Ad Hoc Metric for Correspondence Analysis Between Fuzzy Partitions

Carlos Molina[1], María D. Ruiz[2], Daniel Sánchez[3], and José M. Serrano[1(✉)]

[1] University of Jaén, Jaén, Spain
{carlosmo,jschica}@ujaen.es
[2] University of Cádiz, Cádiz, Spain
mariadolores.ruiz@uca.es
[3] University of Granada, Granada, Spain
daniel@decsai.ugr.es

Abstract. Correspondence analysis is a very common and renowned statistical technique, with applications in data summarization, classification, regression, etc. One particular approach is that of comparing different partitions over the same set of objects. Moreover, it can be interesting to analyze correspondences at different detail levels, not only between partitions, but between classes in these partitions. In addition, the case of fuzzy partitions over data is still a researching milestone in development. In this work we propose a novel measure following a previous definition of an alternate methodology in terms of data mining tools, in order to overcome some limitations of the former one for the case of considering partial and global correspondences between fuzzy partitions.

Keywords: Fuzzy correspondence analysis · Fuzzy partitions comparison · Ad hoc metrics

1 Introduction

Correspondence analysis [3] is a well-known statistical technique that can be commonly applied to obtain and describe existing relations between two categorical variables. It is a helpful tool for data dimensionality reduction, as an initial step before more complex processes such as classification, regression, discriminant analysis, etc. Further extensions and applications of this technique can be found throughout the literature [9,12].

Nevertheless, since it is based on distances and graphical representations, the interpretation can be subjective and sometimes confusing. As a way to overcome this, an alternative to classical correspondence analysis based on data mining techniques was introduced in [19]. This approach allows to obtain local, partial, and global correspondences, according to the required detail level. In contrast to the usual graphical interpretation of distances, correspondences are expressed in terms of data mining tools such as association rules and approximate dependencies, and as a consequence, we can apply the same metrics to interpret and measure the original correspondences.

S. Moral et al. (Eds.): SUM 2017, LNAI 10564, pp. 412–419, 2017.
DOI: 10.1007/978-3-319-67582-4_31

Furthermore, it must be taken into account the fact that in most of real world problems, unclear boundaries between partitions can be found, as some particular elements, due to their nature, may belong to more than one class, with different degrees, inside a same partition. Fuzzy logic allows us to extend existing techniques such as classification, clustering, etc., in order to cope with this issue. As a result, techniques for comparing sets of partitions have been extended in the same way. Renowned metrics as the Rand [17] or Jaccard indices [14] meet their counterparts in fuzzy contexts as, for example, approaches as those of Campello [8], Frigui et al. [11], Brouwer [6], Hüllermeier and Rifqi [13] and Anderson et al. [2]. In [1] the reader may find a more extensive comparison of the cited indices.

Similarly, in [7] the mentioned methodology for correspondence analysis in terms of data mining tools is extended to the fuzzy case, and in [16] an initial comparison with some of the previous measures is discussed. Nevertheless, as it is discussed in [7], some restrictions apply in the original definition of fuzzy partial and global correspondences, as non-atomic values (i.e., elements belonging to more than one partition) are not fully allowed. This paper is intended to continue this research line, introducing an ad hoc measure, in order to overcome the cited drawback. The document is structured as follows. After this introduction, the original proposal for (fuzzy) correspondence analysis in terms of data mining tools is recalled. Following this, we define our new index, and some examples of use are discussed. Concluding remarks as well as future works proposals end the paper.

2 Correspondences as Data Mining Tools

Correspondence analysis is usually applied as an early stage for integration or fusion of different classifications over a same set of objects. In classical correspondence analysis, partitions are displayed by means of a contingency table. Instead, we represent partitions by means of a relational table. For sake of brevity, we will refer directly to the fuzzy case, since the crisp case is easy to particularize from the former one.

Let O be a finite set of objects, and $\widetilde{\mathcal{P}} = \{\widetilde{P}_1, \ldots, \widetilde{P}_p\}$ and $\widetilde{\mathcal{Q}} = \{\widetilde{Q}_1, \ldots, \widetilde{Q}_q\}$ be two fuzzy partitions over O. Let $\widetilde{T}_{\widetilde{\mathcal{P}}\widetilde{\mathcal{Q}}}$ be the fuzzy transactional table associated to O, where each transaction represents an object, that is, $|\widetilde{T}_{\widetilde{\mathcal{P}}\widetilde{\mathcal{Q}}}| = |O|$. Table 1 shows an example of representation (let us remark that, in this particular example, partitions are not in Ruspini form). Given $o \in O$, $\widetilde{P}_i \in \widetilde{\mathcal{P}}$ and $\widetilde{Q}_j \in \widetilde{\mathcal{Q}}$, we noted for $\widetilde{P}_i(o)$ (respectively, $\widetilde{Q}_j(o)$) the membership degree of o in $\widetilde{P}_i$ (respectively, $\widetilde{Q}_j$). Each object must belong to at least one class of each partition, that is, $\forall o \in O, \exists \widetilde{P}_i \in \widetilde{\mathcal{P}} / \widetilde{P}_i(o) > 0$, and each class must contain at least one object, that is, $\widetilde{P}_i$, $\widetilde{Q}_j \neq \emptyset$. Let us note that, for sake of simplicity, each class in $\widetilde{\mathcal{P}}$ (resp. $\widetilde{\mathcal{Q}}$) can be associated to a single column. Without loss of generality, we can say that columns $\widetilde{P}_1 \ldots \widetilde{P}_p$ (resp. $\widetilde{Q}_1 \ldots \widetilde{Q}_q$) represent the set of possible classes in $\widetilde{\mathcal{P}}$ (resp. $\widetilde{\mathcal{Q}}$).

Table 1. Example of fuzzy transactional table $\widetilde{T}_{\widetilde{\mathcal{P}}\widetilde{\mathcal{Q}}}$

O	$\widetilde{\mathcal{P}}$			$\widetilde{\mathcal{Q}}$		
	$\widetilde{P}_1$	$\widetilde{P}_2$	$\widetilde{P}_3$	$\widetilde{Q}_1$	$\widetilde{Q}_2$	$\widetilde{Q}_3$
o_1	0.81	0	0	0.47	0.63	0
o_2	0.72	0.35	0	0	0.93	0
o_3	0.41	0.65	0	0	1.0	0
o_4	0.09	0.9	0	0	1.0	0.02
o_5	0	0.69	0.1	0	0.78	0.51
o_6	0	0	0.7	0	0.52	0.89
o_7	0	0	0.89	0	0.02	0.63

Let us remark that this approach allows us to consider not only perfect correspondences, but also those with possible exceptions. Hence, we are concerned with measuring the accuracy of correspondences between partitions.

2.1 Local, Partial, and Global Correspondences

Due to space restriction issues, we will recall only the definitions regarding the fuzzy case. A complete discussion about crisp correspondence analysis by means of data mining tools can be found in [19]. One of the advantages of this approach is that correspondences can be measured with the same metrics as those of data mining tools. In particular, certainty factor [20] returns a value between -1 (perfect, negative correspondence) and 1 (perfect, positive correspondence).

This methodology was later extended in order to manage correspondences between fuzzy partitions in [7]. Representing fuzzy partitions as in Table 1, the following types of fuzzy correspondences can be defined.

Definition 1 ([7]) ***Fuzzy local correspondence.*** *Let $\widetilde{P}_i \in \widetilde{\mathcal{P}}$ and $\widetilde{Q}_j \in \widetilde{\mathcal{Q}}$. There exists a fuzzy local correspondence from $\widetilde{P}_i$ to $\widetilde{Q}_j$, noted $\widetilde{P}_i \Rightarrow \widetilde{Q}_j$, if $\widetilde{P}_i \subseteq \widetilde{Q}_j$, that is, $\forall o \in O$, $\widetilde{P}_i(o) \leq \widetilde{Q}_j(o)$.*

Fuzzy local correspondences can be obtained in terms of fuzzy association rules (e.g., following the formal model proposed in [10]). Fuzzy partial and global correspondences were defined as well, following the model for fuzzy approximate dependencies introduced in [5]. But, as it is addressed in [7], in these cases, we must manage not classes, but partitions. It would be necessary to define an overall membership degree of an object regarding a whole partition, that is, $\widetilde{\mathcal{A}}(o)$. This issue introduced a multidimensionality problem and, hence, objects were limited to belong to only one class in every partition, for example, that one with the highest membership degree.

Definition 2 ([7]) ***Fuzzy partial correspondence.*** *There exists a fuzzy partial correspondence from $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$, noted $\widetilde{\mathcal{P}} \Rrightarrow \widetilde{\mathcal{Q}}$, when $\forall \widetilde{P}_i \in \widetilde{\mathcal{P}}$ $\exists \widetilde{Q}_j \in \widetilde{\mathcal{Q}}$ such that $\widetilde{P}_i \subseteq \widetilde{Q}_j$, that is, $\forall o \in O / t_o[\widetilde{\mathcal{P}}] = \widetilde{P}_i$ implies $t_o[\widetilde{\mathcal{Q}}] = \widetilde{Q}_j$ and $\widetilde{\mathcal{P}}(o) \dot{\leq} \widetilde{\mathcal{Q}}(o)$.*

$\dot{\leq}$ defines a vectorial order relation that, for this particular case, corresponds to a classic order relation. Finally, the step from fuzzy partial correspondences to fuzzy global correspondences is straightforward.

Definition 3 ([7]) ***Fuzzy global correspondence.*** *There exists a fuzzy global correspondence between $\widetilde{\mathcal{P}}$ and $\widetilde{\mathcal{Q}}$, noted $\widetilde{\mathcal{P}} \equiv \widetilde{\mathcal{Q}}$, when $\widetilde{\mathcal{P}} \Rrightarrow \widetilde{\mathcal{Q}}$ and $\widetilde{\mathcal{Q}} \Rrightarrow \widetilde{\mathcal{P}}$.*

In order to continue and complete this approach, in the following section we propose a new index, specifically intended for measuring fuzzy partial (and global) correspondences between fuzzy partitions.

3 Ad Hoc Index for Fuzzy Partial Correspondences

According to Definition 2, there is a fuzzy partial correspondence between two partitions, when we find that classes from the first partition are included, to some extent, in classes from the second partition. Hence, if we are capable of measure these inclusions for each pair of classes and aggregate the obtained values into a general index, we could measure a partial (and later, global) correspondence between these two partitions. With this idea in mind, we define our index as follows:

Definition 4. *Let $O = \{o_1, \ldots, o_n\}$, be again a set of objects, with $\widetilde{\mathcal{P}} = \{\widetilde{P}_1, \ldots, \widetilde{P}_p\}$ and $\widetilde{\mathcal{Q}} = \{\widetilde{Q}_1, \ldots, \widetilde{Q}_q\}$, two fuzzy partitions over O. There is a partial correspondence from $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$ when all classes from partition $\widetilde{\mathcal{P}}$ are included in classes from $\widetilde{\mathcal{P}}$, to some extent, which we measure by means of the following index:*

$$adhoc(\widetilde{\mathcal{P}}, \widetilde{\mathcal{Q}}) = AGGR_{i=1}^{p}\left(\bigoplus_{j=1}^{q}\left(AVG_{k=1}^{n}\left(\widetilde{\mathcal{P}}_i(o_k) \otimes \widetilde{\mathcal{Q}}_j(o_k)\right)\right)\right) \tag{1}$$

where $\otimes$ is a t-norm, $\oplus$ a t-conorm, $AGGR$ is an aggregation operator, and AVG is an averaging operator.

The reasoning behind this definition is that, for each pair $\widetilde{\mathcal{P}}_i \in \widetilde{\mathcal{P}}, \widetilde{\mathcal{Q}}_j \in \widetilde{\mathcal{Q}}$, we check to what extent is the former one included in the latter one according to all objects in O, by means of the t-norm $\otimes$. In our experiments we have considered $a \otimes b = min(a, b)$. Next, by means of an averaging operator (in our case, an average mean), we aggregate all these values for each $\widetilde{\mathcal{Q}}_j \in \widetilde{\mathcal{Q}}$ in order to obtain an estimated inclusion degree. Among all these degrees, we select the most representative one for each $\widetilde{\mathcal{P}}_i \in \widetilde{\mathcal{P}}$ (we took $\oplus = \max$). Finally, we obtain our index as an aggregation ($AGGR = sum$, in our case) of the previous values. The closer the value to 1, the more similar the partitions are. In fact, $adhoc(\widetilde{\mathcal{P}}, \widetilde{\mathcal{Q}}) = 1$, if $\widetilde{\mathcal{P}} = \widetilde{\mathcal{Q}}$. Algorithm 1 describes the process in a more formal way.

It must be remarked that, reviewing the literature, a similar index has been already proposed by Beringer and Hüllermeier in [4], where similarities between classes within partitions, instead of objects, are taken into account.

4 Experiments

As an initial but illustrative example, let us remember the example shown in Table 1. Following the original approach for fuzzy partial correspondences introduced in [7], a certainty factor $CF = 0.80$ (resp., 0.20) was returned for the fuzzy partial correspondence $\widetilde{\mathcal{P}} \Rrightarrow \widetilde{\mathcal{Q}}$ (resp., $\widetilde{\mathcal{Q}} \Rrightarrow \widetilde{\mathcal{P}}$). Our index returned a value of 0.839 (resp., 0.641). Apart from this, we have compared different set of partitions. Starting from randomly generated values, we compare a 5-classes fuzzy partition with a 7-classes one over an hypothetical set of 400 objects. Let $\widetilde{\mathcal{A}_5}$ be the former one, and $\widetilde{\mathcal{A}_7}$, the latter one. We measured fuzzy partial correspondence $\widetilde{\mathcal{A}_5} \Rrightarrow \widetilde{\mathcal{A}_7}$ (resp. $\widetilde{\mathcal{A}_7} \Rrightarrow \widetilde{\mathcal{A}_5}$) with a value of our index of $adhoc(\widetilde{\mathcal{A}_5}, \widetilde{\mathcal{A}_7}) = 0.571$ (resp. 0.787). This first experimental instance was mainly intended to test the behavior of the metric.

Table 2. Fuzzy partitions computed over wiki4HE dataset

	$\widetilde{\mathcal{W}_1}$	$\widetilde{\mathcal{W}_2}$	$\widetilde{\mathcal{W}_3}$	$\widetilde{\mathcal{W}_4}$
Distance	Euclidean		Manhattan	
Clusters	19	11	19	11
Error	1.3715	3.2144	0.4853	0.8590

In second place, we took *wiki4HE* Dataset [15] from UCI Machine Learning Repository, and applied different FCM (R package *e1071*) executions in order to generate different partitions (Table 2). Two different metrics (Euclidean and Manhattan) were applied, and for each one, two possible partitions were computed, with different number of classes. It is expected that, since both metrics are relatively similar, our index should reflect this with a high value. Moreover, high values for fuzzy partial correspondences are expected from more detailed (higher number of classes) partitions to more general (lower number of classes) ones, and vice versa.

Our index, together with the proposed one in [4], were computed between those partitions, in order to measure the fuzzy partial correspondences between them. The results are summarized in Table 3, the first value being that of our index, and the second one, Beringer and Hüllermeier's.

Table 3. Fuzzy partial correspondences between partitions (row $\Rrightarrow$ column)

	$\widetilde{\mathcal{W}_1}$	$\widetilde{\mathcal{W}_2}$	$\widetilde{\mathcal{W}_3}$	$\widetilde{\mathcal{W}_4}$
$\widetilde{\mathcal{W}_1}$	1.000/1.000	0.998/0.974	0.819/0.974	0.988/0.974
$\widetilde{\mathcal{W}_2}$	0.468/0.943	1.000/1.000	0.476/0.943	0.763/0.943
$\widetilde{\mathcal{W}_3}$	0.853/0.971	0.955/0.971	1.000/1.000	0.955/0.971
$\widetilde{\mathcal{W}_4}$	0.483/0.947	0.903/0.947	0.521/0.947	1.000/1.000

It must be noticed how fuzzy partial correspondences $\widetilde{\mathcal{W}_1} \Rrightarrow \widetilde{\mathcal{W}_2}$ and $\widetilde{\mathcal{W}_3} \Rrightarrow \widetilde{\mathcal{W}_4}$ are strong (index value close to 1), since the latter ones are summarizations of the former ones. That is, a reduction in the number of clusters induces that the former clusters are included, to some degree, in the latter ones. The opposite correspondences have a lower index value, which, according to the previous reasoning, seems logical. Since this issue is not detected in Beringer and Hüllermeier's proposal, whose index shows similar values for each pair of partitions, a deeper study should be conducted in order to explain it.

Finally, we also computed our index over the same partitions considered in [7], and found an interesting issue; our ad hoc index returned a value higher than 1. This could be due to the fact that one of the partitions was not in Ruspini [18] form. This situation may suggest that fuzzy operators in Eq. 1 needs to be properly adjusted.

Algorithm 1. Algorithm AdHoc

Input : $O = \{o_1, \ldots, o_n\}$, a set of objects, $\widetilde{\mathcal{P}} = \{\widetilde{P}_1, \ldots, \widetilde{P}_p\}$ and $\widetilde{\mathcal{Q}} = \{\widetilde{Q}_1, \ldots, \widetilde{Q}_q\}$, two fuzzy partitions over O.
Output: $adhoc(\widetilde{\mathcal{P}}, \widetilde{\mathcal{Q}})$, measure of the fuzzy partial correspondence from $\widetilde{\mathcal{P}}$ to $\widetilde{\mathcal{Q}}$, $\widetilde{\mathcal{P}} \Rrightarrow \widetilde{\mathcal{Q}}$.

$V_P \leftarrow \emptyset$
foreach $\widetilde{P}_i \in \widetilde{\mathcal{P}}$ **do**
  $V_Q \leftarrow \emptyset$
  foreach $\widetilde{Q}_j \in \widetilde{\mathcal{Q}}$ **do**
    /* Consider how $\widetilde{P}_i$ is included in every $\widetilde{Q}_j$ according to O */
    $V_Q[j] \leftarrow AVG_{o \in O}\left(\left(\widetilde{P}_i(o) \otimes \widetilde{Q}_j(o)\right)\right)$
  end
  /* For each $\widetilde{P}_i$, select the most representative value in V_q */
  $V_P[i] \leftarrow \bigoplus_1^q (V_Q[j])$
end
/* Finally, aggregate all values in V_P */
$adhoc(\widetilde{\mathcal{P}}, \widetilde{\mathcal{Q}}) \leftarrow AGGR_{i=1}^{p} (V_P[i])$

5 Concluding Remarks and Further Works

In this work our intention has been to continue a previous methodology for fuzzy correspondence analysis. To this purpose, we have proposed a new *ad hoc* index (in absence of a better name) to measure fuzzy partial, and global, correspondences between two fuzzy partitions, based on the extent to which the classes of a partition are included in the classes of the second partition, according to every object in a collection. First experiments suggest that the obtained results seem reasonable (values close to 1 where expected, and vice versa), although a deeper analysis, interesting properties study, and comparison with existing indices is still pending in order to validate and refine our proposal. They will be properly addressed in a future extension of this paper.

Acknowledgements. This work has been partially supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund (FEDER) under projects TIN2015-64776-C3-1-R and TIN2014-58227-P, and by the Energy IN TIME project funded from the European Union in the Seventh Framework Programme under grant agreement No. 608981.

References

1. 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). doi:10.1007/978-3-642-14055-6_46
2. Anderson, D.T., Bezdek, J.C., Popescu, M., Keller, J.M.: Comparing fuzzy, probabilistic, and possibilistic partitions. IEEE Trans. Fuzzy Syst. **18**(5), 906–918 (2010)
3. Benzécri, J.: Cours de Linguistique Mathématique. Faculté des Sciences, Université de Rennes (1964)
4. Beringer, J., Hüllermeier, E.: Fuzzy clustering of parallel data streams. In: Advances in Fuzzy Clustering and Its Application, pp. 333–352 (2007)
5. Berzal, F., Blanco, I., Sánchez, D., Serrano, J., Vila, M.: A definition for fuzzy approximate dependencies. Fuzzy Sets Syst. **149**(1), 105–129 (2005). Fuzzy Sets in Knowledge Discovery
6. Brouwer, R.K.: Extending the rand, adjusted rand and jaccard indices to fuzzy partitions. J. Intell. Inform. Syst. **32**(3), 213–235 (2009)
7. Calero, G., Delgado, J., Serrano, J., Sánchez, D., Vila, M.: A proposal of fuzzy correspondence analysis based on flexible data mining techniques. In: Soft Methodology and Random Information Systems, pp. 447–454. Springer, Heidelberg (2004)
8. Campello, R.J.: A fuzzy extension of the rand index and other related indexes for clustering and classification assessment. Pattern Recogn. Lett. **28**(7), 833–841 (2007)
9. Cox, M.A.A., Cox, T.F.: Multidimensional Scaling, pp. 315–347. Springer, Heidelberg (2008)
10. Delgado, M., Ruiz, M., Sánchez, D., Serrano, J.: A formal model for mining fuzzy rules using the RL representation theory. Inform. Sci. **181**(23), 5194–5213 (2011)
11. Frigui, H., Hwang, C., Rhee, F.C.-H.: Clustering and aggregation of relational data with applications to image database categorization. Pattern Recogn. **40**(11), 3053–3068 (2007)
12. Greenacre, M.: Correspondence Analysis in Practice, 3rd edn. CRC Press, Boca Raton (2016)
13. Hullermeier, E., Rifqi, M., Henzgen, S., Senge, R.: Comparing fuzzy partitions: a generalization of the rand index and related measures. IEEE Trans. Fuzzy Syst. **20**(3), 546–556 (2012)
14. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles **37**, 547–579 (1901)
15. Meseguer Artola, A., Aibar Puentes, E., Lladós Masllorens, J., Minguillón Alfonso, J., Lerga Felip, M.: Factors that influence the teaching use of wikipedia in higher education (2014)

16. Molina, C., Prados, B., Ruiz, M.-D., Sánchez, D., Serrano, J.-M.: Comparing partitions by means of fuzzy data mining tools. In: Hüllermeier, E., Link, S., Fober, T., Seeger, B. (eds.) SUM 2012. LNCS (LNAI), vol. 7520, pp. 337–350. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33362-0_26
17. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**(336), 846–850 (1971)
18. Ruspini, E.H.: A new approach to clustering. Inform. Control **15**(1), 22–32 (1969)
19. Sánchez, D., Serrano, J.M., Vila, M.A., Aranda, V., Calero, J., Delgado, G.: Using data mining techniques to analyze correspondences between user and scientific knowledge in an agricultural environment. In: Piattini, M., Filipe, J., Braz, J. (eds.) Enterprise Information Systems IV, pp. 75–89. Kluwer Academic Publishers, Hingham (2003)
20. Shortliffe, E., Buchanan, B.: A model of inexact reasoning in medicine. Math. Biosci. **23**, 351–379 (1975)