Hierarchical clustering of modal ordinal symbolic data objects

Carmen Bravo*, José M. García-Santesmases**

* Universidad Complutense de Madrid, Servicio Informático de Apoyo al Usuario - Investigación, Edificio Vicerrectorado de Alumnos, 28040 Madrid, Spain. mcbravo@ucm.es
** Universidad Complutense de Madrid, Dpto. de Estadística e Inv. Operativa Facultad de Ciencias Matemáticas, 28040 Madrid, Spain. josemgar@ucm.es

Abstract. The problem of analyzing the dispersion of a set of objects described by ordinal modal symbolic data is addressed in order to obtain homogeneous groups, which are evaluated by a consensus measure. Based on a generalized φ function a consensus measure for objects and for sets of objects described by modal ordinal data is defined. A variability measure for sets of subsets of objects based in the consensus measure of their members is proposed. A dissimilarity measure between objects and between set of objects based on this consensus variability measure is also given. It is proven that the Leik consensus measure is a φ function. An ascending hierarchical clustering algorithm is presented. The criterion to be minimized in each step is based on the decrease of the consensus variability. An example with modal ordinal data of 34 teachers that were evaluated by their students is presented.

Introduction

This paper proposes an ascending hierarchical clustering algorithm for modal ordinal symbolic data using a dissimilarity measure based on consensus variability. One common meaning of consensus is a general agreement among the members of a given group and can be seen as a function of shared team feelings towards an issue. A common way to analyse it is to use consensus measures to evaluate the strength of consensus in a class of individuals. As introduced by Leik (1966) the conception of consensus is simply a lack of dispersion, and a consensus measure provides a way to measure the dispersion in ordinal scales. In García-Santesmases and Bravo (2010) and García-Santesmases et al. (2010), three specific consensus measures for groups of individuals based on a single issue are given and are proved that satisfy the requirements given by Tastle (2005). They are extended to several issues and to symbolic data. In this paper our main contribution is to give a characterization of a consensus measure for a group of individuals based on a single issue through the introduction of a φ function with some properties. This characterization covers all the requirements given by Tastle (2005) for a consensus measure to be considered viable. It can be used to build consensus measures for symbolic data objects and sets, variability consensus measures and distances based on these variabilities. The present work does not make any assumptions regarding the ordinal scale. Related work on hierarchical clustering for modal symbolic categorical data, although not modal ordinal data, can be found in Kim and Billard (2012). Chavent (2000) proposes a divisive clustering algorithm considering the ordinal property of modal ordinal symbolic variables to split a partition at each step. Nevertheless, variability for modal ordinal variables is not addressed.

In section 1 basic concepts and notation are introduced, including the definition of the φ function. Section 2 introduces consensus measures for a single object as well as for a set of objects described by modal ordinal data. This section gives also the variability of a partition of a set of modal ordinal symbolic data objects and introduces consensus variability. In section 3 it is proven that the Leik measure is a φ function. Section 4 gives and agglomerative hierarchical clustering based on the minimum decrease of consensus variability. Some indexes for cluster and partition interpretation are also introduced. Section 5 gives an example.

1 Basic concepts and notation

In this section we present the input data and give the definition of the φ function with specific properties to build from it consensus measures for modal ordinal symbolic data objects and for sets of modal ordinal symbolic data objects.

1.1 Input data

Let $\mathfrak{S}$ be a set of objects described by symbolic modal ordinal variables y_j with domain on an ordinal scale $Y_j = \left\{r_1^j, r_2^j, ..., r_{k_j}^j\right\}$, $j = 1, ..., p$ and let $\mathcal{S} = \{s_1, s_2, ..., s_m\} \subset \mathfrak{S}$ be a subset of elements of $\mathfrak{S}$. The description of s_i is given by $(D_i^1, ..., D_i^p)$ with $D_i^{j^\intercal} = (r_1^j(w_{1i}^j), r_2^j(w_{2i}^j), ..., r_{k_j}^j(w_{k_j i}^j))$, where $\mathbf{w}_i^{j^\intercal} = (w_{1i}^j, w_{2i}^j, ..., w_{k_j i}^j)$ ($\sum\limits_{l=1,...,k_j} w_{li}^j = 1$) are the probability or weight values associated to $r_1^j, r_2^j, ..., r_{k_j}^j$. These values may represent the distribution of the ratings of individual preferences of a group of individuals (see Bock and Diday, 2000).

Let q_{s_i} be the relative weight of s_i in $\mathcal{S}$, $\sum\limits_i q_{s_i} = 1$, and

$$\mathbf{W} = \begin{pmatrix} w_{11}^1 & ... & w_{k_1 1}^1 & ... & ... & w_{11}^p & ... & w_{k_p 1}^p \\ w_{12}^1 & ... & w_{k_1 2}^1 & ... & ... & w_{12}^p & ... & w_{k_p 2}^p \\ ... & ... & ... & ... & ... & ... & ... & ... \\ w_{1m}^1 & ... & w_{k_1 m}^1 & ... & ... & w_{1m}^p & ... & w_{k_p m}^p \end{pmatrix}$$

the associated symbolic data table and $\mathbf{w}_i = (\mathbf{w}_i^{1^\intercal}, \mathbf{w}_i^{2^\intercal}, ..., \mathbf{w}_i^{p^\intercal})$, the i-th row of W, the complete s_i description.

1.2 The φ function

A general φ function that captures the properties given by Tastle et al. (2005) for consensus measures is defined.

Definition 1.1. *The φ function is defined on the set of all l-tuples $\boldsymbol{p}^\intercal = (p_1, p_2, ..., p_l)$, with $p_j \geq 0$ and $\sum_j p_j = 1$, each one representing a probability distribution of a set of ordered categories. This function satisfies the following properties:*

1. The φ function is maximum only for the l-tuples: $(1, 0, ..., 0)$, $(0, 1, ..., 0)$, ..., $(0, 0, ..., 1)$.
2. The φ function is minimum only for the l-tuple $(\frac{1}{2}, 0, ..., 0, \frac{1}{2})$.
3. $\varphi(p_1, p_2, ..., p_l) = \varphi(p_l, p_{l-1}, ..., p_1), \forall \boldsymbol{p}^\intercal = (p_1, p_2, ..., p_l)$.
4. The φ function is convex: $\varphi(\alpha \boldsymbol{p}_1^\intercal + (1-\alpha)\boldsymbol{p}_2^\intercal) \leq \alpha\varphi(\boldsymbol{p}_1^\intercal) + (1-\alpha)\varphi(\boldsymbol{p}_2^\intercal), \forall \alpha \in [0, 1]$.

This φ function should measure the dispersion of categories, being the complementary of a variability measure. For normalization, in this paper it will be considered that the image of φ is the $[0, 1]$ interval. The φ function could be defined for whatever l dimension.

2 Consensus measures for modal ordinal data

In this section we define consensus measures for objects described by modal ordinal data and also for set of them, based on the φ function. We also define the variability for subsets of $\mathcal{S}$ and a dissimilarity measure for elements of $\mathcal{S}$ and for subsets of $\mathcal{S}$, both of them based on the consensus measures and consequently on the φ function.

Given that for a set of individuals described by mono-evaluated data, consensus measures are defined on empirical probability distributions of ordinal categories, they are easily extended to modal ordinal symbolic data.

2.1 Consensus measure for objects described by symbolic data

A consensus measure for an object $s \in \mathcal{S}$ described by $\mathbf{w}^{j\intercal} = (w_1^j, w_2^j, ..., w_{k_j}^j)$ for issue y_j is defined as $\varphi(w_1^j, w_2^j, ..., w_{k_j}^j)$. This value may measure the consensus of a set of individuals described by mono-evaluated data on Y_j with empirical probability distribution given by $\mathbf{w}^{j\intercal}$. This measure is extended to p issues by:

$$c_\varphi(s) = \frac{1}{p}\sum_{j=1}^{p} \varphi(w_1^j, w_2^j, ..., w_{k_j}^j) \tag{1}$$

2.2 Consensus measure for sets of objects described by symbolic data

Given a set $G \subseteq \mathcal{S}$ of modal ordinal data objects, the consensus of G is based on the φ function and it is defined as:

$$C_\varphi(G) = 1 - (\sum_{s_l \in G} \frac{q_{s_l}}{q_G} c_\varphi(s_l) - c_\varphi(g)) \tag{2}$$

where $q_G = \sum\limits_{s_l \in G} q_{s_l}$ and $g = \sum\limits_{s_l \in G} \frac{q_{s_l}}{q_G} s_l$ is the centroid of G. The symbolic description of g is given by $\sum\limits_{s_l \in G} \frac{q_{s_l}}{q_G} \mathbf{w}_l$. This consensus measure defines a variability measure as the complementary of $C_\varphi(G)$, which is given by:

$$Q_\varphi(G) = \sum_{s_l \in G} \frac{q_{s_l}}{q_G} c_\varphi(s_l) - c_\varphi(g) \tag{3}$$

The $Q_\varphi(G)$ value measures the weighted average of G element consensus values with respect to the consensus value of the G centroid. It is verified that $0 \leq C_\varphi(G), Q_\varphi(G) \leq 1, \forall G \subseteq \mathcal{S}$ given that φ is convex and because of φ normalization in the $[0, 1]$ interval. It is also verified that $C_\varphi(\{s_i\}) = 1, Q_\varphi(\{s_i\}) = 0, \forall s_i \in \mathcal{S}$.

2.3 Consensus variability for sets of symbolic data objects

Given that our approach to the clustering problem is based on partitions, the functions defined in this paper for sets of subsets of $\mathcal{S}$ will consider that these subsets are disjoint. These considerations are made for the variability function U_φ and for the distance function D_φ defined below. Even though, extensions for non-disjoint subsets can be easily done.

Let $\mathcal{P} = \{G_1, G_2, ..., G_k\}$ be a set of disjoint subsets $G_k \subseteq \mathcal{S}$ with weights $q_1, q_2, ..., q_k$ ($q_r = \sum\limits_{s_l \in G_r} q_{s_l}$) and centroids $g_1, g_2, ..., g_k$. The consensus variability of $\mathcal{P}$ based on the φ function is defined by:

$$U_\varphi(\mathcal{P}) = Q_\varphi(\bigcup_{r=1}^{k} G_r) - \sum_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} Q_\varphi(G_r) \tag{4}$$

with $q_\mathcal{P} = \sum\limits_{r=1}^{k} q_r$. The consensus variability of $\mathcal{P}$ is the $\bigcup\limits_{r=1}^{k} G_r$ consensus variability minus the weighted average consensus variability of the G_r. It is easily deduced that:

$$U_\varphi(\mathcal{P}) = \sum_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} C_\varphi(G_r) - C_\varphi(\bigcup_{r=1}^{k} G_r) \tag{5}$$

Thus, the $U_\varphi(\mathcal{P})$ value is the weighted average of G_r consensus values minus the consensus value of $\bigcup\limits_{r=1}^{k} G_r$.

Proposition 2.1. *The value $U_\varphi(\mathcal{P})$ can be expressed by:*

$$U_\varphi(\mathcal{P}) = Q_\varphi(\{g_1, g_2, ..., g_k\}) = \sum_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} c_\varphi(g_r) - c_\varphi(g) \tag{6}$$

where g is the centroid of $\bigcup\limits_{r=1}^{k} G_r$ and also the centroid of $\{g_1, g_2, ..., g_k\}$.

Proof. $U_\varphi(\mathcal{P}) = Q_\varphi(\bigcup\limits_{r=1}^{k} G_r) - \sum\limits_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} Q_\varphi(G_r) = \sum\limits_{s_l \in \bigcup\limits_{r=1}^{k} G_r} \frac{q_{s_l}}{q_\mathcal{P}} c_\varphi(s_l) - c_\varphi(g) -$

$\sum\limits_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} \left[\sum\limits_{s_l \in G_r} \frac{q_{s_l}}{q_r} c_\varphi(s_l) - c_\varphi(g_r) \right] = \sum\limits_{r=1}^{k} \frac{q_r}{q_\mathcal{P}} c_\varphi(g_r) - c_\varphi(g) = Q_\varphi(\{g_1, g_2, ..., g_k\})$ □

Thus, the consensus variability of $\mathcal{P}$ is the consensus variability of the G_r centroid set. This suggests that the centroid of a set of symbolic data objects is a suitable representative of the set. The $U_\varphi(\mathcal{P})$ value is in the interval $[0, Q_\varphi(\mathcal{S})]$. This value is minimum for $\mathcal{P} = \mathcal{S}$ ($U_\varphi(\mathcal{P}) = 0$) and maximum for the trivial partition $\mathcal{P} = \{\{s_1\}, \{s_2\}, ..., \{s_m\}\}$ ($U_\varphi(\mathcal{P}) = Q_\varphi(\mathcal{S})$).

In particular, if $\mathcal{P}$ is a partition of $\mathcal{S}$ applying proposition 2.1 to equation (4) the decomposition of $\mathcal{S}$ consensus variability in terms of the $Q_\varphi(.)$ function is:

$$Q_\varphi(\mathcal{S}) = Q_\varphi(\{g_1, g_2, ..., g_k\}) + \sum_{r=1}^{k} q_r Q_\varphi(G_r) \tag{7}$$

showing that consensus variability of $\mathcal{S}$ is the sum of between consensus variability of G_r and the weighted within consensus variabilities of the G_r.

Proposition 2.2. *For any partition $\mathcal{P} = \{G_1, G_2, ..., G_{k-1}, G_k\}$ of $\mathcal{S}$ it is verified that:*

$$U_\varphi(\mathcal{P}) = U_\varphi(\{G_1, G_2, ..., G_{k-1} \cup G_k\}) + q_{k-1,k} U_\varphi(\{G_{k-1}, G_k\}) \tag{8}$$

where $q_{k-1,k} = q_{k-1} + q_k$. This result is valid for whatever pair of sets union.

Proof. For notation, it will be considered that $k_1 = k - 1$ and $q_{k_1 k} = q_{k-1} + q_k$. Let $g_{k_1 k}$, g be the centroids of $\{s_{k-1}, s_k\}$, $\mathcal{S}$.

$U_\varphi(\{G_1, G_2, ..., G_{k-1} \cup G_k\}) = \sum\limits_{r=1}^{k-2} q_r c_\varphi(g_r) + q_{k_1 k} c_\varphi(g_{k_1 k}) - c_\varphi(g) = \sum\limits_{r=1}^{k-2} q_r c_\varphi(g_r)$

$+ q_{k_1 k} \left[c_\varphi(g_{k_1 k}) + \frac{q_{k-1}}{q_{k_1 k}} c_\varphi(g_{k-1}) - \frac{q_{k-1}}{q_{k_1 k}} c_\varphi(g_{k-1}) + \frac{q_k}{q_{k_1 k}} c_\varphi(g_k) - \frac{q_k}{q_{k_1 k}} c_\varphi(g_k) \right] - c_\varphi(g)$

$= \sum\limits_{r=1}^{k} q_r c_\varphi(g_r) - c_\varphi(g) + q_{k_1 k} \left[c_\varphi(g_{k_1 k}) - \frac{q_{k-1}}{q_{k_1 k}} c_\varphi(g_{k-1}) - \frac{q_k}{q_{k_1 k}} c_\varphi(g_k) \right] =$

$U_\varphi(\mathcal{P}) - q_{k_1 k} U_\varphi(\{G_{k-1}, G_k\})$ □

2.4 Dissimilarity measures for symbolic data objects

We introduce here a dissimilarity measure that is based on the φ function. Let $s_l, s_t \in \mathcal{S}$ be two modal ordinal symbolic data objects, the dissimilarity between these two objects is defined by:

$$d_\varphi(s_l, s_t) = Q_\varphi(\{s_l, s_t\}) = \frac{q_{s_l}}{q_{s_l s_t}} c_\varphi(s_l) + \frac{q_{s_t}}{q_{s_l s_t}} c_\varphi(s_t) - c_\varphi(g_{s_l s_t}) \tag{9}$$

with $q_{s_l s_t} = q_{s_l} + q_{s_t}$ and $g_{s_l s_t}$ the centroid of $\{s_l, s_t\}$. This d_φ function is a dissimilarity given that the φ function is convex. The dissimilarity between two elements s_l, s_t of $\mathcal{S}$ is the consensus variability of $\{s_l, s_t\}$ set and it takes values in the $[0, 1]$ interval.

This dissimilarity function can be easily extended to sets of modal ordinal symbolic data objects. Let $A = \{a_1, a_2, ..., a_{k_a}\}$, $B = \{b_1, b_2, ..., b_{k_b}\}$ be two disjoint subsets of $\mathcal{S}$. The dissimilarity between A and B is defined by:

$$D_\varphi(A, B) = U_\varphi(\{A, B\}) \tag{10}$$

The dissimilarity between two subsets A, B of $\mathcal{S}$ is the consensus variability of $\{A, B\}$ set. Let a, b be the centroids of A, B. As it is verified that $U_\varphi(\{A, B\}) = Q_\varphi(\{a, b\}) = d_\varphi(a, b)$, then:

$$D_\varphi(A, B) = d_\varphi(a, b) \tag{11}$$

Thus, the dissimilarity between two subsets of $\mathcal{S}$ is the dissimilarity between their centroids.

For any partition $\mathcal{P} = \{G_1, G_2, ..., G_{k-1}, G_k\}$ of $\mathcal{S}$, it is deduced from proposition 2.2 that:

$$\begin{aligned} & U_\varphi(\{G_1, G_2, ..., G_{k-1}, G_k\}) - U_\varphi(\{G_1, G_2, ..., G_{k-1} \cup G_k\}) \\ = \quad & q_{k-1,k} U_\varphi(\{G_{k-1}, G_k\}) = q_{k-1,k} D_\varphi(G_{k-1}, G_k) \end{aligned} \tag{12}$$

The decreasing consensus variability of a partition $\mathcal{P}$ of $\mathcal{S}$, when two of its members are joined to form a new subset, is proportional to the dissimilarity between the two sets joined.

The c_φ and d_φ functions, defined above on $\mathcal{S}$, are actually defined on $\mathfrak{S}$. The Q_φ function defined below for subsets of $\mathcal{S}$, is actually defined for countable subsets of $\mathfrak{S}$.

3 The Leik measure

In this section we select a specific φ function to be used in the clustering method that we propose in section 4. Among the different consensus measures in the literature (see García-Santesmases et al., 2010) we focus our attention into the Leik measure because it does not make any assumptions on ordinal categories and can be applied to any ordinal scale.

Let $\boldsymbol{p}^\intercal = (p_1, p_2, ..., p_l)$ be a probability distribution associated to a set of increasing ordered categories, let $\boldsymbol{F}^\intercal = (F_1, F_2, ..., F_l)$ be the cumulative distribution associated to $\boldsymbol{p}^\intercal$ distribution and

$$\begin{aligned} d_j &= F_j \text{ if } F_j \leq 0.5 \\ d_j &= 1 - F_j \text{ otherwise} \end{aligned}$$

The sum $\sum_j d_j$ (see Leik, 1966) is a dispersion index. Standardizing this sum by the value $max\{\sum_j d_j | p_j \geq 0, \sum_j p_j = 1\} = \frac{l-1}{2}$, and taking its complementary to one, then the Leik measure of $\boldsymbol{p}^\intercal$ is:

$$lk(\boldsymbol{p}^\intercal) := 1 - \frac{2\sum_j d_j}{l-1} \tag{13}$$

Proposition 3.1. *The lk function is a φ function.*

Proof. The four properties of definition 1.1 are proven:

1. max $lk(\boldsymbol{p}^\intercal) = 1 \Leftrightarrow \sum_j d_j = 0 \Leftrightarrow \forall j,\ F_j = 0$ or $F_j = 1 \Leftrightarrow \exists\ j\prime | p_{j'} = 1$ and $\forall j \neq j', p_j = 0$.

2. $\min lk(\boldsymbol{p}^{\intercal}) = 0 \Leftrightarrow \sum_j d_j = \frac{l-1}{2} \Leftrightarrow d_j = \frac{1}{2}$, for $j = 1, ..., l-1$ and $d_l = 0 \Leftrightarrow \boldsymbol{p}^{\intercal} = (\frac{1}{2}, 0, ..., 0, \frac{1}{2})$.
3. $lk(p_1, p_2, ..., p_l) = lk(p_l, p_{l-1}, ..., p_1)$, $\forall \boldsymbol{p}^{\intercal} = (p_1, p_2, ..., p_l)$. Let $p_j^2 = p_{l-j+1}$ be the $j-th$ position of $\boldsymbol{p^{2}}^{\intercal} = (p_l, p_{l-1}, ..., p_1)$ and $\boldsymbol{F}^{\intercal} = (F_1, ..., F_l)$, $\boldsymbol{F}^{2^{\intercal}} = (F_1^2, ..., F_l^2)$ be the cumulative distributions associated to $\boldsymbol{p}^{\intercal}$, $\boldsymbol{p^{2}}^{\intercal}$. For notation $F_0 = F_0^2 = 0$. We are going to demonstrate that $F_j^2 = 1 - F_{l-j}, \forall j$. The equality $p_j^2 = p_{l-j+1} = F_{l-j+1} - F_{l-j}$ is true $\forall j$, in particular $F_1^2 = p_1^2 = p_l = F_l - F_{l-1} = 1 -F_{l-1}$, thus $F_1^2 = 1 - F_{l-1}$. If for $j-1$ is verified that $F_{j-1}^2 = 1 - F_{l-(j-1)}$ then we demonstrate that this is true for j: $F_j^2 = F_{j-1}^2 + p_j^2 = 1 - F_{l-(j-1)} + F_{l-j+1} - F_{l-j} = 1 - F_{l-j}$. We demonstrate that $d_j = d_{l-j}^2$ for $j = 1, ..., l-1$ and trivially $d_l = d_l^2 = 0$. For $j = 1, ..., l-1$, if $d_j = F_j \leq 0.5$, then $F_{l-j}^2 = 1 - F_j \geq 0.5$, and $d_{l-j}^2 = 1 - F_{l-j}^2 = F_j = d_j$; in a similar way it is demonstrated that if $d_j = 1 - F_j$, then $d_{l-j}^2 = d_j$. Thus, $\sum\limits_{j=1}^{l} d_j = \sum\limits_{j=1}^{l} d_j^2$ and consequently, $lk(\boldsymbol{p}^{\intercal}) = lk(\boldsymbol{p}^{2^{\intercal}})$.
4. lk is convex. Let $\boldsymbol{p}^{1^{\intercal}} = (p_1^1, p_2^1, ..., p_l^1)$, $\boldsymbol{p}^{2^{\intercal}} = (p_1^2, p_2^2, ..., p_l^2)$ and d_j^1, F_j^1, d_j^2, F_j^2 defined as above. Let $\alpha \in [0, 1]$ and $\boldsymbol{p}^{\intercal} = \alpha \boldsymbol{p}^{1^{\intercal}} + (1-\alpha)\boldsymbol{p}^{2^{\intercal}}$. The $\boldsymbol{F}^{\intercal}$ associated to $\boldsymbol{p}^{\intercal}$ is $\boldsymbol{F}^{\intercal} = \alpha \boldsymbol{F}^{1^{\intercal}} + (1-\alpha)\boldsymbol{F}^{2^{\intercal}}$. To evaluate d_j $(j = 1, ..., l)$ associated to $\boldsymbol{F}^{\intercal}$, four cases are distinguished:
   (a) If $F_j^1 \leq 0.5$ and $F_j^2 \leq 0.5$ then $d_j = F_j = \alpha F_j^1 + (1-\alpha) F_j^2 = \alpha d_j^1 + (1-\alpha) d_j^2 (\leq 0, 5)$.
   (b) If $F_j^1 \geq 0.5$ and $F_j^2 \geq 0.5$ then $F_j \geq 0.5$ and $d_j = 1 - F_j = 1 - (\alpha F_j^1 + (1-\alpha) F_j^2) = \alpha(1 - F_j^1) + (1-\alpha)(1 - F_j^2) = \alpha d_j^1 + (1-\alpha) d_j^2 (\geq 0.5)$.
   (c) When $F_j^1 \geq 0.5$ and $F_j^2 \leq 0.5$ we again distinguish two cases:
      i. If $F_j \leq 0.5$ then $d_j = F_j = \alpha F_j^1 + (1-\alpha) F_j^2 = \alpha(1 - d_j^1) + (1-\alpha) d_j^2 \geq \alpha d_j^1 + (1-\alpha) d_j^2$.
      ii. If $F_j \geq 0.5$ then $d_j = 1 - F_j = 1 - (\alpha F_j^1 + (1-\alpha) F_j^2) = \alpha(1 - F_j^1) + (1-\alpha)(1 - F_j^2) = \alpha d_j^1 + (1-\alpha)(1 - d_j^2) \geq \alpha d_j^1 + (1-\alpha) d_j^2$.
   (d) For $F_j^1 \leq 0.5$ and $F_j^2 \geq 0.5$ it is proven that $d_j \geq \alpha d_j^1 + (1-\alpha) d_j^2$ in a similar way as in case 4c.

Thus, $\forall j \in \{1, ..., l\}$, $d_j \geq \alpha d_j^1 + (1-\alpha) d_j^2$ and then $1 - \frac{2\sum_j d_j}{l-1} \leq \alpha(1 - \frac{2\sum_j d_j^1}{l-1}) + (1-\alpha)(1 - \frac{2\sum_j d_j^2}{l-1})$. Therefore, $lk(\boldsymbol{p}^{\intercal}) \leq \alpha lk(\boldsymbol{p}^{1^{\intercal}}) + (1-\alpha) lk(\boldsymbol{p}^{2^{\intercal}})$. □

Corollary 3.2. $lk(\boldsymbol{p}^{\intercal}) = \alpha\, lk(\boldsymbol{p}^{1^{\intercal}}) + (1-\alpha) lk(\boldsymbol{p}^{2^{\intercal}}) \Leftrightarrow \boldsymbol{p}^{\intercal}, \boldsymbol{p}^{1^{\intercal}}, \boldsymbol{p}^{2^{\intercal}}$ *have the same median.*

Proof. The sufficient condition is almost trivial. With respect to the necessary condition, let suppose that the medians are not equal. This is only possible if the median associated to $\boldsymbol{p}^{1^{\intercal}}$ is different from the median associated to $\boldsymbol{p}^{2^{\intercal}}$. This is only possible in the following case: $\exists j \in \{1, ..., l\}$ for which $F_j^1 \geq 0.5$, $F_j^2 < 0.5$ or $F_j^1 > 0.5$, $F_j^2 \leq 0.5$ or $F_j^1 \leq 0.5$, $F_j^2 > 0.5$ or $F_j^1 < 0.5$, $F_j^2 \geq 0.5$. In those cases then it is verified that $d_j > \alpha d_j^1 + (1-\alpha) d_j^2$ and consequently $lk(\boldsymbol{p}^{\intercal}) > \alpha\, lk(\boldsymbol{p}^{1^{\intercal}}) + (1-\alpha) lk(\boldsymbol{p}^{2^{\intercal}})$. □

Corollary 3.3. *When the φ function is the lk function then the derived d_φ dissimilarity function satisfies the following: If $\mathrm{d}_\varphi(s_l, s_t) = 0$ for $s_l, s_t \in \mathcal{S}$, then s_l, s_t have the same median for each y_j issue, $j = 1, ..., p$.*

Proof. This result is a consequence of the corollary 3.2. □

4 Hierarchical clustering algorithm

An ascending hierarchical clustering procedure (see Ward, 1963) is proposed to build a sequence of partitions $\mathcal{P}_m, \mathcal{P}_{m-1}, ..., \mathcal{P}_1$ of $\mathcal{S}$ in the following way: $\mathcal{P}_m = \{\{s_1\}, ..., \{s_m\}\}$ is the trivial partition, $\mathcal{P}_{k-1}$ is derived from the $\mathcal{P}_k$ partition by joining the pair of its members that minimizes the decrease of consensus variability when going from the $\mathcal{P}_k$ partition to the $\mathcal{P}_{k-1}$ partition.

In each step, when $\mathcal{P}_k = \{G_1, G_2, ..., G_k\}$ the union of every possible pair of sets is considered. The pair of sets $G_{j_1}$and G_{j_2}, with j_1, $j_2 \in \{1, ..., k\}$ that are finally merged are those whose union results in the minimum decrease of consensus variability. This decrease of consensus variability (proposition 2.2) is:

$$U_\varphi(\mathcal{P}_k) - U_\varphi(\mathcal{P}_{k-1}) = q_{j_1 j_2} U_\varphi(\{G_{j_1}, G_{j_2}\}) = q_{j_1 j_2} D_\varphi(G_{j_1}, G_{j_2})$$

with $q_{j_1 j_2}$ the sum of G_{j_1}, G_{j_2} weights. Thus, in each step the most similar pair of sets are merged to form the new partition. Given that $D_\varphi(G_{j_1}, G_{j_2})$ is non-negative the non-increasing monotonicity of between cluster consensus variabilities is assured.

4.1 Interpretation of clusters

Let $\mathcal{P}_k = \{G_1, G_2, ..., G_k\}$ be a partition of $\mathcal{S}$ in k classes, the quality of a cluster G_r is given by:

$$Q_\varphi(G_r) = \sum_{s_l \in G_r} \frac{q_{s_l}}{q_r} c_\varphi(s_l) - c_\varphi(g_r)$$

where $q_r = \sum_{s_l \in G_r} q_{s_l}$ and g_r is the centroid of G_r cluster. As told before, $Q_\varphi(G_r)$ is a measure of consensus variability that measures the within-cluster consensus variability. The lower this measure is, the higher the quality of G_r cluster is. The quality of G_r cluster is minimum when $Q_\varphi(G_r)$ takes its maximum value. The G_r cluster quality is maximum when $Q_\varphi(G_r) = 0$. When the φ function is the lk function, the value $Q_\varphi(G_r)$ is null when all the elements of G_r have the same median for each issue y_j.

The quality of a cluster regarding issue y_j is:

$$Q^j_\varphi(G_r) = \sum_{s_l \in G_r} \frac{q_{s_l}}{q_r} \varphi(\mathbf{w}_l^{j\intercal}) - \varphi(\sum_{s_l \in G_r} \frac{q_{s_l}}{q_r} \mathbf{w}_l^{j\intercal})$$

where $\mathbf{w}_l^j$ is the modal ordinal data description of $s_l \in G_r$ regarding issue y_j.

An useful criterion to interpret a cluster regarding a variable y_j is given by:

$$\frac{Q^j_\varphi(G_r)}{Q^j_\varphi(\mathcal{S})} \leq \frac{Q_\varphi(G_r)}{Q_\varphi(\mathcal{S})}$$

in the sense that those issues y_j of G_r cluster that verify this inequality are the issues that characterize the cluster. The lower the $\frac{Q^j_\varphi(G_r)}{Q^j_\varphi(\mathcal{S})}$ ratio value is for a variable y_j, the more homogenous in consensus variability this variable is in G_r cluster.

4.2 Interpretation of the partition

Let $\mathcal{P} = \{G_1, G_2, ..., G_k\}$ be a partition of $\mathcal{S}$ in k clusters with weights $q_1, q_2, ..., q_k$ and centroids $g_1, g_2, ..., g_k$. The quality of a partition is given by:

$$U_\varphi(\mathcal{P}) = \sum_{r=1}^{k} q_r c_\varphi(g_r) - c_\varphi(g) = Q_\varphi(\{g_1, g_2, ..., g_k\})$$

with g the centroid of $\mathcal{S}$. The value of $U_\varphi(\mathcal{P})$ measures the consensus variability of the G_r centroid set and it is a measure of the between-cluster consensus variability. The higher this measure is, the higher the quality of $\mathcal{P}$ is.

The quality of a partition regarding issue y_j is:

$$U^j_\varphi(\mathcal{P}) = \sum_{r=1}^{k} q_r \varphi(\mathbf{w}^{j^\intercal}_{g_r}) - \varphi(\sum_{r=1}^{k} q_r \mathbf{w}^{j^\intercal}_{g_r})$$

where $\mathbf{w}^j_{g_r} = \sum_{s_l \in G_r} \frac{q_{s_l}}{q_r} \mathbf{w}^j_l$ is the modal ordinal data description of g_r regarding issue y_j.

These measures can be normalized into the $[0, 1]$ interval by using the $\frac{U_\varphi(\mathcal{P})}{Q_\varphi(\mathcal{S})}$ and $\frac{U^j_\varphi(\mathcal{P})}{Q^j_\varphi(\mathcal{S})}$ ratios. These ratios measure the proportions of consensus variabilities of $\mathcal{S}$ that are explained by the $\mathcal{P}$ partition globally and for a y_j issue, respectively.

5 Example

To illustrate the proposed method we apply it to a data set $\mathcal{S}$ composed of 34 teachers described by modal ordinal symbolic data. They were rated by their students (1350) on 12 items on the ordinal scale: poor, average, good, excellent. The items were: y_1, initial subject presentation; y_2, teacher setting to course syllabus; y_3, well time management; y_4, evoking interest in the students about the subject; y_5, use of practical examples; y_6, stimulating students to be active in class and readiness to clear their doubts; y_7, readiness to give advice in academic development; y_8, degree of respect between students and teacher; y_9, subject knowledge; y_{10}, stimulating students to read books, journals and magazines; y_{11}, communications skills; and, y_{12}, ability to clear students' doubts. Figure 1 represents the teachers' modal ordinal symbolic data: for each teacher, empirical probability distributions of y_j issues are represented by vertical bar charts.

For the first step (the trivial partition with $k = 34$) and for the last four steps of the ascending hierarchical clustering algorithm, the values of the proportion of consensus variability of $\mathcal{S}$ explained by each partition are shown in table 1 as well as the between-cluster consensus variabilities of partitions. The solution chosen is $\mathcal{P}_3 = \{G_1, G_2, G_3\}$ for $k = 3$, that explains

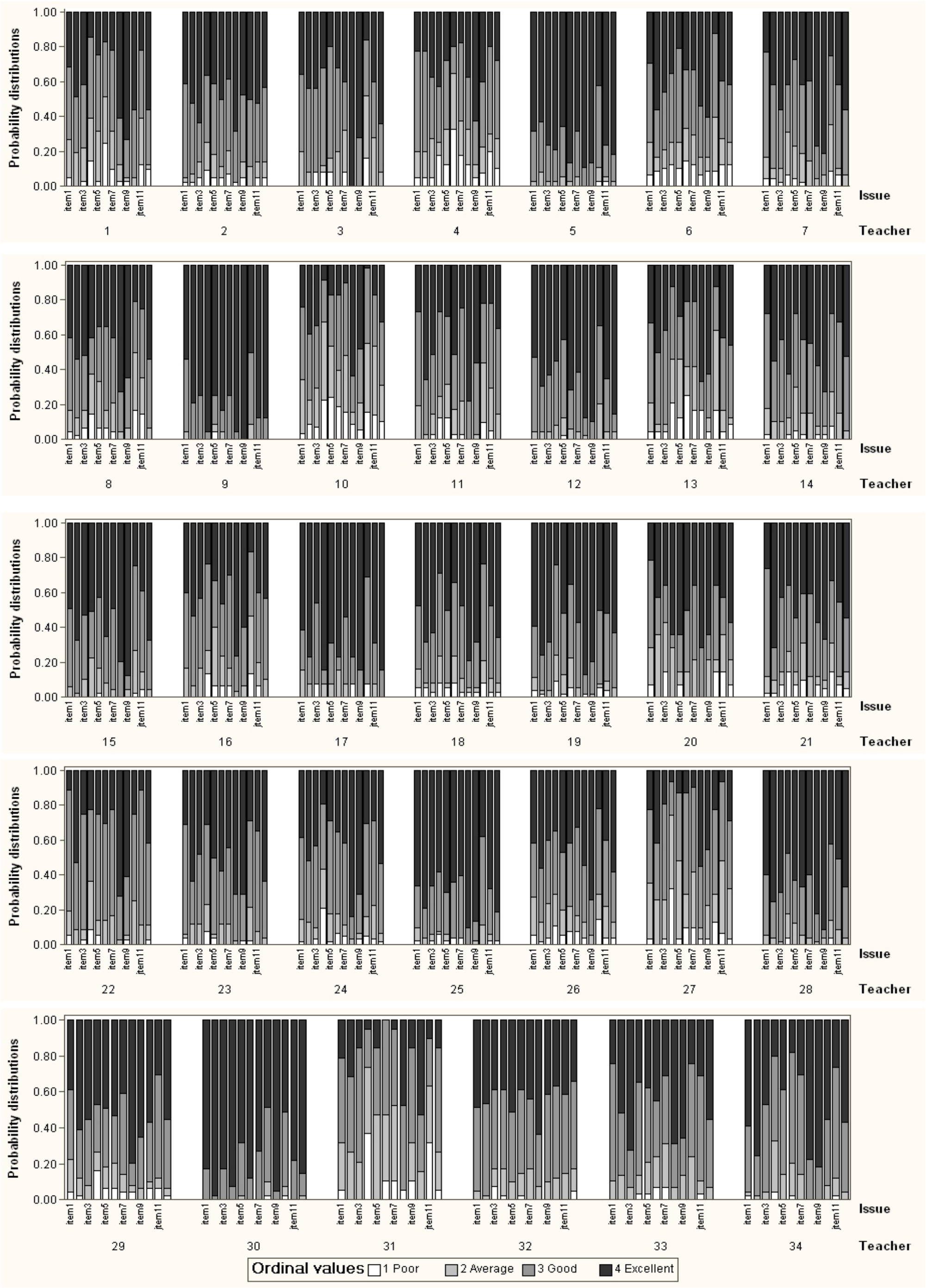

FIG. 1 – *Representation of teachers' modal ordinal symbolic data*

$71,2\%$ of $\mathcal{S}$ consensus variability. This step is also chosen because there is not a big difference when taking $k = 4$ clusters (75%).

To evaluate the quality of the partitions obtained, we have randomly generated 1000 partitions of size 2, 3, 4, respectively and measured their consensus variabilities. The generated partitions have the same cluster sizes that those obtained by our method. Table 1 shows mean and standard deviations of these consensus variabilities as well as the p-values of the permutation tests that evaluates the significance of the partitions obtained by our method. For size 3 we have also generated 1000 partitions with no restrictions on cluster sizes. Their consensus variability mean value was 0.00283 and their standard deviation 0.0022. All partition consensus variabilities were lower than 0.037.

k	$\frac{U_\varphi(\mathcal{P}_k)}{Q_\varphi(\mathcal{S})}$	$U_\varphi(\mathcal{P}_k)$	$Mean\,U_\varphi(.)$	$Std\,U_\varphi(.)$	$p - value$
34	1	0.052			
...	...	...	...	...	...
...	...	...	...	...	...
4	0.75	0.039	0.0045	0.0033	< 0.001
3	0.712	0.037	0.00288	0.00243	< 0.001
2	0.596	0.031	0.00156	0.00756	< 0.001
1	0	0			

TAB. 1 – *Quality of partitions in algorithm steps. Consensus variability mean and standard deviation values of the simulated partitions. P-values of permutation tests.*

The G_1, G_2, G_3 clusters are composed of 23, 4 and 7 teachers, respectively. The cluster members are $G_1 = \{s_{26}, s_7, s_{33}, s_{14}, s_8, s_{21}, s_{16}, s_2, s_{23}, s_3, s_6, s_{24}, s_{29}, s_{32}, s_{11}, s_{34}, s_{18}, s_{20}, s_{15}, s_{19}, s_1, s_{13}, s_{22}\}$, $G_2 = \{s_4, s_{31}, s_{10}, s_{27}\}$ and $G_3 = \{s_5, s_9, s_{12}, s_{17}, s_{25}, s_{28}, s_{30}\}$. Figure 2 represents the modal ordinal symbolic data of the G_r cluster centroids of the $\mathcal{P}_3$ partition. Generally speaking, G_3 is composed of teachers with the highest evaluations and G_2 with the lowest evaluations whilst cluster G_1 is composed of the teachers with intermediate evaluations.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	global
$Q^j_\varphi(\mathcal{S})$	0.044	0.035	0.085	0.1	0.053	0.09	0.043	0.005	0.02	0.016	0.064	0.062	0.052
$Q^j_\varphi(G_1)$	0.01	0.015	0.049	0.015	0.013	0.015	0.004	0	0.007	0.019	0.001	0.037	0.015
$Q^j_\varphi(G_2)$	0	0	0	0.08	0.01	0.053	0.01	0.05	0.04	0.026	0.06	0	0.027
$Q^j_\varphi(G_3)$	0	0	0.007	0.005	0.01	0	0	0	0	0.003	0	0	0.002
$U^j_\varphi(\mathcal{P}_3)$	0.037	0.025	0.05	0.08	0.04	0.07	0.04	0	0.01	0	0.05	0.03	0.037
$\frac{U^j_\varphi(\mathcal{P}_3)}{Q^j_\varphi(\mathcal{S})}$	0.841	0.714	0.588	0.8	0.755	0.778	0.93	0	0.5	0	0.781	0.484	0.712

TAB. 2 – *Consensus variability values of $\mathcal{S}$, G_r and $\mathcal{P}_3$. Proportion of $\mathcal{S}$ consensus variability values explained by the $\mathcal{P}_3$ partition*

In table 2 the consensus variability values of $\mathcal{S}$, G_r clusters and $\mathcal{P}_3$ partition are shown for y_j issues and for all issues globally. There are also included the values of the proportion of $\mathcal{S}$ consensus variability explained by the partition for y_j and for all issues globally. The issues

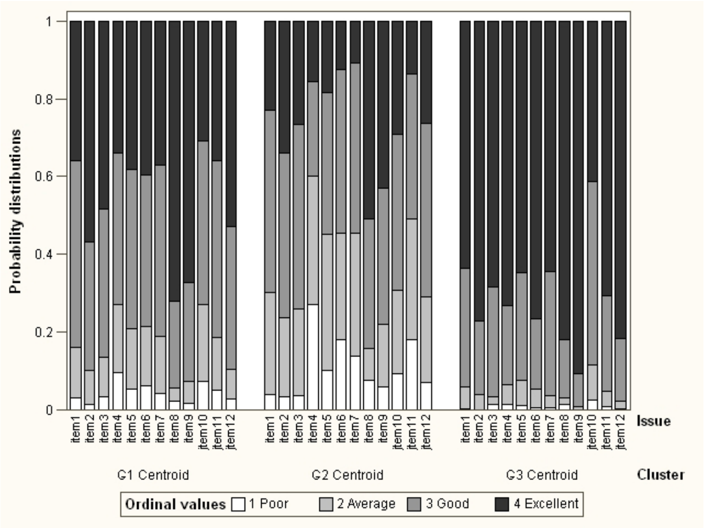

FIG. 2 – *Representation of the $\mathcal{P}_3$ partition cluster centroids*

that better explain the partition are those whose $\frac{U_\varphi^j(\mathcal{P}_3)}{Q_\varphi^j(\mathcal{S})}$ values are bigger than 0.712. These issues are y_1, y_4 to y_7 and y_{11}.

In table 3, the $\frac{Q_\varphi^j(G_r)}{Q_\varphi^j(\mathcal{S})}$ and $\frac{Q_\varphi(G_r)}{Q_\varphi(\mathcal{S})}$ ratios are shown. Those values for which $\frac{Q_\varphi^j(G_r)}{Q_\varphi^j(\mathcal{S})} \leq \frac{Q_\varphi(G_r)}{Q_\varphi(\mathcal{S})}$ are in bold type. These values correspond to the issues of each cluster that characterize the cluster in the sense that for these issues within consensus variabilities are lower. Looking at this table to the ratios relating to the issues that best explain the partition (y_1, y_4 to y_7, y_{11}) we can summarize that y_1 and y_7 are the variables that better explain the partition, that is, the *initial subject presentation* and the *readiness to give advice in academic development* are the most discriminant properties among the groups. *Stimulating students to be active in class and readiness to clear their doubts* and the *communication skills* (y_6, y_{11}) discriminate well the best teachers from the intermediate teachers (G_1 and G_3 clusters) and the *use of practical examples* (y_5) discriminates well the worst teachers from the intermediate teachers (G_1 and G_2 clusters).

We can also observe in table 3 an outlier in ratios that corresponds to cluster G_2 and issue y_8. This issue has very little consensus variability values in the original set ($Q_\varphi^8(\mathcal{S}) = 0.005$, table 2), null within-$G_1$ and G_3 consensus variabilities ($Q_\varphi^8(G_k) = 0$, $k = 1, 3$, table 3 and all consensus variability is in cluster G_2 ($Q_\varphi^8(G_2) = 0.05$, table 3). This is due to teacher 4 who receives the lowest rates for *degree of respect between students and teacher*.

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}	global
$\frac{Q^j_\varphi(G_1)}{Q^j_\varphi(\mathcal{S})}$	**0.23**	0.42	0.58	**0.15**	**0.24**	**0.17**	**0.09**	**0**	0.33	1.13	**0.02**	0.6	0.30
$\frac{Q^j_\varphi(G_2)}{Q^j_\varphi(\mathcal{S})}$	**0**	**0**	**0**	0.8	**0.18**	0.59	**0.23**	8.5	1.92	1.58	0.92	**0**	0.53
$\frac{Q^j_\varphi(G_3)}{Q^j_\varphi(\mathcal{S})}$	**0**	**0**	0.093	0.059	0.25	**0**	**0**	**0**	**0**	0.23	**0**	**0**	0.049

TAB. 3 – *Quality of clusters for the $\mathcal{P}_3$ partition*

Conclusion

In this paper we have introduced a general φ function to characterize a consensus measure defined for probability distributions for a set of ordinal categories. We extend this measure to sets of modal ordinal symbolic data objects and define a dissimilarity measure for these sets based in the consensus variability of their centroids.

We have presented an ascending hierarchical clustering algorithm for modal ordinal data. In each step, the two clusters joined are those with the minimum distance between their centroids, the same criterion applied in the Ward algorithm who used the Euclidean distance for continuous mono-evaluated data (see Ward, 1963). As an example of a φ function we have chosen the Leik measure that is suitable for any ordinal scale and we have applied the proposed method to analyze data coming from the evaluation of a set of teachers by their students.

References

Bock, H.H. and E. Diday (Eds.) (2000). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Heidelberg: Springer Verlag.

Chavent, M. (2000). Criterion-based divisive clustering for symbolic data. In: Bock, H.H. and E. Diday (Eds.). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data*. Heidelberg: Springer Verlag, 299-311.

Garcia-Santesmases, J.M. and M.C. Bravo (2010). Consensus analysis through modal symbolic objects. *Compstat 2010 Proceedings*. Springer, ISBN 978-3-7908-2603-6, 1055–1062.

García-Santesmases, J.M., C. Franco and J. Montero (2010). Consensus measures for symbolic data. *Computer Engineering and Information Science 4*, 651–658.

Kim, J and L. Billard (2012). Dissimilarity measures and divisive clustering for symbolic multimodal-valued data. *Computational Statistics and Data Analysis 56*, 2795-2808.

Leik, K.R. (1966). A measure of ordinal consensus. *The Pacific Sociological Review 9*, 85–90.

Tastle, W.J., M.J. Wierman and U.R. Dumdum (2005). Ranking ordinal scales using the consensus measure. *Issues in Information Systems 6 (2)*, 96–102.

Ward, J. (1963). Hierarchical grouping to optimize an objective. *Journal of the American Statistical Association 58 (301)*, 236–244.