlos Aguado[b]

[a] *ESADE Business School, Universitat Ramon Llull, Barcelona, Spain*
[b] *Automatic Control Department, Universitat Politècnica de Catalunya, Barcelona, Spain*
[c] *Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, UK*
[d] *DMU Interdisciplinary Group in Intelligent Transport Systems, Faculty of Technology, De Montfort University, Leicester, UK*

Abstract

This paper addresses the problem of choosing the most appropriate classification from a given set of classifications of a set of patterns. This is a relevant topic on unsupervised systems and clustering analysis because different classifications can in general be obtained from the same data set. The provided methodology is based on five fuzzy criteria which are aggregated using an Ordered Weighted Averaging (OWA) operator. To this end, a novel multi-criteria decision making (MCDM) system is defined, which assesses the degree up to which each criterion is met by all classifications. The corresponding single evaluations are then proposed to be aggregated into a collective one by means of an OWA operator guided by a fuzzy linguistic quantifier, which is used to implement the concept of fuzzy majority in the selection process. This new methodology is applied to a real marketing case based on a business to business (B2B) environment to help marketing experts during the segmentation process. As a result, a segmentation containing three segments consisting of 35, 98 and 127 points of sale respectively is selected to be the most suitable to endorse marketing strategies of the firm. Finally, an analysis of the managerial implications of the proposed methodology solution is provided.

Keywords: fuzzy selection criteria, OWA operator, classification selection, market segmentation, linguistic quantifier

1. Introduction

The use of unsupervised learning systems allows the behaviour of certain phenomena to be identified without relying on expert knowledge or information from past situations. Indeed, the main characteristic of this type of learning systems is that it works with patterns without explicitly knowing their output. Because of this, unsupervised learning systems have been considered in the literature as systems capable to capture knowledge from complex structures [1–3].

Choosing the most appropriate classification from a given set of classifications of a set of patterns is an important topic on unsupervised systems and, in particular, on clustering analysis. In most cases, the use of these techniques leads to several classifications as outputs, i.e. various classifications are compatible with the set of given patterns. For this reason, research in this area aims to develop suitable tools and models for selecting classifications [4–6].

Previous research in this direction uses selection criteria as filters: a set of criteria is applied sequentially to all the obtained classifications [6–9]. All those classifications failing to meet a particular criterion are discarded and not taken into account in the application of the subsequent criterion. The following drawback can be associated with this type of methodology: because a true-false decision is applied in the application of each criterion, this could result in classifications being discarded and not taken into account when they marginally fail to meet one particular criterion but meet other criteria with a high score. Therefore, a classification might be discarded prematurely when its 'overall' score, with respect to the set of criteria, would have been high. In an extreme case, this methodology could produce no result because none of the classifications meet a particular criterion, which could indicate that the criterion in particular might not have been the most adequate or taken into account.

An alternative approach to the sequential approach described above, which has been successfully applied in multi-criteria decision making (MCDM), is that of evaluating the degree up to which each criterion is met by all classifications, i.e. the use of fuzzy criteria, and, only after this, obtaining an overall aggregated value for each classification reflecting the degree up to which the whole set of criteria is satisfied by each classification. Note

Email addresses: german.sanchez@esade.edu (Germán Sánchez-Hernández), chiclana@dmu.ac.uk (Francisco Chiclana), nuria.agell@esade.edu (Núria Agell), juan.carlos.aguado@upc.edu (Juan Carlos Aguado)

Cite as: **German Sanchez-Hernandez, Francisco Chiclana, Nuria Agell, Juan Carlos Aguado: Ranking and selection of unsupervised learning marketing segmentation. Knowledge-Based Systems 44 (2013) 20–33, doi: 10.1016/j.knosys.2013.01.012**

that the objective of the aggregation step is to combine a set of criteria in such a way that the final aggregation output takes into account all the single fuzzy criterion [10]. The final selection of classifications naturally derives from this set of overall degrees, and the drawback mentioned above does not apply.

Many different families of aggregation operators have been studied [10–20]. Among them the Ordered Weighted Averaging (OWA) operator proposed by Yager [19] is one of the most widely used. Among the reasons to support this extensive use of the OWA operator is that it allows the implementation of the concept of fuzzy majority in the aggregation phase by means of a fuzzy linguistic quantifier [21] representing the proportion of satisfied criteria 'necessary for a good solution' [22]. This is done by using the linguistic quantifier in the computation of the weights associated with the OWA operator. In addition, Marichal [23] investigated the aggregation of dependent criteria and the fuzzy integral was found to be the appropriate aggregation operator in these cases. The most representative fuzzy integrals are the Choquet integral and the Sugeno integral. It is well known that the OWA operator is a particular case of Choquet integral, and consequently it is not necessary to assume independence of criteria when using the OWA operator.

From the application point of view, unsupervised systems have been relevant in a wide range of domains, among which it is worth mentioning: text categorisation, images recognition, telecommunications fraud detection, stock price forecasting, bioinformatics, fault diagnosis, pollution classification and clinical or socio-economic systems [24–34]. In the marketing field, finding new and creative solutions is valuable because these allow for the definition of new strategies and innovation. The use of unsupervised learning algorithms allows us to suggest segmentations that are, in principle, not trivial. In this sense, behavioural patterns of 'interesting' profiles could be established by using this type of algorithm and these may reveal new customer profiles not yet known to experts [35–39].

This paper presents a novel classification selection methodology based on a set of fuzzy criteria and the MCDM approach described above. This MCDM approach uses an aggregation function based on OWA operators defined via a linguistic qualifier to summarise the information gathered through the set of fuzzy criteria. This new methodology has been implemented in the statistical computing tool R [40] and applied to a real marketing problem.

The paper is structured as follows. In the next section five selection criteria related to market segmentation are defined, and their fuzzy nature and interpretation are considered. Following that, in Section 3, the MCDM approach is introduced and the OWA operator and fuzzy linguistic quantifier concepts are provided. A case study to select a segmentation from a real business situation is described in Section 4, and results obtained by applying the proposed new methodology are analysed. In Section 5 conclusions are drawn and suggestions for further future research work are given.

2. Fuzzy criteria for selecting classifications

The use of unsupervised learning algorithms enables to find out non trivial classifications. However, when many different classifications are obtained, how to choose the best one with respect to the proposed objective? In this section methods and criteria for the evaluation of clustering results are reviewed. Below five fuzzy indicators, adapted and extended from criteria introduced by Sánchez-Hernández et al. [8] to help solve this problem, are described and defined. For each fuzzy criterion, a membership function describing the degree up to which it is verified by a particular classification is proposed.

2.1. Clustering validation

This section reviews criteria and methods to evaluate classifications derived from the application of any of the available clustering techniques. There are mainly three types of clustering validation criteria [41, 42]: internal, external and relative. An internal criterion tries of determine if the classification structure is intrinsically appropriate for the data. An external criterion of validation compares the considered classification with an *a priori* structure: either a previously known partition of the analysed dataset, typically provided by some domain experts, or an external variable not participating in the clustering process. Finally, a relative criterion measures the relative similarity between two classifications.

Several works reviewing cluster validation indexes have been published [6, 43–45]. These works and other using or defining new criteria are shown in Table 1. Criteria associated with the *compactness* concept compute how closely related the individuals in a cluster are, being usually based on indexes measuring density or variance of the data; *separability* criteria determine how distinct or well-separated a cluster is from other clusters; criteria related to the *prediction strength* of the clusters usually calculate the accuracy rate of a model constructor from them [6, 46]; some criteria are based on the number of important *features* [6]; criteria quantifying the achievement of *goals* can be very heterogeneous, from applying economic theories [7], being assessed by graphical visualisations [47], or checking the existence of outliers clusters or pairs of variables [6]. External criteria require the existence of an *a priori* external variable or classification defined for each of the individuals. The computation of an index associated with external criteria can be performed by any of the following indexes: Rand statistic,

Jaccard coefficient, Fowlkes and Mallows index, Hubert's statistic and so on. The computation of relative criteria implies the pairwise comparison between clusters, usually performed by some domain experts.

Paper	Comments	Internal criteria					External criteria	Relative criteria
		Compactness	Separability	Accuracy	Features	Goals		
[48] Ramze Rezaee et al., 1998	One index for fuzzy c-Means	Yes: compactness	Yes: separation	No	No	No	No	No
[49] Cheng et al., 1999	Subspace clustering	Yes: high density	No	No	No	No	No	No
[43] Haldiki et al., 2001	Review	Yes: several		No	No	No	Yes: several	Yes: several
[7] Choi et al., 2005	Association rules	No	No	No	No	Yes: Recency Freq. & M.V.	No	No
[46] Tibshirani & Walther, 2005	Validation by prediction strength	Yes: variance	Yes: bias	Yes: prediction strength	No	No	No	No
[45] Yatskiv & Gusarova, 2005	Review	No	No	No	No	No	Yes: several	Yes: several
[47] Bittmann & Gelbard, 2009	Visualisation of hierarchical clustering	Yes: minimal hetereogeneity	No	No	No	Yes: visualisation	No	No
[50] Wang et al., 2009	Clinical application	Yes: Davies-Bouldin & relapse-free		No	No	No	No	No
[51] Wu et al., 2009	External criteria for k-Means	No	No	No	No	No	Yes: several	No
[52] Xiong et al., 2009	k-Means	Yes: Sum of Squared Errors	Yes: entropy and Coef. of Variation	Yes: F-measure	No	No	No	No
[44] Liu et al., 2010	Internal criteria review	Yes: several	Yes: several	No	No	No	No	No
[6] Osei Bryson, 2010	Review	No	No	Yes: accuracy	Yes: # of important variables	Yes: outliers, Max/Min	No	Yes: pairwise, full & partial expert
Method presented	Review & application	Yes: I_C (coherence)		Yes: I_A (accuracy)	No	Yes: I_U & I_B (usefulness & balanced)	Yes: I_D (dependency)	No

Table 1: Clustering validation criteria

Although there are some methods to guide the search of which comparisons should be made for minimising their number, relative criteria have not been taken into account in this work due to the usual difficulty in getting this feedback from the experts. All the analysed papers review or define criteria based on a few concepts used for clustering evaluation, while almost all concepts are covered in this work.

2.2. First criterion: useful number of classes

The usability of a classification is based on its informativeness and manageability: it is worthwhile examining classifications that have a sufficient number of classes to generate new knowledge, but are small enough to produce an easy and manageable model. For instance, in marketing environments in which these classifications are used to extract behavioural patterns to design market strategies, the number of classes distinguished is usually taken to be between three and five [53]. This is because marketing campaigns with less than three segments may not be informative; while those with more than five segments may not be manageable.

The assumption of a classification with a number of classes M between K_1 and K_2 to be considered useful for a given problem does not imply that a classification with a number of classes lower than K_1 or higher than K_2 should be automatically discarded. This is specially true in those cases when there is enough evidence to suggest that such classifications perform well with respect to the rest of criteria. A natural approach in these cases would be that of associating a value to each classification to indicate how well they fit with the criterion 'useful number of classes'. By doing this, we move from a crisp to a fuzzy interpretation of the criterion 'useful number of classes', i.e. we move from the use of a characteristic function to the use of a membership function.

Note that a classification with a single class is trivial and therefore not useful, while a classification with a number of classes between K_1 and K_2 is considered totally useful. The minimum number of classes in any classification is 1 (contains all the individuals), while the maximum is N (each class contains just 1 individual). These two classifications are not informative and therefore these classification associated usefulness degree should be 0. A classification usefulness degree therefore should increase as the number of classes increases from 1 to K_1 and should decrease when the number of classes increases from K_2 to N. Therefore the general expression of the membership function associated with the criterion 'useful number of classes' is the following

Definition 1. *Given a classification $\mathcal{C}$, the index of usefulness is characterised by the following membership function:*

$$I_{U,K_1,K_2}(\mathcal{C}) = \begin{cases} f_1(M), & \text{if } 1 \leq M < K_1; \\ 1, & \text{if } K_1 \leq M \leq K_2; \\ f_2(M), & \text{if } K_2 < M \leq N, \end{cases} \tag{1}$$

where $M \in \mathbb{N}$ is the number of classes of $\mathcal{C}$; $K_1, K_2 \in \mathbb{N}$ such that $K_1 < K_2$ are two prefixed parameters; and f_1 is a strict increasing function and f_2 is a strict decreasing functions verifying $f_1(1) = f_2(N) = 0$ and $f_1(K_1) = f_2(K_2) = 1$.

This fuzzy interpretation of the criterion 'useful number of classes' covers a larger number of classifications than the classical approach. In the case study presented in Section 4, a left-linear function has been chosen as the simplest example of a strictly increasing function; while a right-exponential function has been selected

because the usefulness of a classification could decrease asymptotically when the number of classes increases. We note that the selection of different strict monotonic functions to the ones proposed here would not produce any change in the final ordering, because any two strict monotonic functions are mathematically equivalent in preserving an ordering.

$$I_{U,K_1,K_2}(\mathcal{C}) = \begin{cases} \frac{M-1}{K_1-1}, & \text{if } 1 \leq M < K_1; \\ 1, & \text{if } K_1 \leq M \leq K_2; \\ \frac{e^{(N-M)}-1}{e^{(N-K_2)}-1}, & \text{if } K_2 < M \leq N. \end{cases} \tag{2}$$

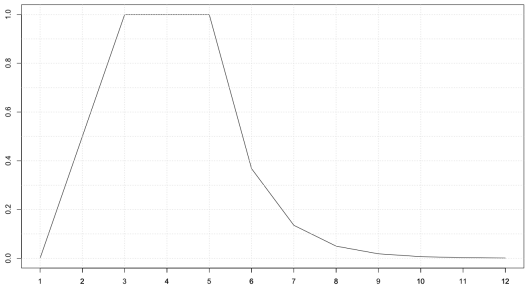

Figure 1: Fuzzy concept 'Useful number of classes' with $K_1 = 3$ and $K_2 = 5$

Figure 1 illustrates such a type of membership function with $K_1 = 3$ and $K_2 = 5$. Obviously, different increasing or decreasing functions could be used depending on the specific problem to solve and the preferences of the user: symmetric behaviour on both tails, linear or curve falls, concave or convex functions, etc.

2.3. Second criterion: balanced classes

The second criterion is based on the distribution of individuals within the obtained classes. For this reason, the variable Y = 'number of elements of each class in a given classification' is considered and its associated dispersion will be used to define the fuzzy concept of 'balanced classification'. Note that in some situations, classifications in which one class encompasses most of the individuals (unbalanced) are worth avoiding because they do not contribute to creating new or relevant knowledge. Nevertheless, in other contexts, unbalanced classifications could be desirable.

Let $N \in \mathbb{N}$ be the number of individuals to be classified, and $M \in \{1, \ldots, N\}$ be the number of classes obtained by the classification Y. Given that different classifications can produce a different number of classes, the coefficient of variation, CV_Y, is considered to be a fairer indicator than the standard deviation, σ_Y, in measuring the concept of 'balanced classification':

$$CV_Y = \frac{\sigma_Y}{\overline{Y}}, \tag{3}$$

with

$$\sigma_Y = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(Y_i - \overline{Y})^2} \quad \text{and} \quad \overline{Y} = \frac{1}{M}\sum_{i=1}^{M} Y_i = \frac{N}{M},$$

where Y_i is the number of individuals within the class i. Note that $CV_Y \geq 0$.

In order to define a normal membership function [51], the *index of balanced classes*, I_B, is proposed and the minimum and maximum values of CV_Y need to be computed. Obviously, the minimum value of CV_Y is zero, since the totally-balanced classification with one individual in each class, $Y_i = 1\ \forall i \in \{1, \ldots, N\}$, has zero coefficient of variation. In the following, the maximum value of CV_Y is computed by considering all the possible classifications for a given set of elements. Specifically, Corollary 1 determines the maximum value for CV_Y, given a set of N individuals and fixing the number of M classes, while Proposition 2 establishes the maximum value of CV_Y.

Lemma 1. *Let $F : \mathbb{R}^M \to \mathbb{R}^+$ be the the following function $F(Y_1, \ldots, Y_M) = \sum_{i=1}^{M} Y_i^2$. The solution to the following problem*

$$\begin{aligned} Max: \quad & F(Y_1, \ldots, Y_M) \\ s.t.: \quad & \textstyle\sum_{i=1}^{M} Y_i = N \in \mathbb{N} \\ & Y_i \geq 1\ \forall i \\ & N > M \end{aligned}$$

is $(Y_{1*}, \ldots, Y_{M*}) = (1, \ldots, 1, N - (M-1))$.

Proof. Let $(Y_1, \ldots, Y_M)$ such that

$$\sum_{i=1}^{M} Y_i = N > M$$

and

$$1 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_M < N - M + 1.$$

We need to prove:

$$F(Y_{1*}, \ldots, Y_{M*}) > F(Y_1, \ldots, Y_M),$$

or equivalently:

$$\sum_{i=1}^{M} [Y_{i*}^2 - Y_i^2] > 0.$$

Denoting $d_i = Y_{i*} - Y_i$, we have

$$\sum_{i=1}^{M} \left[Y_{i*}^2 - Y_i^2\right] = \sum_{i=1}^{M} (Y_{i*} - Y_i) \cdot (Y_{i*} + Y_i) = \sum_{i=1}^{M} d_i \cdot (Y_{i*} + Y_i).$$

It is clear that $\sum_{i=1}^{M} d_i = 0$, $d_i \leq 0 \ \forall i \in \{1, \ldots, M-1\}$ and $d_M > 0$. Also because $Y_{i*} + Y_i < Y_{M*} + Y_i \leq Y_{M*} + Y_M$, we have that $d_i \cdot (Y_{i*} + Y_i) > d_i \cdot (Y_{M*} + Y_M) \ \forall i \in \{1, \ldots, M-1\}$. Thus, we conclude:

$$\sum_{i=1}^{M} d_i \cdot (Y_{i*} + Y_i) > \sum_{i=1}^{M} d_i \cdot (Y_{M*} + Y_M) = (Y_{M*} + Y_M) \cdot \sum_{i=1}^{M} d_i = 0.$$

□

Proposition 1. *Let N be the number of individuals to classify. Considering all the classifications with M classes, those that result in one class with cardinality $N - M + 1$ and the rest of classes with cardinality 1 have the greatest coefficient of variation.*

Proof. Let be Y_* the variable 'number of elements of each class' associated with a classification with one class with cardinality $N - M + 1$ and the rest of classes with cardinality 1. Without lost of generality we can consider:

$$1 = Y_{1*} = Y_{2*} = \cdots = Y_{M-1*} < Y_{M*} = N - M + 1.$$

For any other classification $\mathcal{C}$ with M classes, the range of the variable Y associated with $\mathcal{C}$ can be considered as follows:

$$1 \leq Y_1 \leq Y_2 \leq \cdots \leq Y_M < N - M + 1.$$

Note that $\sum_{i=1}^{M} Y_{i*} = \sum_{i=1}^{M} Y_i = N$, $Y_{i*} \leq Y_i \ \forall i \in \{1, \ldots, M-1\}$, $Y_{M*} \geq Y_M$ and $\overline{Y_*} = \overline{Y} = \frac{N}{M}$.

Therefore, proving $CV_{Y_*} \geq CV_Y$ reduces to prove $\sigma_{Y_*} \geq \sigma_Y$, which in turn reduces to prove $\sum_{i=1}^{M} Y_{i*}^2 \geq \sum_{i=1}^{M} Y_i^2$, which is true according to Lemma 1. □

Corollary 1. *Let $M \in \mathbb{N}$ be the number of classes used to classify N individuals. The maximum value of CV_Y is:*

$$\frac{1}{N}(N - M)\sqrt{M - 1}\ . \tag{4}$$

Proof. From Proposition 1, we have that the maximum value of CV_Y is achieved when the M classes cardinalities are $1 = Y_{1*} = Y_{2*} = \cdots = Y_{M-1*} < Y_{M*} = N - M + 1$. A simple algebraic manipulation yields:

$$\sigma_{Y_*} = \frac{1}{M}(N - M)\sqrt{M - 1}$$

and therefore the maximum value for the coefficient of variation is:

$$\frac{\sigma_{Y_*}}{\overline{Y}} = \frac{1}{N}(N - M)\sqrt{M - 1}\ .$$

□

Proposition 2. *Given a number of individuals N, the maximum value of the coefficient of variation for all classifications is:*

$$max_Y(CV_Y) = \begin{cases} \frac{2N-3}{3N}\sqrt{\frac{N}{3}}, & \text{if } N \equiv 0 (mod\ 3); \\ \frac{2N-2}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 1 (mod\ 3); \\ \frac{2N-1}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 2 (mod\ 3). \end{cases} \tag{5}$$

Proof. Note that the minimum number of classes in any classification is 1 (contains all the individuals), while the maximum is N (each class contains just 1 individual). These two cases produce a value of zero for CV_Y as they are totally balanced classifications. Therefore, from now on, we assume $M \in \{2, \ldots, N-1\}$.

Let's consider the real function $f : [2, N-1] \to \mathbb{R}$:

$$f(x) = \frac{1}{N}(N-x)\sqrt{x-1},$$

defined as a real extension of (4). Figure 2 depicts function f with $N = 260$.

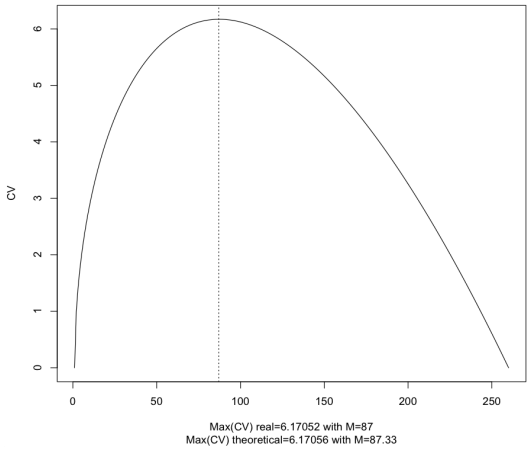

Figure 2: Function f and its maximum coefficients of variation when $N = 260$

The derivative of f is $f'(x) = \frac{N+2-3x}{2N\sqrt{x-1}}$, which is positive in $(2, \frac{N+2}{3})$ and negative in $(\frac{N+2}{3}, N-1)$. Thus, f then reaches its absolute maximum in $\frac{N+2}{3}$.

Nevertheless, the maximum value in the set $\{f(1), f(2), ..., f(N)\}$ depends on whether $\frac{N+2}{3}$ is integer or not. Let a be the integer part of $\frac{N+2}{3}$. Since $a \leq \frac{N+2}{3} < a+1$ and given that f increases in $(1, \frac{N+2}{3})$ and decreases in $(\frac{N+2}{3}, N)$, the maximum is $f(a)$ if $f(a) \geq f(a+1)$ and $f(a+1)$ if $f(a) < f(a+1)$. Simple algebraic manipulations yields:

$$max_Y(CV_Y) = \begin{cases} f(\frac{N+3}{3}) = \frac{2N-3}{3N}\sqrt{\frac{N}{3}}, & \text{if } N \equiv 0(mod\ 3); \\ f(\frac{N+2}{3}) = \frac{2N-2}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 1(mod\ 3); \\ f(\frac{N+1}{3}) = \frac{2N-1}{3N}\sqrt{\frac{N-1}{3}}, & \text{if } N \equiv 2(mod\ 3). \end{cases}$$

□

The minimum and maximum values of CV_Y are finally used to normalise and define the following *index of balanced classes*:

Definition 2. *Given a classification $\mathcal{C}$, the index of balanced classes of $\mathcal{C}$ is defined as:*

$$I_B(\mathcal{C}) = \frac{max_Y(CV_Y) - CV_{\mathcal{C}}}{max_Y(CV_Y)}, \tag{6}$$

where $CV_{\mathcal{C}}$ is the coefficient of variation associated with $\mathcal{C}$ and $max_Y(CV_Y)$ is given as per proposition 2.

The range of index I_B is $[0, 1]$ and it can be interpreted as the membership function associated with the vague concept 'balanced classification'. The higher the value of I_B the more balanced is the classification. When unbalanced classifications are preferred in a specific context, the index to use is:

$$I_{\overline{B}} = 1 - I_B\ . \tag{7}$$

2.4. Third criterion: coherent classification

The notion of adequacy of one individual to a class, which is modelled via a membership function, is used to establish the concept of 'coherent classification'. A set of P qualitative and/or quantitative descriptors $\{D_1, \ldots, D_P\}$ are defined. Each individual to be classified will be represented as $X = (x_1, \ldots, x_P)$, where x_i is the observed value of X for descriptor D_i. Given a classification $\mathcal{C}$ consisting of M classes $\{C_1, \ldots, C_M\}$, the

marginal adequacy degree of individual X to class C_i according to descriptor D_k, $\mathrm{MAD}_{C_i}(x_k)$, is defined as follows;

$$\mathrm{MAD}_{C_i}(x_k) = \mu_i^k(x_k), \tag{8}$$

where μ_i^k is the marginal distribution of descriptor D_k in the class C_i, $i \in \{1, \ldots, M\}$. The marginal adequacy degree is calculated via the density or frequency with which the specific marginal observation appears in the given class [54]. In the case of a qualitative descriptor, it is computed by taking into account the frequencies of the different modalities that the descriptor exhibits in a certain class. A density function is used if the descriptor is quantitative, in which case the height corresponding to the observed value of the individual inside the density function of the descriptor in the class is measured. The density function to chose has to be estimated for each descriptor, being the three following ones the most frequently used:

Non-parametric Bayesian estimation: This classical distribution function is based on the distribution function of a binomial variable:

$$\mathrm{MAD}_{C_i}(x_k) = {\rho_{C_i,k}}^{x_k} \cdot (1 - \rho_{C_i,k})^{(1-x_k)}, \tag{9}$$

where $\rho_{C_i,k}$ stands for the average value of descriptor D_k in class C_i, and x_k is the normalised observed value of the individual X. Figure 3 shows some examples of the Bayesian function, with different values of ρ: from 0.1 (top-left to bottom-right) to 0.9 (bottom-left to top-right). These functions have one maximum situated at the extremes (except for $\rho = 0.5$ when the distribution function is constant).

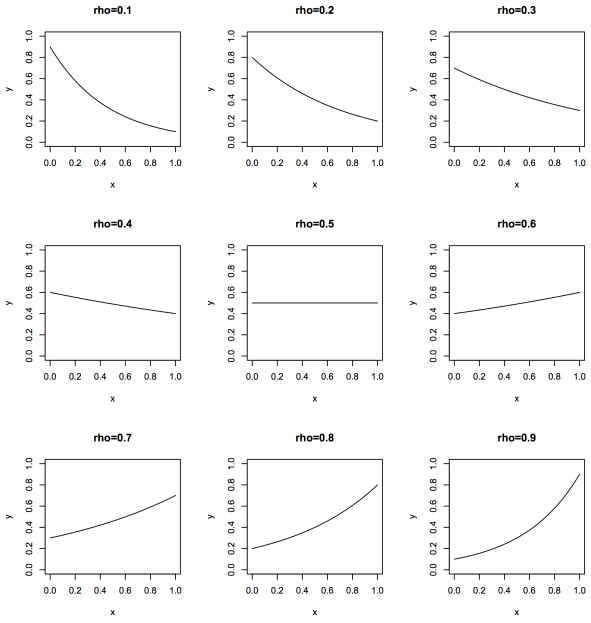

Figure 3: Bayesian function with different values of ρ

Gaussian function Given a normal distribution:

$$f(x; \rho, s^2) = \frac{1}{s\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \frac{(x-\rho)^2}{s^2}},$$

with standard deviation value s and mean value ρ, the following normalised function is used:

$$\mathrm{MAD}_{C_i}(x_k) = e^{-\frac{1}{2} \cdot \frac{(x_k - \rho_{C_i,k})^2}{s^2_{C_i,k}}}, \tag{10}$$

where $s_{C_i,k}$ and $\rho_{C_i,k}$ are the standard deviation and mean values of descriptor D_k in class C_i, while x_k is the normalised observed value of the individual X. Figure 4 shows some examples of the Gaussian function, with a fixed value of $s = 0.1$ and different values of ρ: from 0.1 (left) to 0.9 (right). It is known that the maximum value of the Gaussian functions is located in $x = \rho$.

Waissman Function: In this case, the expression to use is given as follows [55]:

$$\mathrm{MAD}_{C_i}(x_k) = \begin{cases} \rho_{C_i,k}^{d_{i,k}} \cdot (1 - \rho_{C_i,k})^{1-d_{i,k}}, & \text{if } \rho < 0.5; \\ \rho_{C_i,k}^{1-d_{i,k}} \cdot (1 - \rho_{C_i,k})^{d_{i,k}}, & \text{otherwise,} \end{cases} \tag{11}$$

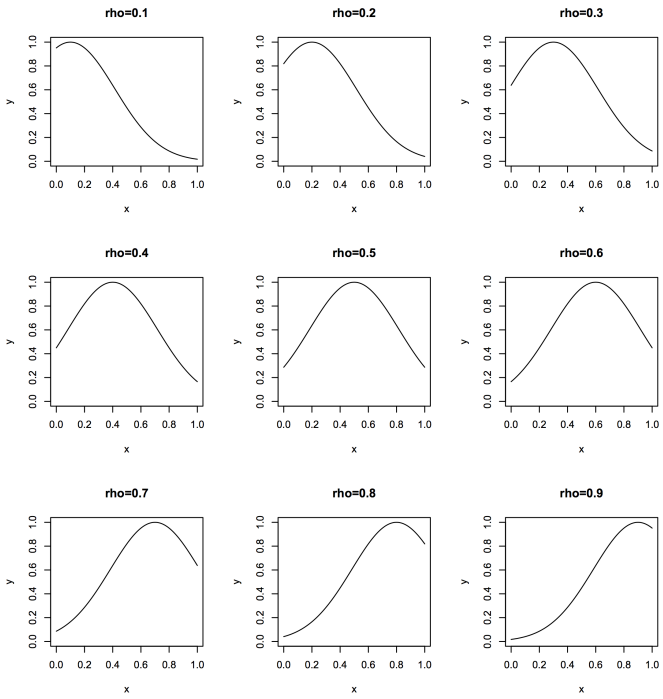

Figure 4: Gaussian function with value $s = 0.1$ and different values of ρ

where $d_{i,k}$ is the distance between the normalised value x_k to the distribution centre $c_{C_i,k}$ ($d_{i,k} = |x_k - c_{C_i,k}|$) and $\rho_{C_i,k}$ is the average value of the distances between each normalised value to the distribution centre, $\rho_{C_i,k} = \frac{\sum_k d_{i,k}}{P}$. Figure 5 shows some examples of the Waissman function with different values of ρ: from 0.1 (top-left) to 0.9 (bottom-right).

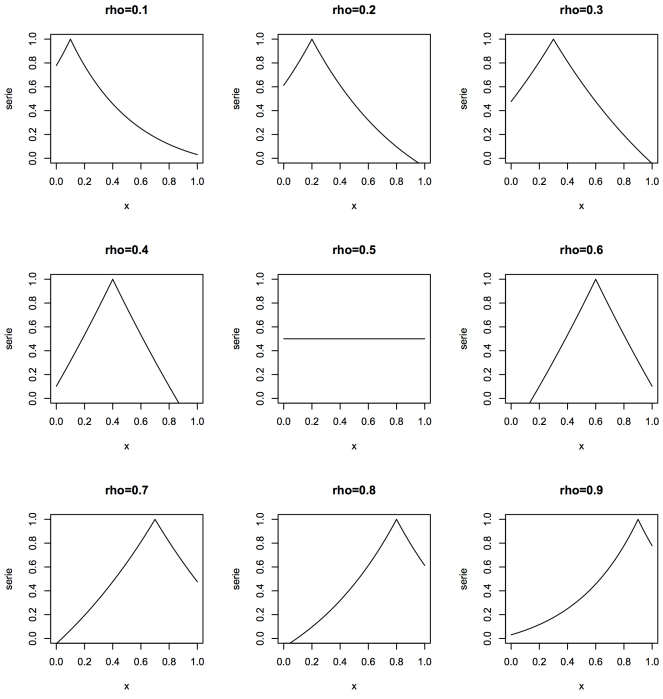

Figure 5: Waissman function with different values of ρ

A classification is considered *coherent* when the differences between the MADs are small enough for each class and each individual. The index of coherence will ensure that the global degrees of adequacy are obtained from similar values of MADs, thus reflecting the fuzzy concept of 'coherent classification'.

Let us consider μ_{ijk} as the MAD of the individual j to class i according to the descriptor k, the *index of coherence* of classification $\mathcal{C}$ is defined as the following mean of differences ($\mathrm{MD}_{\mathcal{C}}$):

$$\mathrm{MD}_{\mathcal{C}} = \frac{\sum_i \sum_j max_{k,k'}|\mu_{ijk} - \mu_{ijk'}|}{M} = \frac{\sum_i \sum_j [max(\mu_{ijk}) - min(\mu_{ijk})]}{M}. \tag{12}$$

When all MADs are equal for each individual, then the coherence index will be 0. If each individual has associated a MAD of zero (0) and a MAD of one (1) for each class then the index of coherence will be N, where N is the number of individuals. Thus, the index of coherence range is $[0, N]$. Given that the lower $\mathrm{MD}_{\mathcal{C}}$ the more coherent is the classification $\mathcal{C}$, the following inverse standardisation function is proposed as the membership function of the *index of coherence* [51].

Definition 3. *The index of coherence of classification is given as follows:*

$$I_C(\mathcal{C}) = 1 - \frac{\sum_i \sum_j [max(\mu_{ijk}) - min(\mu_{ijk})]}{M \cdot N}, \tag{13}$$

where $N \in \mathbb{N}$ is the number of individuals, $M \in \mathbb{N}$ the number of classes of classification $\mathcal{C}$, and μ_{ijk} is the MAD of the individual j to class i according to the descriptor k.

2.5. Fourth criterion: dependency on external variables

In many cases, the relevance of the classifications obtained is evaluated using external variables provided by experts, and known as control variables. The dependency or not of a classification with respect to a control variable can be tested by applying the χ^2 non-parametric test computed using the following contingency table (Table 2) with $\{C_1 \dots C_i \dots C_M\}$ representing the classes of the considered classification; $\{D_1 \dots D_s \dots D_S\}$ the values of the external variable; and q_{is} the number of observations that take the value D_s in class C_i.

Class	Descriptors or intervals				Total
	D_1	D_2	$\dots$	D_S	classes
C_1	q_{11}	q_{12}	$\dots$	q_{1S}	M_{1+}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
C_i	q_{i1}	q_{i2}	$\dots$	q_{iS}	M_{i+}
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
C_M	q_{M1}	q_{M2}	$\dots$	q_{MS}	M_{M+}
Total descriptors	M_{+1}	M_{+2}	$\dots$	M_{+S}	N

Table 2: Contingency table

It is important to note that this criterion can be used directly when the control variables are qualitative. In the case of quantitative control variables, these are previously discretised into intervals (D_s). The discretisation criterion will vary depending on the problem addressed [56–58].

Under the hypothesis of being the variable independent of the classification, the relative frequency with which members of different classes take different control variable values would not differ significantly. This hypothesis is tested using

$$\chi^2 = \sum_{i=1}^{N} \sum_{s=1}^{S} \frac{(q_{is} - e_{is})^2}{e_{is}}, \tag{14}$$

where e_{is} is the number of cases expected under the hypothesis of independence and is defined as

$$e_{is} = \frac{M_{i+} \cdot M_{+s}}{N}.$$

For each classification, the dependency of each of the control variables with respect to the classification is studied, and those classifications that have a high dependency on these external variables will be chosen. For this reason, the statistic χ^2 must have a high value.

The range of χ^2 can vary according to the number of classes of the classification. For this reason, Tschuprow's coefficient [59] is used to evaluate the degree of dependency on the control variable.

Definition 4. *Given a classification $\mathcal{C}$, its index of dependency on a control variable is defined as:*

$$I_D(\mathcal{C}) = \frac{\chi^2}{N \cdot \sqrt{M-1} \cdot \sqrt{S-1}}, \tag{15}$$

where N is the number of individuals, M is the number of classes of $\mathcal{C}$ and S is the number of values of the control variable, if it is qualitative, or the number of considered intervals in the discretisation if it is quantitative.

We note that $0 \leq I_D(\mathcal{C}) \leq 1$, and therefore this degree of dependency of a classification on a control variable could be interpreted as the membership function associated with the fuzzy concept 'dependency on a control variable'. Other possible interpretations of the value offered by this criterion rely on the concept of compatibility between the considered classification and the classification defined by the control variable.

2.6. Fifth criterion: accuracy of the predictive model

A high predictability of the model obtained from a classification ensures new individuals to be classified in the proper cluster. To this end, a criterion based on the achieved accuracy when performing supervised learning from a classification is defined.

Following the well-known concepts of precision and recall in machine learning, the fuzzy concept of accuracy of a particular classification is based on the precision and recall of the model. Firstly, precision of a class is the proportion of individuals assigned to that class that were correctly classified, while recall is the proportion of individuals of that class that have been classified in that class correctly. The precision and recall of a classification can be defined as the weighted average of the precision and recall of its classes, with weights proportional to the cardinality of the classes. The *index of accuracy* of a classification is defined as the harmonic mean of its precision and accuracy values:

Definition 5. *Given a classification $\mathcal{C}$, its index of accuracy is defined as:*

$$I_A(\mathcal{C}) = 2 \cdot \frac{precision(\mathcal{C}) \cdot recall(\mathcal{C})}{precision(\mathcal{C}) + recall(\mathcal{C})}. \tag{16}$$

The range of precision and recall is $[0, 1]$, so $0 \leq I_A(\mathcal{C}) \leq 1$ and therefore this index of accuracy could be understood as the membership function related to the fuzzy concept 'accuracy of the predictive model of the classification'.

3. Fusion of classification criteria values

In this section, a multi-criteria decision making (MCDM) approach based on a fuzzy aggregation function to summarise the given fuzzy criteria is conducted to choose the best classification among a set of feasible ones. MCDM problems normally consist of two steps [13]: *aggregation* and *exploitation*. The aggregation step consists of combining for each alternative the single evaluations into a collective evaluation in such a way that it summarises the conditions expressed in all the evaluations. The exploitation phase transforms the global evaluation of the alternatives into a ranking of the alternatives. This can be done in different ways, the most common being the use of a ranking method to obtain a score function [60–63].

Yager in [19] introduced an aggregation technique based on the Ordered Weighted Averaging (OWA) scheme. Generally speaking, the OWA operator based aggregation process consists of three steps:

(i) the first step is to re-order the input arguments in increasing order. In this way, a particular element for aggregation is not associated with a particular weight, but rather a weight is associated with a particular ordered position of an aggregated object;

(ii) the second step is to determine the weights for the operator in a proper way;

(iii) finally, the OWA weights are used to aggregate the re-ordered arguments.

Among the three steps, the first step introduces non-linearity into the aggregation process by re-ordering the input arguments, which make Yager's OWA operator significantly different from the classical linear weighted averaging operator.

Definition 6. *An OWA operator of dimension n is a mapping $\phi\colon \mathbb{R}^n \to \mathbb{R}$, which has a set of weights $W = (w_1, \cdots, w_n)^T$ associated with it, so that $w_i \in [0,\ 1]$ and $\sum_{i=1}^{n} w_i = 1$,*

$$\phi_W(a_1, \cdots, a_n) = \sum_{i=1}^{n} w_i a_{\sigma(i)} \tag{17}$$

where σ is a permutation function such that $a_{\sigma(i)}$ is the i-th highest value in the set $\{a_1, \cdots, a_n\}$.

The OWA operator exhibits the following desirable properties for an aggregation operation:

1. It is commutative:
$$\phi_W\left(p_{\sigma(1)}, \ldots, p_{\sigma(n)}\right) = \phi_W\left(p_1, \ldots, p_n\right),$$
being σ any permutation of the set $\{1, \ldots, n\}$.

2. It is an *or-and* operator, i.e., it is located between the minimum and the maximum of the arguments to be aggregated:
$$min(a_i) \leq \phi_W(a_1, \cdots, a_n) \leq max(a_i).$$

3. It is idempotent:
$$\phi_W(a, \cdots, a) = a.$$

4. It is monotonic:
$$\phi_W(a_1, \cdots, a_n) \geq \phi_W(e_1, \cdots, e_n), \text{ if } a_i \geq e_i \ \forall i.$$

An issue in the definition of the *OWA* operator is how to obtain the associated weighting vector [19]. In [19] we can find two ways to do this. The first approach is to use a learning mechanism using some sample data; the second approach is to provide some semantics or meaning to the weights. The latter approach enables applications in the area of quantifier guided aggregations [64, 65].

In the process of quantifier guided aggregation, given a collection of n criteria represented as fuzzy subsets of the alternatives X, the OWA operator has been used to implement the concept o fuzzy majority in the aggregation phase by means of a *fuzzy linguistic quantifier* [21] that indicates the proportion of satisfied criteria 'necessary for a good solution' [22]. This implementation is done by using the quantifier to calculate the OWA weights.

Definition 7. *Given a function $Q : [0, 1] \to [0, 1]$ such that $Q(0) = 0$, $Q(1) = 1$ and if $x > y$ then $Q(x) \geq Q(y)$, an OWA aggregation operator guided by Q is given as [19]:*

$$\phi_Q(a_1, \ldots, a_n) = \sum_{i=1}^{n} w_i \cdot a_{\sigma(i)},$$

being $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ a permutation such that $a_{\sigma(i)} \geq a_{\sigma(i+1)}$, $\forall i = 1, \ldots, n-1$, i.e., $a_{\sigma(i)}$ is the i-th largest value in the set $\{a_1, \ldots, a_n\}$; and

$$w_i = Q\left(\frac{i}{n}\right) - Q\left(\frac{i-1}{n}\right), i = 1, \ldots, n. \tag{18}$$

These Q functions are called Basic Unit-interval Monotone (BUM) functions in [66] and 'are particularly useful in situations in which the imperative guiding the OWA aggregation is expressed linguistically by a quantifier'. We make note that in [22] BUM functions are called Regular Increasing Monotone (RIM) quantifiers.

Example 1. *The aggregation of the set of values $\{0.5, 0.07, 0.228, 0.057, 0.482\}$ using an OWA operator guided by the fuzzy linguistic quantifier 'most of' represented via the RIM function $Q(r) = r^{1/2}$ [12], whose corresponding weighting vector using (18) is $(0.447, 0.185, 0.142, 0.120, 0.106)$, yields*

$$\begin{aligned} \phi_{most\ of}(0.5, 0.07, 0.228, 0.057, 0.482) = & \quad 0.447 \cdot 0.5 + 0.185 \cdot 0.482 + 0.142 \cdot 0.228 + 0.129 \cdot 0.07 + 0.106 \cdot 0.057 \\ = & \quad 0.360. \end{aligned}$$

This collective value is interpreted as the value up to which 'most of' the criteria are verified.

This type of aggregation 'is very strongly dependent upon the weighting vector used' [22], and consequently upon the function expression used to represent the fuzzy linguistic quantifier. The RIM function used in this example guarantees that all the individual valuations contribute to the final aggregated value because it is a strictly increasing function. Moreover, the higher the ranking of a value, the higher the weighting value associated with it. This is a consequence of the concavity property – which was proven in [12] to make a RIM function appropriate for conducting aggregation processes in heterogeneous decision-making problems.

4. Marketing case study

In this section, we describe a case study addressing a marketing problem, and solve it by using the methodology introduced above. The case presented shows the relation between the theoretical study done in previous sections and its connection to real applications. The study took place in a business to business (B2B) environment over nine months, where information about the retailers of a commercial firm was provided by the firms' sales representatives. B2B environments are characterized by marketing activities of organizations exchanging commerce transactions with other organizations [67]. These types of environments arise when a firm distributes its products via other firms (retailers).

The main objective of our study was to identify and then segment a set of retailers (points of sales) of an industrial company, considering behavioural, relational and descriptive variables. The presented methodology is used to select the best market segmentation according to marketing experts and firm expectations. The new segmentation obtained will give an opportunity to marketing executives and managers to understand their customers' behaviour. In addition, it will enable them to design or define appropriate and common marketing strategies for each segment.

4.1. Dataset

The study conducted is based on data collected using the observations, knowledge, and experience of the sales representatives working for Textil Seu SA, an outdoor sporting equipment firm (Grifone, http://www.grifone.com) established in La Seu d'Urgell (in Catalonia, Spain) for more than 25 years. Grifone works in a B2B environment and distributes clothes through points of sale and not directly to customers.

This section presents the results obtained from a database of information from 260 shops that distribute Grifone products [8]. According to marketing experts, 16 variables were used to describe these points of sale (3 quantitative, 5 qualitative, and 8 qualitative ordinal, as presented in Table 3). Consequently, each of the points of sale is described by a vector of dimension 16. It should be noticed that the variable *promotions sensitivity* was not used in the unsupervised learning process.

Type	Number	Description
Quantitative	3	*Duration of commercial relationship* *Number of full-time sales assistants* *Assessment by Grifone representatives*
Qualitative	5	*Specialist store* *Geographic location* *Grifone products in the display window* *Thermal product display* *Use of the Internet for e-Commerce*
Ordinal	8	*Level of competition* *Store size* *Store maintenance* *Display window size* *Communicative quality* *Aesthetics quality* *Grifone products' importance* *Promotions sensitivity*

Table 3: Description of variables

4.2. Obtaining segmentations

The unsupervised learning technique used is based on the algorithm LAMDA [68–70]. LAMDA is based on fuzzy hybrid connectives, and employs the interpolation capabilities of logic operators over fuzzy environments [71]. A linearly compensated hybrid connective, considered as an interpolation between a t-norm T and its dual t-conorm T^* is used:

$$H = (1-\beta)T + \beta T^*,$$

where $\beta \in [0,1]$ is known as the level of tolerance of the classification. It can be noted that for $\beta = 0$, the t-norm is obtained, and for $\beta = 1$, the t-conorm is the result.

As a result of the unsupervised learning process, 566 segmentations were obtained with a number of classes between 1 and 303. The used hybrid connectives were minMax, Frank, probabilistic sum and product, and Lukasiewicz. Table 4 shows the number of segmentations obtained through the use of each hybrid connective.

MinMax	**Frank**	**Prob.**	**Lukasiewicz**
244	303	18	1

Table 4: Number of segmentations obtained

4.3. Ranking and selecting segmentations

The criteria defined in previous sections were applied to choose the most appropriate points of sale segmentation according to them. In this marketing environment, the most desirable number of classes is set between three and five [53] and therefore $K_1 = 3$ and $K_2 = 5$ in membership function (2). Balanced classes were required and therefore (6) was used to compute the balanced index with the second part of (5) because $N = 260 \equiv 2(mod\ 3)$. Figure 6 shows the histogram of the quantitative variables used in the third criterion *duration of commercial relationship*, *number of full-time sales assistants* and *assessment by Grifone representatives*, and the graphs of their associated density functions: Waissman, Bayesian and Gaussian, respectively. The variable *promotion*

sensitivity has been used as the control variable to contribute in the computation of the fourth index. Finally, a supervised learning process is performed on the obtained segmentations. This step involves partitioning the dataset by means of a cross-validation process with 10 folds. Support Vector Machines are considered for supervised learning due to their good performance on high dimensional spaces [72].

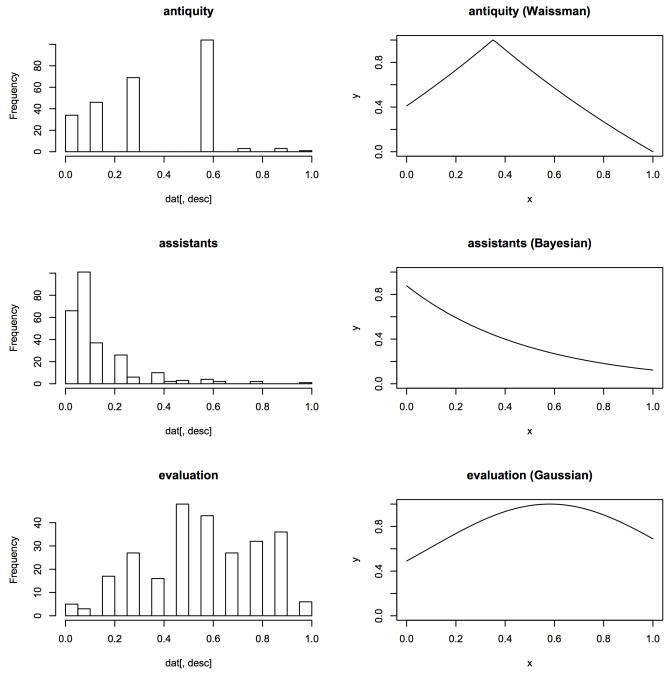

Figure 6: Histogram of the quantitative descriptors and their chosen density functions

Once the five indexes have been computed for each of the segmentations, they were aggregated using an OWA operator guided by the fuzzy linguistic quantifier 'most of', represented using the RIM function $Q(r) = r^{1/2}$, which has associated the following weighting vector is $(0.447, 0.185, 0.142, 0.120, 0.106)$. Table 5 shows an extract of the best segmentations obtained using this methodology.

ID	**Conn.**	**Tol.**	**M**	I_U	I_B	I_C	I_D	I_A	**OWA**
#259	Minmax	0.439	3	1	0.928	0.251	0.528	0.936	0.8423
#260	Minmax	0.469	3	1	0.929	0.255	0.363	0.922	0.8208
#258	Minmax	0.422	4	1	0.885	0.226	0.425	0.875	0.8103
#243	Minmax	0.290	3	1	0.909	0.257	0.008	0.965	0.7868
#244	Minmax	0.304	3	1	0.920	0.256	0.012	0.948	0.7856
#257	Minmax	0.411	3	1	0.933	0.280	0.032	0.856	0.7787
#256	Minmax	0.400	4	1	0.872	0.246	0.064	0.906	0.7752
#253	Minmax	0.359	4	1	0.883	0.232	0.021	0.915	0.7722
#252	Minmax	0.352	4	1	0.884	0.238	0.025	0.904	0.7714
#255	Minmax	0.393	4	1	0.939	0.264	0.063	0.768	0.7687

Table 5: Extract of the best segmentations using fuzzy selection criteria OWA methodology

4.4. Class description and managerial implications

The three classes obtained in segmentation #259 are described below:

Class 1: Consists of 35 points of sale and includes shops with a long-standing commercial relationship with Grifone. By and large, these shops are not specialists in mountain gear while competition between them is generally intense. These points of sale are located in cities or non-mountainous areas, they are medium or large in size and employ many shop assistants. The shops are well-maintained, have a medium-sized window display, and their aesthetic and communicative qualities and abilities are high.

Many of the shops in this group have a thermal product display and usually market their goods on the Internet, which is rather unusual for this type of shop. The importance of the Grifone brand is secondary, in general, and their clients demonstrate a mid-level sensitivity to promotions.

In short, Class 1 might correspond to multi-sports shops, with large points of sale not found in mountain locations.

Class 2: Most shops in this class (98 in total) do not have a long-standing commercial relationship with Grifone. Competitiveness is medium, they are small or medium in size, and have few sales staff. Their maintenance, aesthetic quality, and communicative abilities are generally average or good, and they have a moderately small display window.

Most shops in Class 2 do not display Grifone products; and the importance of Grifone is minor. Almost none have a thermal product display and Grifone representatives give these shops the worst evaluation. Customers of these shops demonstrate a low or mid-level sensitivity to promotions.

It seems that from Grifone's point of view, Class 2 are the worst shops and where its brand is worst placed.

Class 3: This is the largest class with 127 shops. They are usually sector specialists, with fairly strong competition between them, and, perhaps because of this factor, they are primarily located in mountainous populations.

The shops are not large, nor do they have many shop assistants, but they are well-maintained, and have excellent aesthetic and communicative qualities. Normally, they have Grifone clothing in the display window, although they usually do not have a thermal product display.

The importance given to the Grifone brand is usually the highest; Grifone is often the principal product. Their clients are quite sensitive to promotions.

Class 3 shops are Grifone's favourite clients: small elegant specialists in mountain gear. In this segment, shops with a long-standing commercial relationship with Grifone are mixed with others with a more recent relationship. The latter could potentially become favourite shops and be the first possible target of a marketing campaign.

According to the methodology described by Sánchez-Hernández et al. [8], where the criteria are applied as filters, some of the top-10 segmentations of Table 5, obtained with the present methodology, would have been discarded. For instance, the segmentation classes third best (number 258) could have been discarded due to the low value obtained in the index I_B. Also the segmentations classed fourth to tenth should have been discarded due to their limited relation with the control variable 'promotions sensitivity'. In conclusion, the new approach avoids discarding segmentations that could be useful for marketing experts when observed globally.

4.5. Discussion

The real case study presented above illustrates how the proposed methodology can deal with the ambiguity that appears when managing multi-criteria associated to fuzzy concepts. Our approach covers almost all the validation concepts considered in the literature given in Table 1, while other approaches in clustering validation, in general, take into account only some of these concepts.

The proposed criteria are inspired on well-known concepts for clustering validation [6, 43–45]. The *useful number of classes* and *balanced classes* criteria have a marketing background, since they were defined to guarantee manageable segmentations [53]. However, their implementations (I_U and I_B) in a fuzzy environment is a specific contribution of this paper. The *coherence* criterion measures the compactness and separability of a given segmentation, that are common measures for clustering validation. Its implementation (I_C) is defined in a novel way via a normalised distribution function [70]. Regarding the *dependency* criterion, there are different approaches in the literature to estimate the compatibility between the analysed segmentations and an *a priori* segmentation or an external variable. In the methodology presented, the proposed index I_D relies on the concept of dependency given by a χ^2 distribution. Finally, the *accuracy* criterion and its associated index I_A have been defined as an aggregation of the widely-known recall and precision indicators [6, 46, 52].

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#272	Minmax	0.958	2	0.5	0.995	0.284	0.010	0.802	0.6997	74
#189	Minmax	0.782	3	1	0.992	0.278	0.018	0.168	0.6925	124
#172	Minmax	0.772	3	1	0.990	0.284	0.008	0.400	0.7224	46

Table 6: Best segmentations according to I_B

Table 6 shows the best segmentations looking exclusively at the values obtained from the *balanced classes* criterion. Although in this study it is desirable to obtain a balanced segmentation, it is not indispensable, as

indicated by the fact that best overall segmentations (#259, #260 and #258) perform below segmentations #272, #189 and #172. This is obviously due to this criterion discarding the most unbalanced segmentations.

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#214	Minmax	0.904	2	0.5	0.914	0.297	0.006	0.795	0.6633	234
#246	Minmax	0.911	3	1	0.921	0.290	0.013	0.770	0.7635	19
#248	Minmax	0.918	2	0.5	0.989	0.290	0.011	0.843	0.7052	61

Table 7: Best segmentations according to I_C

The best segmentations according to the *coherence* criterion are included in Table 7. Values obtained by this criterion are fairly homogeneous and therefore exhibiting a low influence in the process of selecting the best segmentation in the methodology presented here. Indeed, the classifications discarded with the present methodology are because of their low performance in other criteria, as illustrated by segmentation #246, which performs very well in all criteria but I_D.

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#259	Minmax	0.439	3	1	0.928	0.251	0.528	0.936	0.8423	1
#258	Minmax	0.422	4	1	0.885	0.226	0.425	0.875	0.8103	3
#260	Minmax	0.469	3	1	0.929	0.255	0.363	0.922	0.8208	2

Table 8: Best segmentations according to I_D

The best three segmentations with the present methodology are kept using the *dependency* criterion and shown in Table 8, which indicated a high significance of this criterion in this methodology. However, it is clear that rank reversal phenomenon occurs here [73] as segmentation #258 is ranked lower than segmentation #260 in the overall ranking due to its lower performance in the other criteria.

ID	Conn.	Tol.	M	I_U	I_B	I_C	I_D	I_A	OWA	rank
#003	Frank (0.5)	0.055	2	0.5	0.841	0.197	0.001	0.988	0.6926	121
#002	Frank (1.5)	1	2	0.5	0.841	0.197	0.001	0.988	0.6926	122
#001	Frank (3)	0.5	2	0.5	0.842	0.151	0.001	0.988	0.6873	160

Table 9: Best segmentations according to I_A

Table 9 includes the best segmentations according to their *accuracy*. The facts that these segmentations do not show a high overall rank, and the best overall segmentations in Table 5 exhibit a very high accuracy degree illustrates that in this case study a great number of segmentations that are quite accurate also perform well in the rest of criteria.

The present methodology, on the one hand, avoids the predefinition of arbitrary thresholds to decide which segmentations are taken into account in the application of subsequent criteria. On the other hand, it is clear that the sequential application of the above set of criteria would have prevented any of the three best overall classifications obtained with the methodology to have been ranked in those positions. Therefore, the case study clearly illustrates that the methodology presented avoids discarding segmentations that could be potentially useful for marketing experts when observed globally.

5. Conclusions

In this paper, an unsupervised learning methodology has been employed to automatically generate a set of classifications. A set of fuzzy criteria has been proposed, analysed and modelled using a set of indexes to evaluate the obtained classifications. The properties and usability of the defined criteria have been explained and proven. The indexes were proposed to be aggregated using an OWA operator guided by a fuzzy linguistic quantifier that is used to implement in the process the concept of fuzzy majority. This methodology has been applied to a real marketing case based on a B2B environment, and an analysis of the managerial implications of the proposed methodology solution was provided.

Future work is oriented towards:

- Using the defined criteria to assess unsupervised learning techniques.

- Defining and analysing other selection criteria, as well as other aggregation operators for specific real situations.
- Defining indexes for selection criteria with fuzzy values and their aggregation using fuzzy operators.
- Designing and developing a series of tools that will provide an automatic qualitative description of the chosen classification.
- Applying the considered methodology to other real problems.

Acknowledgements

This research paper has been partially conducted during a three-months visiting period to the Centre for Computational Intelligence (CCI) at De Montfort University in Leicester (UK). This work is supported by the SENSORIAL Research Project (TIN2010-20966-C02-01, 02), funded by the Spanish Ministry of Science and Information Technology.

The participation and interest of the firm Textil Seu SA in the SENSORIAL project is also gratefully acknowledged.

References

[1] R. Duda, P. Hart, D. Stork, Pattern classification (2nd ed.), John Wiley & Sons, New York, 2001.

[2] M. Figueiredo, A. K. Jain, Unsupervised learning of finite mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 381–396.

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

[4] A. Z. Broder, P. Ciccolo, M. Fontoura, E. Gabrilovich, V. Josifovski, L. Riedel, Search advertising using web relevance feedback, in: Proceedings of the 17th ACM conference on Information and knowledge management, New York, NY, 2008, pp. 1013–1022.

[5] M. Kukar, Transductive reliability estimation for medical diagnosis, Artificial Intelligence in Medicine 29 (1-2) (2003) 81–106.

[6] K.-M. Osei-Bryson, Towards supporting expert evaluation of clustering results using a data mining process model, Information Sciences 180 (2010) 414–431.

[7] D. H. Choi, B. S. Ahn, S. H. Kim, Prioritization of association rules in data mining: Multiple criteria decision approach, Expert Systems with Applications 29 (4) (2005) 867–878.

[8] G. Sánchez-Hernández, N. Agell, J. C. Aguado, M. Sánchez, F. Prats, Selection criteria for fuzzy unsupervised learning: Applied to market segmentation, in: Lecture Notes in Computer Science, Vol. 4529, Springer-Verlag, Berlin, 2007, pp. 307–317.

[9] J. Sánchez Almeida, J. A. L. Aguerri, C. Muñoz Tuñon, A. de Vicente, Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra, The Astrophysical Journal 714 (1) (2010) 487–504.

[10] D. Dubois, H. Prade, A review of fuzzy set aggregation connectives, Information Sciences 36 (1985) 85–121.

[11] F. Chiclana, E. Herrera-Viedma, F. Herrera, A. Alonso, Induced ordered weighted geometric operators and their use in the aggregation of multiplicative preference relations, International Journal of Intelligent Systems 19 (2004) 233–255.

[12] F. Chiclana, E. Herrera-Viedma, F. Herrera, A. Alonso, Some induced ordered weighted averaging operators and their use for solving group decision-making problems based on fuzzy preference relations, European Journal of Operational Research 182 (2007) 383–399.

[13] J. Fodor, M. Roubens, Fuzzy Preference Modelling and Multicriteria Decision Support, Kluwer, 1994.

[14] F. Herrera, E. Herrera-Viedma, F. Chiclana, A study of the origin and uses of the ordered weighted geometric operator in multicriteria decision making, International Journal of Intelligent Systems 18 (2003) 689–707.

[15] G. J. Klir, T. A. Folger, Fuzzy Sets, Uncertainty and Information, Prentice-Hall, 1988.

[16] V. Torra, The weighted OWA operator, International Journal of Intelligent Systems 12 (1997) 153–166.

[17] V. Torra, Y. Narukawa, A view of averaging aggregation operators, IEEE Transactions on Fuzzy Systems 16 (2007) 1063–1067.

[18] Z. S. Xu, Q. L. Da, An overview of operators for aggregating information, International Journal of Intelligent Systems 18 (2003) 953–969.

[19] R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decision-making, IEEE Transactions on Systems, Man, and Cybernetics 18 (1988) 183–190.

[20] S.-M. Zhou, F. Chiclana, R. I. John, J. M. Garibaldi, Type-1 OWA operators for aggregating uncertain information with uncertain weights induced by type-2 linguistic quantifiers, Fuzzy Sets and Systems 159 (2008) 3281–3296.

[21] L. A. Zadeh, A computational approach to fuzzy quantifiers in natural languages, Computational & Applied Mathematics 9 (1983) 149–184.

[22] R. R. Yager, Quantifier guided aggregation using OWA operators, International Journal of Intelligent Systems 11 (1996) 49–73.

[23] J.-L. Marichal, Aggregation operators for multicriteria decision aid, Ph.D. thesis, Institute of Mathematics, University of Liège, Liège, Belgium (1998).

[24] Y. Chen, J. Z. Wang, R. Krovetz, Clue: cluster-based retrieval of images by unsupervised learning, IEEE Transactions on Image Processing 14 (2005) 1187–1201.

[25] M. Elati, C. Rouveirol, Unsupervised learning for gene regulation network inference from expression data: A review, in: M. Elloumi, A. Y. Zomaya (Eds.), Algorithms in Computational Molecular Biology: Techniques, Approaches and Applications, John Wiley & Sons, Inc., 2011, pp. 955–978.

[26] S. Constantinos, A. Paris, An application of supervised and unsupervised learning approaches to telecommunications fraud detection, Knowledge-Based Systems 21 (7) (2008) 721–726.

[27] E. Hadavandi, H. Shavandi, A. Ghanbari, Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting, Knowledge-Based Systems 23 (8) (2010) 800–808.

[28] J. Goldsmith, Unsupervised learning of the morphology of a natural language, Computational Linguistics 27 (2) (2001) 153–198.

[29] C. H. Lee, H. C. Yang, Construction of supervised and unsupervised learning systems for multilingual text categorization, Expert Systems with Applications 36 (2009) 2400–2410.

[30] J. C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal, International Journal of Computer Vision 79 (3) (2008) 299–318.

[31] J. L. Oliver, L. Tortosa, J. F. Vicent, A neural network model to develop actions in urban complex systems represented by 2d meshes, International Journal of Computer Mathematics 88 (2011) 3361–3379.

[32] Y. Yang, Y. Liao, G. Meng, J. Lee, A hybrid feature selection scheme for unsupervised learning and its application in bearing fault diagnosis, Expert Systems with Applications 38 (9) (2011) 11311–11320.

[33] D. Ferraretti, G. Gamberoni, E. Lamma, Unsupervised and supervised learning in cascade for petroleum geology, Expert Systems with Applications 39 (10) (2012) 9504–9514.

[34] J. Barrón-Adame, M. Cortina-Januchs, A. Vega-Corona, D. Andi, Unsupervised system to classify so_2 pollutant concentrations in salamanca, Expert Systems with Applications 39 (1) (2012) 107–116.

[35] C.-Y. Chiu, Y.-F. Chen, I.-T. Kuo, H. C. Ku, An intelligent market segmentation system using k-means and particle swarm optimization, Expert Systems with Applications 36 (3, Part 1) (2009) 4558–4565.

[36] T. Hong, E. Kim, Segmenting customers in online stores based on factors that affect the customer's intention to purchase, Expert Systems with Applications 39 (2012) 2127–2131.

[37] J. Mo, M. Y. Kiang, P. Zou, Y. Li, A two-stage clustering approach for multi-region segmentation, Expert Systems with Applications 37 (10) (2010) 7120–7131.

[38] Z. Yao, A. H. Holmbom, T. Eklund, B. Back, Combining unsupervised and supervised data mining techniques for conducting customer portfolio analysis, in: Advances in Data Mining: Applications and Theoretical Aspects, Vol. 6171 of Lecture Notes in Artificial Intelligence, 2010, pp. 292–307.

[39] Z. Lu, S. Wang, X. Li, L. Yang, D. Yang, D. Wu, Online shop location optimization using a fuzzy multicriteria decision model – case study on taobao.com, Knowledge-Based Systems 32 (2012) 76–83.

[40] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0 (2011).
URL `http://www.R-project.org`

[41] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM Computing Surveys 31 (2009) 264–323.

[42] S. Theodoridis, K. Koutroumbas, Pattern Recognition, 4th Edition, Academic Press, Cambridge, United Kingdom, 2008.

[43] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 2 (2001) 107–145.

[44] Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu, Understanding of internal clustering validation measures, in: IEEE 10th International Conference on Data Mining (ICDM), 2010, 2010, pp. 911–916.

[45] I. Yatskiv, L. Gusarova, The methods of cluster analysis results validation, in: Transport and Telecommunication, 2005, pp. 75–80.

[46] R. Tibshirani, G. Walther, Cluster validation by prediction strength, Journal of Computational and Graphical Statistics 15 (3) (2005) 511–528.

[47] R. M. Bittmann, R. Gelbard, Visualization of multi-algorithm clustering for better economic decisions - the case of car pricing, Decision Support Systems 47 (1) (2009) 42–50.

[48] M. Ramze Rezaee, B. Lelieveldt, J. Reiber, A new cluster validity index for the fuzzy c-mean, Pattern Recognition Letters 19 (3-4) (1998) 237–246.

[49] C.-H. Cheng, A. Wai-chee Fu, Y. Zhang, Entropy-based subspace clustering for mining numerical data, in: KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 84–93.

[50] H. Wang, M. Ding, X. Li, B. Shen, Clinical information driven ensemble clustering for inferring robust tumor subtypes, in: 2nd International Conference on Biomedical Engineering and Informatics, 2009, IEEE, 2009, pp. 1–4.

[51] J. Wu, H. Xiong, J. Chen, Adapting the right measures for k-means clustering, in: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, ACM Press, 2009, pp. 877–886.

[52] H. Xiong, J. Wu, J. Chen, K-means clustering versus validation measures: a data-distribution perspective, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 39 (2) (2009) 318–331.

[53] M. Casabayó, Shopping behaviour forecasts: Experiments based on a fuzzy learning technique in the spanish food retailing industry, Ph.D. thesis, University of Edinburgh (2005).

[54] J. Aguilar-Martin, N. Agell, M. Sánchez, F. Prats, Analysis of tensions in a population based on the adequacy concept, in: Topics in Artificial Intelligence, 5th Catalonian Conference on AI, CCIA 2002, Castellón, Spain, October 24-25, 2002, Proceedings, Vol. 2504 of Lecture Notes in Computer Science, Springer, 2002, pp. 17–28.

[55] J. Waissman, J. Aguilar, B. Dahhou, G. Roux, Généralisation du degré d'adequation marginale dans la méthode de classification LAMDA, in: 6èmes Rencontres de la Société Francophone de Classification, Montpellier, France, 1998.

[56] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Proceedings of the 12th International Conference on Machine Learning, San Francisco, CA, 1995, pp. 194–202.

[57] L. A. Kurgan, K. J. Cios, CAIM discretization algorithm, IEEE Transactions on Knowledge and Data Engineering 16 (2004) 145–153.

[58] F. Ruiz, C. Angulo, N. Agell, IDD: A supervised interval distance-based method for discretization, IEEE Transactions on Knowledge and Data Engineering 40 (2008) 1230–1238.

[59] A. A. Tschuprow, M. Kantorowitcsch, Principles of the mathematical theory of correlation, Journal of the American Statistical Association 34 (1939) 755.

[60] F. Chiclana, F. Herrera, E. Herrera-Viedma, Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations, Fuzzy Sets and Systems 97 (1) (1998) 33–48.

[61] Z. Zhang, C. Guo, A method for multi-granularity uncertain linguistic group decision making with incomplete weight information, Knowledge-Based Systems 26 (2012) 111–119.

[62] L. Zhou, H. Chen, A generalization of the power aggregation operators for linguistic environment and its application in group decision making, Knowledge-Based Systems 26 (2012) 216–224.

[63] S. Gramajo, L. Martínez, A linguistic decision support model for QoS priorities in networking, Knowledge-Based Systems 32 (2012) 65–75.

[64] R. R. Yager, Quantifiers in the formulation of multiple objective decision functions, Information Sciences 31 (1983) 107–139.

[65] Z. Pei, D. Ruan, J. Liu, Y. Xu, A linguistic aggregation operator with three kinds of weights for nuclear safeguards evaluation, Knowledge-Based Systems 28 (2012) 19–26.

[66] R. R. Yager, Induced aggregation operators, Fuzzy Sets and Systems 137 (2003) 59–69.

[67] P. W. Turnbull, S. Leek, Business-to-business marketing: Organizational buying behavior, relationships and networks, in: M. Market (Ed.), The marketing book, Butterworth-Heinemann, Oxford, UK, 2003, pp. 142–169.

[68] J. C. Aguado, A mixed qualitative-quantitative self-learning classification technique applied to situation assessment in industrial process control, Ph.D. thesis, Universitat Politècnica de Catalunya (1998).

[69] J. C. Aguado, A. Catalá, X. Parra, Comparison of structure and capabilities between a non-standard classificarion technique and the radial basis bunction neural networks, in: Proceedings of the 13th European Simulation Multiconference (ICQFN 99), Vol. II, Warsaw, Poland, 1999, pp. 442–448.

[70] J. Aguilar, R. López de Mántaras, The process of classification and learning the meaning of linguistic descriptors of concepts, Approximate Reasoning in Decision Analysis (1982) 165–175.

[71] J. Klir, B. Yuan, Fuzzy sets and fuzzy logic, Theory and Applications, Prentice Hall, 1995.

[72] O. Chapelle, P. Haffner, V. Vapnik, Support vector machines for histogram-based image classification, IEEE Transactions on Neural Networks 10 (1999) 1055–1064.

[73] E. Triantaphyllou, Multi-Criteria Decision Making Methods: A Comparative Study, Applied Optimization, Springer, 2000.