Comparing Partitions by Means of Fuzzy Data Mining Tools

Carlos Molina[1], Belén Prados[2], María-Dolores Ruiz[3], Daniel Sánchez[4], and José-Mańia Serrano[1]

[1] Dept of Computer Science, University of Jaén, Spain
{carlosmo,jschica}@ujaen.es
[2] Dept. Software Engineering, University of Granada, Spain
belenps@ugr.es
[3] Dept. Computer Science and A.I., University of Granada, Spain
mdruiz@decsai.ugr.es
[4] European Centre for Soft Computing, Mieres, Spain
daniel.sanchezf@softcomputing.es

Abstract. Rand index is one of the most popular measures for comparing two partitions over a set of objects. Several approaches have extended this measure for those cases involving fuzzy partitions. In previous works, we developed a methodology for correspondence analysis between partitions in terms of data mining tools. In this paper we discuss how, without any additional cost, it can be applied as an alternate computation of Rand index, allowing us not only to compare both crisp and fuzzy partitions, but also classes inside these partitions.

Keywords: fuzzy partitions, fuzzy data mining tools, Rand index.

1 Introduction

Fuzzy models have been extensively used in pattern recognition. In particular, clustering techniques have been extended to determine a finite set of groups or categories, that can be fuzzy (elements being associated to each cluster with a degree of membership), to describe a set of objects with similar features. Developed algorithms have been successfully applied in a wide range of areas including image recognition, signal processing, market segmentation, document categorization and bioinformatics.

The main problems arising the comparison of two fuzzy partitions of a given set are the following: (1) the number of clusters in both partitions are not necessarily the same, (2) the measures for comparing two equivalent partitions, that can be represented by matrices A and B, must be invariant under row permutations.

As far as we know, most of current approaches are only suitable for comparing a fuzzy partition with a crisp one, where the latter represents the "true" partition of data. But in nearly all real cases, there is no such crisp partition giving a perfect matching.

There are three kinds of approaches for evaluating the partitions quality: internal, external and relative criteria [27]. *Internal* criterion is used for evaluating a partition separately, usually for measuring the grade of fit between the partition and the input data. *External* measures compare the obtained partition with a reference partition that

E. Hüllermeier et al. (Eds.): SUM 2012, LNAI 7520, pp. 337–350, 2012.

pertains to the data but which is independent of it. *Relative* measures, also known as relative indices, assess the similarity between two partitions computed by different methods. Our approach belongs to this last group. Our goal in this paper is to define an alternative to the popular Rand index [29], by means of a family of data mining tools, applied to both crisp and fuzzy correspondence analysis between partitions.

The paper is organized as follows. In the following section, we mention some comparison methods between partitions, specially those related to fuzzy cases. Then, we summarize the models for data mining employed as tools for analyzing some types of correspondences, described in the next section. After this, there is our problem approach in terms of the data mining measures applied to correspondence analysis. Finally, some future trends in this work to come are defined as well as we present our conclusions.

2 Rand Index and Other Comparison Measures

Comparison methods include those measures that compare two partitions. When comparing a resulting partition by a clustering process with a referential one, which is considered to be the "true" partition, we will call it an external method. External indices [27] give the expert an indication of the quality of the resulting partition, while when comparing two different partitions we obtain a grade of how similar they are. If both partitions come from different clustering processes the method is considered as relative.

There are many indices to be reviewed for crisp partitions [22] (see also [2]). For fuzzy partitions we will refer to the most important approaches developed until now. Many of them are generalizations of crisp measures.

The Rand index [29] proposed by Rand in 1971 is given in terms of the number of pairwise comparisons of data objects. It is one of the most popular indices. Given A and B two crisp clusters we set:

- a, pairs belonging to the same cluster in A and to the same cluster in B.
- b, pairs belonging to the same cluster in A but to a different cluster in B.
- c, pairs belonging to a different cluster in A but to the same cluster in B.
- d, pairs belonging to different clusters in both A and B.

Then, the Rand index is given by the proportion between the number of agreements and the total number of pairs:

$$I_R(A,B) = \frac{a+d}{a+b+c+d} \tag{1}$$

Campello [13] extends the Rand index for comparing fuzzy partitions. For that purpose, he rewrites the original formulation in terms of the fuzzy partitions. Let X and Y be two fuzzy partitions defined over the set of objects O, we consider:

- $X_1 = \{(o,o') \in O \times O$ that belong to the same cluster in $X\}$.
- $X_0 = \{(o,o') \in O \times O$ that belong to different clusters in $X\}$.
- $Y_1 = \{(o,o') \in O \times O$ that belong to the same cluster in $Y\}$.
- $Y_0 = \{(o,o') \in O \times O$ that belong to the different clusters in $Y\}$.

The Rand index is rewritten in terms of the previous four quantities: $a = |X_1 \cap Y_1|, b = |X_1 \cap Y_0|$, $c = |X_0 \cap Y_1|, d = |X_0 \cap Y_0|$. In the fuzzy case the sets X_i, Y_i are defined by means of a t-norm $\otimes$ and a t-conorm $\oplus$. Let $X_i(o) \in [0,1]$ the degree of membership of element $o \in O$ in the i-th cluster of X. Analogously $Y_i(o) \in [0,1]$ is the degree of membership of element $o \in O$ in the i-th cluster of Y

- $X_1(o,o') = \bigoplus_{i=1}^{k} X_i(o) \otimes X_i(o')$
- $X_0(o,o') = \bigoplus_{1 \le i \ne j \le k} X_i(o) \otimes X_j(o')$
- $Y_1(o,o') = \bigoplus_{i=1}^{l} Y_i(o) \otimes Y_i(o')$
- $Y_0(o,o') = \bigoplus_{1 \le i \ne j \le l} Y_i(o) \otimes Y_j(o')$

The four frequencies taking part in equation (1) are then formulated in terms of the intersection of these sets using the sigma-count principle:

$$\begin{aligned} a &= \sum_{(o,o') \in O \times O} X_1(o,o') \otimes Y_1(o,o') \qquad b = \sum_{(o,o') \in O \times O} X_1(o,o') \otimes Y_0(o,o') \\ c &= \sum_{(o,o') \in O \times O} X_0(o,o') \otimes Y_1(o,o') \qquad d = \sum_{(o,o') \in O \times O} X_0(o,o') \otimes Y_0(o,o') \end{aligned} \tag{2}$$

This is not the only generalization of the Rand index. We can find in the literature the approaches of Frigui et al. [21], Brouwer [11], Hüllermeier and Rifqi [23] and Anderson et al. [2]. In [3,2] the reader may find a more extensive comparison of the cited indices. We will resume the main differences between them:

- Campello was interested in comparing a fuzzy partition with a non-fuzzy one, but its proposal is formulated for comparing two fuzzy partitions.
- Frigui et al. present generalizations for several indices including the Rand index. They also restrict the approach when one of the partitions is a crisp one. When using product for the t-norm and sum for the t-conorm for the Campello's approach we obtain this particular case [21].
- Brouwer presents another generalization by defining a relationship called bonding that describes the degree to which two objects are in the same cluster. Then, bonding matrices are built using previous relation and the cosine distance [11].
- Hüllermeier and Rifqi's approach is defined for every two fuzzy partitions by defining a fuzzy equivalence relation on the set of objects O. This fuzzy relation is then used for defining the degree of concordance or discordance between two objects $o, o' \in O$. The distance obtained using the resulting index satisfies the desirable properties for a pseudo-metric and in some special cases it is a metric [23].

A very similar index was proposed by several authors: the so-called Jaccard coefficient [24] where the participation of the quantity d is suppressed in Campello's index.

The Fowlkes-Mallows index proposed in [20] can be defined as in equation (3) obtaining a value of 1 when clusters are good estimates of the groups.

$$I_F(A,B) = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{3}$$

Previous indices, as well as the Adjusted Rand index of Hubert and Arabie [22], the C statistics [25] and the Minkowski measure [26] can be defined in terms of the four frequencies a, b, c, d and they are related to Rand index. All these indices allow uniquely the evaluation of hard (crisp) clustering partitions, but some authors [13,14,2] have extended all of them in a unified formulation. The first attempt of Campello [13] relies solely on the redefinition of the four frequencies using basic fuzzy set concepts, but it has the shortcoming that one of the partitions must to be hard for keeping the important property of reaching their maximum (unit value) when comparing equivalent partitions. A more recent approach [14] settles this shortcoming by defining a *fuzzy transfer distance* between two fuzzy partitions. In addition, Campello also addresses the problem of how to compare two partitions from different subsamples of data.

Anderson et al. [2] developed a method to generalize comparison indices to all possible cases concerning two different partitions: crisp, fuzzy, probabilistic and possibilistic and for every index that can be expressed in terms of the four mentioned frequencies.

A different proposal by Di Nuovo and Catania [27], called DNC index, is based on a defined measure called *degree of accuracy* which is intended to measure the degree of association of a partition with its reference partition representing the real group.

A quite different approach is that developed by Runkler [31] which is based on the similarities between the resultant subsets by the partitions. The *subset similarity index* is computed in terms of the similarities between all the partitions subsets. This index is reflexive and invariant under row permutations which are desirable properties.

3 Crisp and Fuzzy Data Mining Tools

3.1 Association Rules

Given a set I ("set of items") and a database D constituted by set of transactions ("T-set"), each one being a subset of I, association rules [1] are "implications" of the form $A \Rightarrow B$ that relate the presence of itemsets A and B in transactions of D, assuming $A, B \subseteq I, A \cap B = \emptyset$ and $A, B \neq \emptyset$.

The ordinary measures proposed in [1] to assess association rules are *confidence* (the conditional probability $P(B|A)$) and *support* (the joint probability $P(A \cup B)$). An alternative framework [8,16] measures accuracy by means of Shortliffe and Buchanan's certainty factors [33], showing better properties than confidence, and helping to solve some of its drawbacks. Let $\text{supp}(B)$ be the support of the itemset B, and let $\text{Conf}(A \Rightarrow B)$ be the confidence of the rule. The *certainty factor* of the rule is defined as

$$CF(A \Rrightarrow B) = \begin{cases} \frac{\text{Conf}(A \Rightarrow B) - \text{supp}(B)}{1 - \text{supp}(B)} & \text{if } \text{Conf}(A \Rightarrow B) > \text{supp}(B \\ \frac{\text{Conf}(A \Rightarrow B) - \text{supp}(B)}{\text{supp}(B)} & \text{if } \text{Conf}(A \Rightarrow B) < \text{supp}(B) \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

The certainty factor yields a value in the interval [-1, 1] and measures how our belief that B is in a transaction changes when we are told that A is in that transaction.

3.2 Formal Model for Mining Fuzzy Rules

Many definitions for fuzzy rule can be found in the literature, but in this work, we will apply the formal model developed in [17,19], which allows us to mine fuzzy rules in a straightforward way extending the accuracy measures from the crisp case. Its formalization basically underlies in two concepts: the representation by levels associated to a fuzzy property (RL for short) and the four fold table associated to the itemsets A and B in database D, noted by $\mathscr{M} = 4ft(A,B,D)$.

A RL associated to a fuzzy property P in a universe X is defined as a pair (Λ_P, ρ_P) where $\Lambda_P = \{\alpha_1, \ldots, \alpha_m\}$ is a finite set of levels verifying that $1 = \alpha_1 > \cdots > \alpha_m > \alpha_{m+1} = 0$ and $\rho_P : \Lambda_P \to \mathscr{P}(X)$ is a function which applies each level into the crisp realization of P in that level [32]. The set of crisp representatives of P is the set $\Omega_P = \{\rho_P(\alpha) \,|\, \alpha \in \Lambda_P\}$. The values of Λ_P can be interpreted as values of possibility for a possibility measure defined for all $\rho_P(\alpha_i) \in \Omega_P$ as $Pos(\rho_P(\alpha_i)) = \alpha_i$. Following this interpretation we define the associated probability distribution $m : \Omega_P \to [0,1]$ as in equation (5) which give us information about how representative is each crisp set of the property P in Ω_P.

$$m_P(Y) = \sum_{\alpha_i \,|\, Y = \rho(\alpha_i)} \alpha_i - \alpha_{i+1} \tag{5}$$

For each $Y \in \Omega_P$, the value $m_P(Y)$ represents the proportion to which the available evidence supports claim that the property P is represented by Y. From this point of view, a RL can be seen as a basic probability assignment in the sense of the theory of evidence, *plus a structure indicating dependencies between the possible representations of different properties*.

The four fold table associated to the itemsets involved in a rule $A \Rightarrow B$ detaches the number of transactions in D satisfying the four possible combinations between A and B using the logic connectors $\wedge$ (conjunction) and $\neg$ (negation). So, $\mathscr{M} = 4ft(A,B,D) = \{a,b,c,d\}$ where a is the number of rows of D satisfying $A \wedge B$, b the number of rows satisfying $A \wedge \neg B$, c represents those satisfying $\neg A \wedge B$ and d those satisfying the last possibility $\neg A \wedge \neg B$ [30,18]. Note that $|D| = a+b+c+d = n$. The validity of an association rule is assessed by using $\mathscr{M}$ by means an operator $\approx$ (interestingness measure) called 4ft-quantifier. In particular, known measures of support and confidence are 4ft-quantifiers defined as follows:

$$\begin{aligned} \text{Supp}(A \Rightarrow B) &= \approx_S (a,b,c,d) = \frac{a}{a+b+c+d} \\ \text{Conf}(A \Rightarrow B) &= \approx_C (a,b,c,d) = \frac{a}{a+b} \end{aligned} \tag{6}$$

and we can use them to define the certainty factor $\approx_{CF} (a,b,c,d)$ in terms of the four frequencies of $\mathscr{M}$ (see [18] for its shorter form).

Using these two models we have proposed [19] a framework for fuzzy rules that ables us to extend the interestingness measures for their validation from the crisp to the fuzzy case. Summarizing the model, we can represent the fuzzy sets appearing in the fuzzy rule by the associated RLs $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$, $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$ and for every level in $\Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ we define the associated four fold table as $\mathscr{M}_{\alpha_i} = (a_i, b_i, c_i, d_i)$ whose values are computed using the previous RLs (see [19] for more details).

Using $\mathscr{M}_{\alpha_i}$ and the probability distribution of equation (5) we extend the accuracy measures for fuzzy rules from the crisp case [19]:

$$\sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) (\approx (a_i, b_i, c_i, d_i)) \tag{7}$$

The model is a good generalization of the crisp case, allowing the use of equation (7) in the fuzzy definition of measures in equation (6) as, respectively, FSupp($A \Rightarrow B$), FConf($A \Rightarrow B$), and FCF($A \Rightarrow B$) (see [19] for a complete discussion).

3.3 Approximate Dependencies

Let $RE = \{At_1, \ldots, At_m\}$ be a relational scheme and let r be an instance of RE such that $|r| = n$. Also, let $V, W \subset RE$ with $V \cap W = \emptyset$. A functional dependency $V \rightarrow W$ holds in RE if and only if

$$\forall t, s \in r \;\; if \; t[V] = s[V] \; then \; t[W] = s[W] \tag{8}$$

Approximate dependencies can be roughly defined as functional dependencies with exceptions. The definition of approximate dependence is then a matter of how to define exceptions, and how to measure the accuracy of the dependence [10]. We shall follow the approach introduced in [15,9], where the same methodology employed in mining for association rules is applied to the discovery of approximate dependencies.

Since a functional dependency '$V \rightarrow W$' can be seen as a rule that relates the equality of attribute values in pairs of tuples (see equation (8)), and association rules relate the presence of items in transactions, we can represent approximate dependencies as association rules by using the following interpretations of the concepts of item and transaction:

- An item is an object associated to an attribute of RE. For every attribute $At_k \in RE$ we note it_{At_k} the associated item.
- We introduce the itemset I_V to be $I_V = \{it_{At_k} \,|\, At_k \in V\}$
- T_r is a T-set that, for each pair of tuples $< t, s > \in r \times r$ contains a transaction $ts \in T_r$ verifying $it_{At_k} \in ts \;\Leftrightarrow\; t[At_k] = s[At_k]$ It is obvious that $|T_r| = |r \times r| = n^2$.

Then, an approximate dependence $V \rightarrow W$ in the relation r is an association rule $I_V \Rightarrow I_W$ in T_r [15,9]. The support and certainty factor of $I_V \Rightarrow I_W$ measure the interest and accuracy of the dependence $V \rightarrow W$.

3.4 Fuzzy Approximate Dependencies

In [7] a definition integrating both approximate and fuzzy dependencies features is introduced. In addition to allowing exceptions, the relaxation of several elements of equation (8) is considered. In particular, we associate membership degrees to pairs $< attribute, value >$ as in the case of fuzzy association rules, as well as the equality of the rule is smoothed as a fuzzy similarity relation.

Extending the crisp case above, fuzzy approximate dependencies in a relation are defined as fuzzy association rules on a special fuzzy T-set obtained from that relation.

Let $I_{RE} = \{it_{At_k} | At_k \in RE\}$ be the set of items associated to the relational schema RE. We define a fuzzy T-set $\widetilde{T}_r$ as follows: for each pair of rows $< t,s >$ in $r \times r$ we have a fuzzy transaction ts in $\widetilde{T}_r$ defined as

$$\forall it_{At_k} \in \widetilde{T}_r, ts(it_{At_k}) = min(At_k(t), At_k(s), S_{At_k}(t(At_k), s(At_k))) \quad (9)$$

This way, the membership degree of a certain item it_{At_k} in the transaction associated to tuples t and s takes into account the membership degree of the value of At_k in each tuple ($At_k(t)$) and the similarity between them (S_{At_k}). The latter represents the degree to which tuples t and s agree in At_k. According to this, let $X, Y \subseteq RE$ with $X \cap Y = \emptyset$ and $X, Y \neq \emptyset$. The fuzzy approximate dependence[7] $X \rightarrow Y$ in r is defined as the fuzzy association rule $I_X \Rightarrow I_Y$ in $\widetilde{T}_r$.

Analogously to the crisp case, we measure the importance and accuracy of the fuzzy approximate dependence $X \rightarrow Y$ as the support and certainty factor of the fuzzy association rule $I_X \Rightarrow I_Y$ (see section 3.2).

4 Correspondence Analysis in Terms of Data Mining Tools

Correspondence analysis [6] describes existing relations between two nominal variables, by means of a contingency table, obtained as the cross-tabulation of both variables. It can be applied to reduce data dimension, prior to a subsequent statistic processing (classification, regression, discriminant analysis, ...). In particular, it can be helpful in the integration or matching of different partitions over a set of objects.

4.1 Crisp Correspondences

In [5], we introduced an alternate methodology to classic correspondence analysis, centered in the interpretation of a set of rules and/or dependencies. For that, we represent the possible correspondences between objects as a relational table, where the value of a cell for a given object (row) and partition (column) means the class in the partition where the object is.

Let O be a finite set of objects, and $\mathscr{A} = \{A_1, A_2, \ldots, A_p\}$, $\mathscr{B} = \{B_1, B_2, \ldots, B_q\}$ two partitions of O, i.e., $A_i, B_j \subseteq O$ and $A_i, B_j \neq \emptyset$, $A_{i_1} \cap A_{i_2} = \emptyset \; \forall i_1, i_2 \in \{1, \ldots, p\}$ and $B_{j_1} \cap B_{j_2} = \emptyset \; \forall j_1, j_2 \in \{1, \ldots, q\}$. Also, $\bigcup_{A_i \in \mathscr{A}} A_i = \bigcup_{B_j \in \mathscr{B}} B_j = O$.

We represent partitions $\mathscr{A}$ and $\mathscr{B}$ by means of a table, $r_{\mathscr{A}\mathscr{B}}$ (see table 1), and we shall use the notation for relational databases. Each row (tuple) and column (attribute) of $r_{\mathscr{A}\mathscr{B}}$ will be associated to an object and a partition, respectively. This way, we assume $|r_{\mathscr{A}\mathscr{B}}| = |O|$.

We shall note t_o the tuple associated to object o, and $X_{\mathscr{P}}$ the attribute associated to partition $\mathscr{P}$. The value for tuple t_o and attribute $X_{\mathscr{P}}$, $t_o[X_{\mathscr{P}}]$, will be the class for o following $\mathscr{P}$, i.e., $t_o[X_{\mathscr{P}}] \in \mathscr{P}$.

Let us remark that we are interested not only in perfect correspondences, but also in those with possible exceptions. Hence, we are concerned with measuring the accuracy of correspondences between partitions.

Definition 1 ([5]). Local correspondence. *Let $A_i \in \mathscr{A}$ and $B_j \in \mathscr{B}$. There exists a local correspondence from A_i to B_j when $A_i \subseteq B_j$.*

Table 1. Table $r_{\mathscr{A}\mathscr{B}}$

Object	tuple	$X_{\mathscr{A}}$	$X_{\mathscr{B}}$
o_1	t_{o_1}	A_1	B_2
o_2	t_{o_2}	A_2	B_2
o_3	t_{o_3}	A_1	B_1
...	...	...	...

The analysis of local correspondences can be performed by looking for association rules in the table $r_{\mathscr{A}\mathscr{B}}$. Rules $[X_{\mathscr{A}} = A_i] \Rightarrow [X_{\mathscr{B}} = B_j]$ and $[X_{\mathscr{B}} = B_j] \Rightarrow [X_{\mathscr{A}} = A_i]$ tell us about possible local correspondences between classes A_i and B_j.

Definition 2 ([5]). Partial correspondence. *There exists a partial correspondence from $\mathscr{A}$ to $\mathscr{B}$, noted $\mathscr{A} \Rrightarrow \mathscr{B}$, when $\forall A_i \in \mathscr{A}\ \exists B_j \in \mathscr{B}$ such that $A_i \subseteq B_j$.*

Definition 3 ([5]). Global correspondence. *There exists a global correspondence between $\mathscr{A}$ and $\mathscr{B}$, noted $\mathscr{A} \equiv \mathscr{B}$, when $\mathscr{A} \Rrightarrow \mathscr{B}$ and $\mathscr{B} \Rrightarrow \mathscr{A}$.*

The analysis of partial correspondences can be performed by looking for approximate dependencies in $r_{\mathscr{A}\mathscr{B}}$ [5]. If the dependence $X_{\mathscr{A}} \rightarrow X_{\mathscr{B}}$ holds, there is a partial correspondence from $\mathscr{A}$ to $\mathscr{B}$. The certainty factor of the dependence measures the accuracy of the correspondence. As we are interested in using the same measure to assess global correspondences, this leads to define the certainty factor of $\mathscr{A} \equiv \mathscr{B}$ as the minimum between $CF(\mathscr{A} \Rrightarrow \mathscr{B})$ and $CF(\mathscr{B} \Rrightarrow \mathscr{A})$, since it is usual to obtain the certainty factor of a conjunction of facts as the minimum of the certainty factors of the facts.

4.2 Fuzzy Correspondences

Consider the case of establishing correspondences between diseases and symptoms. A certain disease can be described by several symptoms, at a given degree, and also a symptom can be related to different diseases. Since the original correspondence analysis [6] is not able to manage such cases in which partitions boundaries are not so clear, we extended the alternate methodology discussed in section 4.1 in order to manage correspondences between fuzzy partitions [12].

Let $O = \{o_1, \ldots, o_n\}$ be again a finite set of objects. Let $\widetilde{\mathscr{A}} = \{\widetilde{A}_1, \ldots, \widetilde{A}_p\}$ and $\widetilde{\mathscr{B}} = \{\widetilde{B}_1, \ldots, \widetilde{B}_q\}$ be two fuzzy partitions over O. Let $\widetilde{T}_{\widetilde{\mathscr{A}}\widetilde{\mathscr{B}}}$(Table 2) be the fuzzy transactional table associated to O, each transaction representing an object, that is, $|\widetilde{T}_{\widetilde{\mathscr{A}}\widetilde{\mathscr{B}}}| = |O|$. Given $o \in O, \widetilde{A}_i \in \widetilde{\mathscr{A}}$ and $\widetilde{B}_j \in \widetilde{\mathscr{B}}$, we noted for $\widetilde{A}_i(o)$ (respectively, $\widetilde{B}_j(o)$) the membership degree of o in $\widetilde{A}_i$ (respectively, $\widetilde{B}_j$). Each object must belong to at least one class of each partition, that is, $\forall o \in O, \exists \widetilde{P}_i \in \widetilde{\mathscr{P}} / \widetilde{P}_i(o) > 0$, and each class must contain at least one object, that is, $\widetilde{A}_i, \widetilde{B}_j \neq \emptyset$.

As we manage fuzzy partitions, we can relax the condition of disjoint classes within a partition. Also, we do not consider the case of partitions being necessarily normalized.

Definition 4 ([12]). ***Fuzzy local correspondence.*** *Let $\widetilde{A}_i \in \widetilde{\mathscr{A}}$ and $\widetilde{B}_j \in \widetilde{\mathscr{B}}$. There exists a fuzzy local correspondence from $\widetilde{A}_i$ to $\widetilde{B}_j$, noted $\widetilde{A}_i \Rightarrow \widetilde{B}_j$, if $\widetilde{A}_i \subseteq \widetilde{B}_j$, that is, $\forall o \in O$,*

Table 2. Fuzzy transactional table $\widetilde{T}_{\widetilde{\mathscr{A}}\widetilde{\mathscr{B}}}$

Object	$\widetilde{A}_1$	...	$\widetilde{A}_p$	$\widetilde{B}_1$	...	$\widetilde{B}_q$
o_1	$\widetilde{A}_1(o_1)$	...	$\widetilde{A}_p(o_1)$	$\widetilde{B}_1(o_1)$	...	$\widetilde{B}_q(o_1)$
o_2	$\widetilde{A}_1(o_2)$	...	$\widetilde{A}_p(o_2)$	$\widetilde{B}_1(o_2)$	...	$\widetilde{B}_q(o_2)$
o_3	$\widetilde{A}_1(o_3)$	...	$\widetilde{A}_p(o_3)$	$\widetilde{B}_1(o_3)$	...	$\widetilde{B}_q(o_3)$
...	...	...	...	...	...	...

$\widetilde{A}_i(o) \leq \widetilde{B}_j(o)$. *This time, we can obtain fuzzy local correspondences in terms of fuzzy association rules.*

When analyzing fuzzy partial and global correspondences, we must manage not classes, but partitions. It would be necessary to define a membership degree of an object in a partition, that is, $\widetilde{\mathscr{A}}(o)$. This defines a multidimensionality problem, already addressed in [12], and it is a pending task currently under researching. For sake of simplicity, we will reduce to the case in which an object is associated to only one class in every partition, for example, that with the highest membership degree.

We shall represent partitions $\widetilde{\mathscr{A}}$ and $\widetilde{\mathscr{B}}$ by means of a fuzzy relational table, $\widetilde{r}_{\widetilde{\mathscr{A}}\widetilde{\mathscr{B}}}$ (Table 3). Each row (object) is related to a column (partition) with a certain membership degree. The value corresponding to tuple t_o and attribute $X_{\widetilde{\mathscr{A}}}$, $t_o[X_{\widetilde{\mathscr{A}}}]$, will be the class for o according to partition $\widetilde{\mathscr{A}}$, that is, $t_o[X_{\widetilde{\mathscr{A}}}] \in \widetilde{\mathscr{A}}$. We shall note as $X_{\widetilde{\mathscr{A}}}(o)$ the membership degree of o in $t_o[X_{\widetilde{\mathscr{A}}}]$. As discussed before, we shall note this as $\widetilde{\mathscr{A}}(o)$.

Table 3. Fuzzy relational table, $\widetilde{r}_{\widetilde{\mathscr{A}}\widetilde{\mathscr{B}}}$

Object	$X_{\widetilde{\mathscr{A}}}$	$X_{\widetilde{\mathscr{B}}}$
t_{o_1}	$\widetilde{A}_{i1}, \widetilde{\mathscr{A}}(o_1)$	$\widetilde{B}_{j1}, \widetilde{\mathscr{B}}(o_1)$
t_{o_2}	$\widetilde{A}_{i2}, \widetilde{\mathscr{A}}(o_2)$	$\widetilde{B}_{j2}, \widetilde{\mathscr{B}}(o_2)$
t_{o_3}	$\widetilde{A}_{i3}, \widetilde{\mathscr{A}}(o_3)$	$\widetilde{B}_{j3}, \widetilde{\mathscr{B}}(o_3)$
...	...	...

Definition 5 ([12]). ***Fuzzy partial correspondence.*** *There exists a fuzzy partial correspondence from $\widetilde{\mathscr{A}}$ to $\widetilde{\mathscr{B}}$, noted $\widetilde{\mathscr{A}} \Rrightarrow \widetilde{\mathscr{B}}$, when $\forall \widetilde{A}_i \in \widetilde{\mathscr{A}}\ \exists \widetilde{B}_j \in \widetilde{\mathscr{B}}$ such that $\widetilde{A}_i \subseteq \widetilde{B}_j$, that is, $\forall o \in O / t_o[\widetilde{\mathscr{A}}] = \widetilde{A}_i$ implies $t_o[\widetilde{\mathscr{B}}] = \widetilde{B}_j$ and $\widetilde{\mathscr{A}}(o) \dot{\leq} \widetilde{\mathscr{B}}(o)$.*

$\dot{\leq}$ defines a vectorial order relation that, for this particular case, corresponds to a classic order relation.

Definition 6 ([12]). ***Fuzzy global correspondence.*** *There exists a fuzzy global correspondence between $\widetilde{\mathscr{A}}$ and $\widetilde{\mathscr{B}}$, noted $\widetilde{\mathscr{A}} \equiv \widetilde{\mathscr{B}}$, when $\widetilde{\mathscr{A}} \Rrightarrow \widetilde{\mathscr{B}}$ and $\widetilde{\mathscr{B}} \Rrightarrow \widetilde{\mathscr{A}}$.*

Fuzzy partial and global correspondences relate fuzzy partitions, and both can be obtained by means of fuzzy approximate dependencies.

5 Our Proposal. Discussion

In this section, we give an alternate approach to Rand index which is valid for both crisp and fuzzy partitions, in terms of the measures employed in the tools described in section 3, and using the tabular representation described in section 4. Hence, let O be again a finite set of objects, $|O| = n$, with $\mathscr{A} = \{A_1, A_2, \ldots, A_p\}$ and $\mathscr{B} = \{B_1, B_2, \ldots, B_q\}$, two different partitions over O.

Analogously to the approach described in [15,9] for approximate dependencies, let T be a transactional table where each row represents an ordered pair of objects $(o, o') \in O \times O$, $|T| = \frac{n(n-1)}{2}$. Let $I_{\mathscr{A}}$ (resp., $I_{\mathscr{B}}$) be an item that indicates that both objects o, o' belong to the same class in partition $\mathscr{A}$ (resp., $\mathscr{B}$). According to this, we can redefine the Rand index parameters, a, b, c, d in terms of the support measure as:

- $a = |T| \cdot supp(I_{\mathscr{A}} \cap I_{\mathscr{B}})$,
- $b = |T| \cdot (supp(I_{\mathscr{A}}) - supp(I_{\mathscr{A}} \cap I_{\mathscr{B}}))$,
- $c = |T| \cdot (supp(I_{\mathscr{B}}) - supp(I_{\mathscr{A}} \cap I_{\mathscr{B}}))$, and
- $d = |T| \cdot (1 - supp(I_{\mathscr{A}} \cup I_{\mathscr{B}})) = |T| \cdot (1 - supp(I_{\mathscr{A}}) - supp(I_{\mathscr{B}}) + supp(I_{\mathscr{A}} \cap I_{\mathscr{B}}))$.

Thus, we can rewrite the Rand index as,

$$I_R(\mathscr{A}, \mathscr{B}) = \frac{a+d}{a+b+c+d} = \frac{|T| \cdot (supp(I_{\mathscr{A}} \cap I_{\mathscr{B}}) - (1 - supp(I_{\mathscr{A}} \cup I_{\mathscr{B}})))}{|T|} = \\ = 1 - (supp(I_{\mathscr{A}}) + supp(I_{\mathscr{B}}) - 2supp(I_{\mathscr{A}} \cap I_{\mathscr{B}})) \tag{10}$$

Let us notice in first place, how, from this expression, it is trivial that $I_R(\mathscr{A}, \mathscr{B}) = I_R(\mathscr{B}, \mathscr{A})$, since only support is involved in equation 10. Moreover, it is easy to see that these parameters are proportionally equivalent to those of the four fold table used in the model described in section 3.2, allowing us to relate $I_R(\mathscr{A}, \mathscr{B})$ in some way with the measures of support, confidence, and certainty factor (see definitions for 4ft-quantifiers $\approx_S (a,b,c,d)$, $\approx_C (a,b,c,d)$, and $\approx_{CF} (a,b,c,d)$, respectively).

Following this, and taking into account our approach for correspondence analysis (section 4.1), we can establish a direct relation between Rand index and the measurement of partial and global correspondences between two different partitions (by means of approximate dependencies). Even more, we can define a similar measure to analyze not only these types of correspondences, but also local correspondences (by means of association rules).

As for the case of fuzzy partitions, our model for fuzzy correspondences (section 4.2) also allows to obtain measures as informative as the Rand index, again considering both partitions (in terms of fuzzy approximate dependencies) as well as classes (as relations expressed as fuzzy association rules).

Then, according to equation (10), we can redefine the Rand index in terms of the support measure, allowing us to distinguish several interpretations of this measure, based on the data mining tool used as source. Let us consider the following family of indices:

- $I_{AR}(A_i, B_j)$, for comparing classes $A_i \in \mathscr{A}$ and $B_j \in \mathscr{B}$, from association rules,
- $I_{AD}(\mathscr{A}, \mathscr{B})$, for comparing partitions $\mathscr{A}$ and $\mathscr{B}$, from approximate dependencies,

- $I_{FAR}(\widetilde{A}_i, \widetilde{B}_j)$, for comparing fuzzy classes $\widetilde{A}_i \in \widetilde{\mathscr{A}}$ and $\widetilde{B}_j \in \widetilde{\mathscr{B}}$, from fuzzy association rules, and
- $I_{FAD}(\widetilde{\mathscr{A}}, \widetilde{\mathscr{B}})$, for comparing fuzzy partitions $\widetilde{\mathscr{A}}$ and $\widetilde{\mathscr{B}}$, defined in terms of fuzzy approximate dependencies.

Let us remark that our approach allows not only to take into account these proposed measures, but also the already well-defined and popular measures of support, confidence and certainty factor. A very interesting issue could be a deeper study of the combined information obtained by all these values.

From our point of view, these definitions open a new framework in the problem of partition comparison, specially in those cases where the boundaries between classes are unclear, able to be managed by means of fuzzy partitions. In [5], we briefly discussed some interesting properties as, for example, that of how our approach can be applied to the study of relations between more than two partitions. The analysis of the relevance of this and some other properties is a pending task, and future works will be devoted to their study and development.

5.1 A Brief Example

In order to illustrate our proposal, but due to lack of space, we are showing a little example of our methodology, extending the results over the same dataset used in [12]. Here, fuzzy correspondence analysis between different partitions over a set of 211 agricultural zones is addressed and discussed. The first fuzzy partition was obtained as widely discussed in [4] from users (farmers) knowledge, and classified the examples into 19 classes. Let $userclass = \{\widetilde{A}_1, \ldots, \widetilde{A}_{19}\}$ be this classification. A scientific classification was previously presented in [28]. Here, a total of 21 land types, called soil maps units, are found, only 19 being suitable for olive trees cultivation. Let $sciclass = \{\widetilde{B}_1, \ldots, \widetilde{B}_{21}\}$ be this other classification.

Fuzzy local correspondences between classes were computed between *userclass* and *sciclass*, and those more interesting ($CF > 0.65$) are shown in table 4. Each cell in the table shows the CF for the fuzzy local correspondence (fuzzy association rule) of the type $\widetilde{B}_j \Rightarrow \widetilde{A}_i$ (as discussed in [12], the inverse fuzzy local correspondences were found to be not interesting regarding CF). It must be remarked that these results were validated and properly interpreted by soil experts.

Table 5 shows the I_{FAR} value for the same correspondences in table 4. Let us recall that $I(\mathscr{A}, \mathscr{B}) = I(\mathscr{B}, \mathscr{A})$ for any two partitions (or classes, as it is the case). That is, this index tells us about the relation between $\mathscr{A}$ and $\mathscr{B}$, but gives no information about the direction of this relation. From our point of view, this relation is not necessarily symmetric, since one partition class can be partially included in other partition class, but the opposite might not hold. In this sense, these first results suggest that CF measure seems to be more valuable than I_{FAR}. Hence, a more exhaustive and complete analysis of the relation between these measures, considering additional sets of examples, appears to be necessary, and will be properly addressed in a future extension of this work.

Table 4. Fuzzy local correspondences between *sciclass* (rows) and *userclass* (columns) classes, $\widetilde{B}_j \Rightarrow \widetilde{A}_i$ ($CF > 0.65$)

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_{10}	A_{11}	A_{12}	A_{14}	A_{15}	A_{16}	A_{17}
B_1			0.722	0.783				0.687	0.672						0.690
B_3	0.870	0.859	0.793	0.660		0.689	0.655	0.771		0.783			0.759		
B_5	0.684	0.733	0.795	0.744				0.686		0.736					
B_6	0.895	0.913	0.792	0.675		0.661		0.784		0.803			0.813		
B_8	0.886	0.932													
B_9		0.650	0.719	0.687				0.712		0.675					
B_{10}	0.653		0.718	0.728		0.715		0.705	0.711						
B_{11}	0.691	0.764	0.743	0.721	0.653		0.676	0.670		0.813			0.666		
B_{13}				0.700	0.662				0.687						
B_{15}	0.761	0.760	0.842	0.737				0.687	0.681	0.683			0.674		0.664
B_{16}	0.744	0.853	0.812	0.756		0.729	0.734	0.769	0.688	0.696			0.808		
B_{20}			0.803	0.868	0.871			0.667	0.802	0.742	0.700	0.706		0.811	0.770

Table 5. Rand index $I_{FAR}(\widetilde{A}_i, \widetilde{B}_j)$ for fuzzy local correspondences between *sciclass* (rows) and *userclass* (columns) classes

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_{10}	A_{11}	A_{12}	A_{14}	A_{15}	A_{16}	A_{17}
B_1			0.241	0.292				0.292	0.341						0.416
B_3	0.354	0.331	0.283	0.299		0.402	0.438	0.333		0.336			0.398		
B_5	0.346	0.341	0.326	0.354				0.346		0.363					
B_6	0.297	0.277	0.234	0.271		0.366		0.289		0.291			0.358		
B_8	0.288	0.270													
B_9		0.317	0.305	0.339				0.357		0.346					
B_{10}	0.289		0.241	0.286		0.381		0.294	0.346						
B_{11}	0.293	0.281	0.244	0.285	0.422		0.420	0.290		0.307			0.357		
B_{13}				0.287	0.427				0.347						
B_{15}	0.332	0.311	0.293	0.315				0.316	0.365	0.316			0.380		0.434
B_{16}	0.292	0.280	0.242	0.281		0.376	0.419	0.293	0.337	0.288			0.364		
B_{20}			0.235	0.284	0.428			0.281	0.341	0.287	0.626	0.568		0.560	0.411

6 Further Works and Concluding Remarks

Many measures based on the Rand index have been proposed and developed for the study of partitions comparison. A subset of them can be applied also in those cases involving fuzzy partitions. In this work, we have applied a previously developed methodology for correspondence analysis, in terms of fuzzy data mining tools, to the problem of partition comparison, expressed in the form of a measure such as the Rand index. Our approach offers the advantage of being capable of managing both crisp and fuzzy partitions, and, in addition, it allows to compare not only different partitions, but also classes inside these partitions. We have shown an example combining an accuracy measure as CF with the Rand index. Moreover, we have seen how CF, in comparison to Rand index, allows to determine the direction in which the relation between partitions (or classes) is stronger.

Finally, some interesting properties arise from the proposed measures, and a deeper and more complete study and development will be the main topic in future extensions of this work. Practical works covering the discussion of our methodology applied to real world problems, as fuzzy image segmentation (for the comparison of different methods), and classification in medical cases, are also in progress..

Acknowledgements. The research reported in this paper was partially supported by the Andalusian Government (Junta de Andalucía) under project P07-TIC03175 and from the former Spanish Ministry for Science and Innovation by the project grants TIN2006-15041-C04-01 and TIN2009-08296.

References

1. Agrawal, R., Imielinski, T., Swami, A.: Mining Association Rules between Sets of Items in Large Databases. In: Procs. of ACM SIGMOD Conf., Washington DC, USA, pp. 207–216 (1993)
2. Anderson, D.T., Bezdek, J.C., Popescu, M., Keller, J.M.: Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Transactions on Fuzzy Systems 18(5), 906–918 (2010)
3. Anderson, D.T., Bezdek, J.C., Keller, J.M., Popescu, M.: A Comparison of Five Fuzzy Rand Indices. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. CCIS, vol. 80, pp. 446–454. Springer, Heidelberg (2010)
4. Aranda, V., Calero, J., Delgado, G., Sánchez, D., Serrano, J., Vila, M.A.: Flexible land classification for olive cultivation using user knowledge. In: Proceedings of 1st. Int. ICSC Conf. On Neuro-Fuzzy Technologies (NF 2002), La HaBana, Cuba, Enero 16–19 (2002)
5. Aranda, V., Calero, J., Delgado, G., Sánchez, D., Serrano, J.M., Vila, M.A.: Using Data Mining Techniques to Analyze Correspondences Between User and Scientific Knowledge in an Agricultural Environment. In: Enterprise Information Systems IV, pp. 75–89. Kluwer Academic Publishers (2003)
6. Benzécri, J.P.: Cours de Linguistique Mathématique. Université de Rennes, Rennes (1963)
7. Berzal, F., Blanco, I., Sánchez, D., Serrano, J.M., Vila, M.A.: A definition for fuzzy approximate dependencies. Fuzzy Sets and Systems 149(1), 105–129 (2005)
8. Berzal, F., Delgado, M., Sánchez, D., Vila, M.A.: Measuring accuracy and interest of association rules: A new framework. Intelligent Data Analysis 6(3), 221–235 (2002)
9. Blanco, I., Martín-Bautista, M.J., Sánchez, D., Serrano, J.M., Vila, M.A.: Using association rules to mine for strong approximate dependencies. Data Mining and Knowledge Discovery 16(3), 313–348 (2008)
10. Bosc, P., Lietard, L., Pivert, O.: Functional Dependencies Revisited Under Graduality and Imprecision. In: Annual Meeting of NAFIPS, pp. 57–62 (1997)
11. Brouwer, R.K.: Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. Journal of Intelligent Information Systems 32, 213–235 (2009)
12. Calero, J., Delgado, G., Sánchez, D., Serrano, J.M., Vila, M.A.: A Proposal of Fuzzy Correspondence Analysis based on Flexible Data Mining Techniques. In: Soft Methodology and Random Information Systems, pp. 447–454. Springer (2004)
13. Campello, R.J.G.B.: A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment. Pattern Recognition Letters 28, 833–841 (2007)
14. Campello, R.J.G.B.: Generalized external indexes for comparing data partitions with overlapping categories. Pattern Recognition Letters 31, 966–975 (2010)
15. Delgado, M., Martín-Bautista, M.J., Sánchez, D., Vila, M.A.: Mining strong approximate dependencies from relational databases. In: Procs. of IPMU 2000 (2000)
16. Delgado, M., Marín, N., Sánchez, D., Vila, M.A.: Fuzzy Association Rules: General Model and Applications. IEEE Transactions on Fuzzy Systems 11(2), 214–225 (2003)
17. Delgado, M., Ruiz, M.D., Sánchez, D.: A restriction level approach for the representation and evaluation of fuzzy association rules. In: Procs. of the IFSA-EUSFLAT, pp. 1583–1588 (2009)

18. Delgado, M., Ruiz, M.D., Sánchez, D.: Studying Interest Measures for Association Rules through a Logical Model. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems 18(1), 87–106 (2010)
19. Delgado, M., Ruiz, M.D., Sánchez, D., Serrano, J.M.: A Formal Model for Mining Fuzzy Rules Using the RL Representation Theory. Information Sciences 181, 5194–5213 (2011)
20. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. J. of American Statistical Society 78, 553–569 (1983)
21. Frigui, H., Hwang, C., Rhee, F.C.H.: Clustering and aggregation of relational data with applications to image database categorization. Pattern Recognition 40, 3053–3068 (2007)
22. Hubert, L.J., Arabie, P.: Comparing partition. J. Classification 2, 193–218 (1985)
23. Hüllermeier, E., Rifqi, M., Henzgen, S., Senge, R.: Comparing fuzzy partitions: A generalization of the Rand index and related measures. IEEE Transactions of Fuzzy Systems 20(3), 546–556 (2012)
24. Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901)
25. Jain, A., Dubes, R.: Algorithms for Clustering Data. Prentice Hall (1988)
26. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene-expression data: A survey. IEEE Trans. Knowledge Data Engineering 16, 1370–1386 (2004)
27. Di Nuovo, A.G., Catania, V.: On External Measures for Validation of Fuzzy Partitions. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.) IFSA 2007. LNCS (LNAI), vol. 4529, pp. 491–501. Springer, Heidelberg (2007)
28. Pérez-Pujalte, A., Prieto, P.: Mapa de suelos 1:200000 de la provincia de Granada y memoria explicativa. Technical report, CSIC (1980)
29. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. of the American Statistical Association 66(336), 846–850 (1971)
30. Rauch, J., Simunek, M.: Mining for 4ft Association Rules. In: Morishita, S., Arikawa, S. (eds.) DS 2000. LNCS (LNAI), vol. 1967, pp. 268–272. Springer, Heidelberg (2000)
31. Runkler, T.A.: Comparing Partitions by Subset Similarities. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) IPMU 2010. LNCS, vol. 6178, pp. 29–38. Springer, Heidelberg (2010)
32. Sánchez, D., Delgado, M., Vila, M.A., Chamorro-Martínez, J.: On a non-nested level-based representation of fuzziness. Fuzzy Sets and Systems 192, 159–175 (2012)
33. Shortliffe, E., Buchanan, B.: A model of inexact reasoning in medicine. Mathematical Biosciences 23, 351–379 (1975)