Graphical exploratory analysis of fuzzy data as a teaching tool

Inés Couso, Luis Junco, José Otero, and Luciano Sánchez

Abstract

Graphical exploratory analysis for fuzzy data allows us to represent sets of individuals whose attributes are perceived with imprecision on a map so that the degree of dissimilarity between two objects is somehow compatible with the distances between their respective representations. This study will discuss the use of this tool to jointly analyze the evolution of a group of students during a course, and to select the most suitable personnel of a company to receive a training course, according to a catalog of competencies and considering the reliability of information sources.

Key words: Exploratory graphical analysis, coarse data, imprecise statistical data, teaching tool, possibility measures, fuzzy sets

1 Introduction

Graphical exploratory analysis consists of the projection of a set of individuals on a plane, so that the similarities between pairs of individuals are compatible with the distances between their corresponding representations [9]. There are different techniques to perform this projection, depending on the model used to link the similarities between objects to the distances between their representations on the map. The most frequent method is applied to objects described by a vector of numerical properties, and uses the Euclidean distance both to calculate the distances between objects and between their projections.

Different generalizations of graphical exploratory analysis to the case of imprecise data have been considered in the literature. More concretely, the

University of Oviedo couso@uniovi.es, lajunco@uniovi.es, jotero@uniovi.es, luciano@uniovi.es

case where every instance is characterized by means of a vector of fuzzy numbers has been considered by different authors [5, 6, 7, 13]. Just as in the exploratory analysis of crisp data each individual is associated with a point on the map, the projection of a fuzzy vector is a geometric figure that depends on the transformations between the spatial distances and the distances on the map.

Exploratory data analysis techniques are part of the general knowledge and routinely used in a multitude of knowledge discovery problems. However, the use of exploratory analysis of fuzzy data is not very widespread. In relation to information mining in a teaching context, the algorithm defined in [11] has been applied to the analysis of tests solved by groups of children with learning difficulties (early diagnosis of dyslexia [12]) and to the analysis of questionnaires of follow-up by the students of different undergraduate and master's degrees. In both cases, in addition to obtaining visual information about how many different types of students are in each group, it is possible to measure the variations between the learning success of the different subgroups, and the evolution of their relative returns over the course of the course. This study reviews these applications and introduces a new one, selecting the right people to attend in-company training courses, taking into account different sources of information about each employee's competencies (statement internal examinations, questionnaires, previous work, etc.) which are, in turn, affected by the credibility of each source.

The rest of the paper is organized as follows: Section 2 describes the usefulness of graphical data analysis techniques in a teaching context. Section 3 discusses the need for a representation based on fuzzy sets, and describes some technical aspects of the algorithm. Section 4 discusses several case studies. The paper ends with some concluding remarks and future work.

2 Usefulness of graphical exploratory analysis in teaching problems

When organizing training courses for employees of a company, it is important to study what training skills need to be filled, so that you can choose the most appropriate content for the courses. A similar study also serves to select the best attendees for a course with a limited number of places, or to compare the results of the study before and after the end of the course, in order to check whether this effort is paying off and translates into an improvement of the global capacity of the company.

As mentioned in the Introduction, the different graphical exploratory analysis algorithms (Sammon maps, PCA, Multidimensional Scaling (MDS), Self Organized Maps (SOM), etc. [5]) are statistical techniques that project objects as points on a plane, so that the proximity of the projections of two objects (instances) on the map reflects the similarity between their respective properties (seen as functions of the corresponding vectors of attributes). If each of these objects is associated with an employee, and the properties of the employee are assumed to consist of numerical measures of their proficiency level in a competency catalog, the graphic exploratory analysis is one of the most adequate techniques to evidence groups of employees with different skills. By adding fictional individuals (hypothetical employees with perfect knowledge of one technology and none other than other technologies), the positions of actual employees over those of fictitious employees make it possible to detect training gaps. Finally, comparing several maps of the same individuals on different dates, it is possible to evaluate the impact of the courses received.

Notwithstanding, these techniques are not directly applicable to the case where there is some uncertainty about the values of some of the attributes of an individual (eg missing data) and also do not consider the reliability of the different sources of information used to characterize individuals. For example, one of the most accessible sources for checking the level of formation of a group is the follow-up questionnaire [10]. Unlike the exam or interview, it is the student or employee who declares their knowledge, so this information may be inaccurate: an individual can either declare an “advanced” knowledge of English in the curriculum or having passed an examination of that level; in the first case, the uncertainty about her language proficiency is greater.

A simple way to quantify the uncertainty associated with a questionnaire is to associate different questions with the same item (a value of a single attribute for a single individual); the dispersion of the corresponding responses is an indication of the reliability of the test. For example, if a student is approached about his or her knowledge of probability theory, he may declare a “high” knowledge of the concept of the “density function” and “null” knowledge about the notion of “Radon-Nikodym derivative”, while another one can answer “medium” to both questions: the dispersion of the answers associated with the same item is an indication of their reliability. Clearly, if just a central location measure is selected in order to summarize all the responses associated with the same property, valuable information is lost.

Another frequent problem with the data collection phase is the problem with missing data [8]. The most frequent solution are either to remove the individual from the sample or to follow some imputation technique. The latter is the preferred solution when the sample size is not sufficiently high, and generally consists in finding the closest individuals to calculate their average values. Again, the variability of these values is being ignored, which may mean that the distribution of the completed data is probably far from reality, so the analysis would be distorted. Other imputation methods do not affect the variance but only work well under some assumptions about the coarsening process [2, 3].

3 Use of fuzzy sets in competency analysis

The use of fuzzy sets allows a homogenous representation of all previous types of uncertainty in the data. By means of the possibilistic interpretation of the membership function of a fuzzy set [4, 1], each value of the attribute for a specific object can be associated with a fuzzy number $\tilde{X}$ whose α-cuts are interrpreted as nested confidence intervals in the sense that:

$$P([\tilde{X}]_\alpha \ni x) \geq 1 - \alpha, \ \forall \, \alpha \in (0, 1).$$

Thus, for example, a missing value can be replaced by a fuzzy set that models the distribution of the corresponding attribute in other similar objects (even though these in turn are perceived imprecisely), and the statements of a knowledge “Advanced” English or “Average” theory of probability will be associated with two fuzzy numbers, whose specificities will be linked to the reliabilities of information sources. Therefore, in this study it will be considered that the knowledge about an individual can be quantified by means of a vector of fuzzy numbers. This generalization has two fundamental consequences in order to perform a graphical analysis of the population:

1. The spatial coordinates of each individual are unknown, except for a nested family of (multivariate) confidence intervals.
2. The Euclidean distance between two individuals whose coordinates are uncertain, is uncertain in turn.

From the first statement it can be concluded that the projections of each individual on the map will not be points, but families of nested sets, whose form will depend on the distortion of the spatial geometry, proper to each technique, in the flat projection. From the second, it follows that it is not reasonable to use a distance between fuzzy sets and calculate a numerical array of distances between individuals, nor between their projections. In general, a distance like that will not induce a total order among projections that is consistent with the ordering between the actual values of the attributes, since such an order between the actual values is just partially known.

In previous works, different simplifications have been made to achieve an approximate projection. For example, in the method described in [5, 6] multidimensional scaling (MDS) is extended to allow distance matrices to contain ranks or fuzzy numbers. The standard version of MDS consists of finding the scatter plot that minimizes a stress function, defined by the quadratic difference between the matrix of distances between the data and the matrix of distances of the points included in the scatter plot. In the generalized version, a fuzzy-valued stress function is defined, which measures the fit between the set of distances compatible with the map figures and the set of distances between the fuzzy descriptions of the individuals. In this method it is assumed that the projections are circles on the map, which is not always correct, since the attributes are not allowed to have different levels of uncertainty. Subsequent extensions [13] removed this restriction, as explained below.

3.1 How to determine the shape of projections

Let $\tilde{X}_i = (\tilde{X}_{i1} \times \ldots \times \tilde{X}_{if})$ and $\tilde{X}_j = (\tilde{X}_{j1} \times \ldots \times \tilde{X}_{jf})$ two tuples of fuzzy sets representing our incomplete knowledge about the f attributes of individuals number i and j. Let $[\tilde{X}_i]_\alpha = [x^-_{i1}, x^+_{i1}] \times \ldots \times [x^-_{if}, x^+_{if}]$ and $[\tilde{X}_j]_\alpha = [x^-_{j1}, x^+_{j1}] \times \ldots \times [x^-_{jf}, x^+_{jf}]$ be in turn two cuts at the same level α of $\tilde{X}_i$ y $\tilde{X}_j$.

The set of possible values for the distances at a $1-\alpha$-confidence level is:

$$D^\alpha_{ij} = \left\{ \sqrt{\textstyle\sum_{k=1}^{f} (x_{ik} - x_{jk})^2} \mid x_{ik} \in [x^-_{ik}, x^+_{ik}], x_{jk} \in [x^-_{jk}, x^+_{jk}], 1 \leq k \leq f \right\}. \qquad (1)$$

Some authors have used a distance similar to this one before [6], and further assumed that the shape of the projection of an imprecise case was a circle. We have found that, in our problem, this last is a too restrictive hypothesis. Instead, and according to [13], we propose to approximate the shape of the projections by a polygon (see Figure 1) whose radii R^+_{ij} and R^-_{ij} are not free variables, but depend on the distances between the cases.

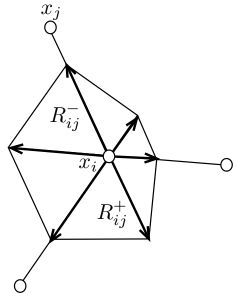

Fig. 1 The α-cuts of the projected data are polygons defined by the distances R_{ij} in the directions that pairwise join the examples.

Let us now consider a multivariate tuple of imprecise data $(\tilde{X}_1, \ldots, \tilde{X}_N)$, where $\overline{x}_i$ is the mode of $\tilde{X}_i$ and let $\{(z_{11}, \ldots, z_{1r}), \ldots, (z_{N1}, \ldots, z_{Nr})\}$ be the projection on a map of dimension r of that N-dimensional vector.

We propose that the radii R^+_{ij} and R^-_{ij} depend on the distance between $[\tilde{X}_i]_\alpha$ and $\overline{x}_j$ (see Figure 2 for a graphical explanation) as follows:

$$R^+_{ij} = d_{ij} \left(\frac{\delta^+_{ij}}{\delta_{ij}} - 1 \right) \quad R^-_{ij} = d_{ij} \left(\frac{\delta_{ij}}{\delta^-_{ij}} - 1 \right) \qquad (2)$$

where $d_{ij} = \sqrt{\sum_{k=1}^{r}(z_{ik} - z_{jk})^2}$, $\delta_{ij} = \{d(\overline{x}_i, \overline{x}_j)\}$, $\delta_{ij}^{+} = \max\{d(x, \overline{x}_j) \mid x \in [\tilde{X}_i]_\alpha\}$, and $\delta_{ij}^{-} = \min\{d(x, \overline{x}_j) \mid x \in [\tilde{X}_i]_\alpha\}$.

3.2 Stress function

According to the above, the available knowledge about the value of the effort function associated with the projection of the data is given by the following fuzzy-valued function, defined by its cuts:

$$
\begin{aligned}
S_\alpha = \Bigg\{ & \sum_{i=1}^{N} \sum_{j=i+1}^{N} ||d(t,u) - \beta|| \mid \\
& t \in [\tilde{X}_i]_\alpha, u \in [\tilde{X}_j]_\alpha, \\
& \beta \in [d_{ij} - R_{ij}^{-} - R_{ji}^{-}, d_{ij} + R_{ij}^{+} + R_{ji}^{+}] \Bigg\}.
\end{aligned} \tag{3}
$$

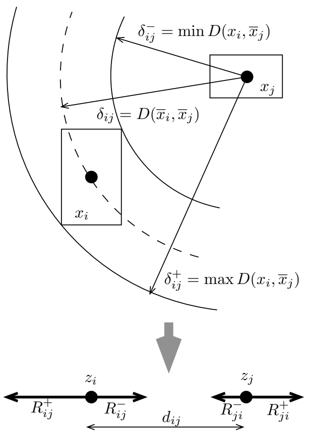
$\delta_{ij}^{-} = \min D(x_i, \overline{x}_j)$
x_j
$\delta_{ij} = D(\overline{x}_i, \overline{x}_j)$
x_i
$\delta_{ij}^{+} = \max D(x_i, \overline{x}_j)$
z_i
z_j
R_{ij}^{+}
R_{ij}^{-}
d_{ij}
R_{ji}^{-}
R_{ji}^{+}

Fig. 2 The distance between the respective projections of $[\tilde{X}_i]_\alpha$ and $[\tilde{X}_j]_\alpha$ is between the values $d_{ij} - R_{ij}^{-} - R_{ji}^{-}$ and $d_{ij} + R_{ij}^{+} + R_{ji}^{+}$.

As an alternative to minimizing the previous fuzzy-valued function, it is possible to define a measure that quantifies how different the collection of spatial distances and the collection of the distances between their projections are, in terms of their corresponding rankings.

Let $D(X_i, X_j)$ be the fuzzy set whose α-cuts are the sets of distances between individuals, and let $D'(Z_i, Z_j)$ be the set of distances between their respective projections, Z_i and Z_j.

If the map correctly reflects the distances between individuals of the population, it must be observed that the rank of the distance between the i -th and j -th objects are the same as the rank of the distance between their projections, in the corresponding matrix. Therefore, the number of pairs of objects for which this does not it is true that it defines an alternative cost function.

In our case, the rank of a fuzzy-valued distance within its own matrix is ??not completely defined, since there are non-comparable pairs of distances. However, for each level α a relation between the $\alpha-$ distances (which are intervals) can be defined; given two interval-valued distances $[d^-, d^+]$ and $[e^-, e^+]$, they are not comparable when

$$[d^-, d^+] \parallel [e^-, e^+] \iff (d^+ > e^-) \wedge (e^+ \geq d^-). \tag{4}$$

Otherwise, we can say that one of them precedes the other (i.e., either $[d^-, d^+] \prec [e^-, e^+]$ or $([e^-, e^+] \prec [d^-, d^+]$. The rank of a distance will be defined, for the α level, according to the following iterative procedure: We take all the distances and we select those that are not preceded by any other in the collection. All of them are assigned rank equal to 1. We remove those distances from the initial collection. We take the remaining ones and iterate the process, by assigning a rank equal to 2 to those that are not dominated by any other one. We continue with the process until we get the empty set.

The purpose of the numerical algorithm (which will not be made explicit, due to extension limitation) is to obtain a map for which the ranks of each of the terms of the matrices of distances between individuals and between projections coincide for every α level. The value of the stress function is the infimum of those α levels for which both collections of ranks do coincide.

4 Numerical and graphical results

In this section we will illustrate, with the help of three real-world datasets, how to identify groups of students and how to stack two maps from the same individuals at different times, for showing the temporal evolution of the learning.

4.1 Variation of individual capacities in the same group and between groups

In the left part of Figure 3 a diagram for 30 students of subject “Statistics” in *Ingenieria Telematica* at Oviedo University, taken at the beginning of the 2009-2010 course is shown. This survey is related to students’ previous knowledge in other subjects. In particular, this survey evaluates previous knowledge in Algebra (A), Logic (B), Electronics (C), Numerical Analysis

(D), Probability (E) and Physics (F). The positions of the characteristic points have been marked with labels. Those points are of the type “A” (all the questions about the subject “A” are correct, the others are erroneous) “NO A” (all the questions except “A” ones are correct, the opposite situation), etc.

In the right part of Figure 3 we have plotted together the results of three different groups, attending lectures by the same teacher. Each intensification has been coded with a distinctive colour. This teacher has evaluated, as before, the initial knowledge of the students in subjects that are a prerequisite. From the graphic in that figure the most relevant fact is that the students of the intensification coded in red (*Ingenieria Industrial*) consider themselves better prepared than those coded in blue (*Ingeniera Tecnica Industrial Electrica*), with the green group in an intermediate position, closer to red (*Ingeniera Tecnica Industrial Quimica*). All the students of all the groups have a neutral orientation to math subjects, and some students in the blue group think that their background is adequate only in subjects C (Operating Systems) and D (Internet).

4.2 Evaluation of learning results

Ten pre-doctoral students in Computer Science, Physics and Mathematics attending a research master were analyzed. The background of these students is heterogeneous. In the survey the students were asked about 36 subjects classified in “Control Algorithms” (A), “Statistical Data Analysis” (B), “Numerical Algorithms” (C) and “Lineal Models” (D). At the top of the figure 4 we can see that there is a large dispersion between the initial knowledges. Since the subject had strong theoretic foundations, students from technical degrees like Computer Science evaluated themselves with the lowest scores (shapes in the right part of each figure).

The same survey, at the end of the course, shows that all the students moved to the left, closer to characteristic point “EVERYTHING”. Additionally, the displacement has been larger for the students in the group at the right. This displacement can be seen clearly in the right part of the same figure, where the shapes obtained from the final survey were replaced by arrows that begin in the initial position and end in the final center. The length of the arrows is related with the progress of the student during the course.

5 Conclusions

We have proposed the use of graphical exploratory maps to analyze the characteristics of groups of students, when those attributes are observed with some uncertainty either due to inconsistencies in the collection of data or missing data. The map of a group consists of several figures and a list of characteristic points. The proximity of an individual to one of these points means that the balance of such an individual with respect to different areas of knowledge resembles the value represented by this indicator. This technique can be used to corroborate the improvement of the abilities after receiving a training course: combining in the same graph the results of two tests, separated in time, it is possible to determine the displacement of each individual towards other characteristic points, and thus to detect the individuals who have best taken advantage of the course.

Acknowledgements We thank the editors of this volume for their kind invitation to participate. The research in this work has been supported by projects TIN2014-56967-R and FC-15-GRUPIN14-073.

References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic vs. epistemic views. International Journal of Approximate Reasoning 55: 1502–1518.
2. Couso I, Dubois D, A general framework for maximizing likelihood under incomplete data, under review.
3. Couso I, Dubois D, Hüllermeier E, Maximum Likelihood Estimation and Coarse Data. In: Moral et al (eds) Proceedings of the 11th International Conference on Scalable Uncertainty Management, SUM 2017.
4. Couso I, Sánchez L (2008) Higher order models for fuzzy random variables, Fuzzy Sets and Systems 159: 237–258.
5. Denœux T, Masson M-H (2000) Multidimensional scaling of interval-valued dissimilarity data, Pattern Recognition Letters: 21, 83–92.
6. Hebert PA, Masson M-H, Denœux T (2006) Fuzzy multidimensional scaling. Computational Statistics and Data Analysis 51: 335–359.
7. Honda K, Ichihashi H (2006) Fuzzy local independent component analysis with external criteria and its application to knowledge discovery in databases, International Journal of Approximate Reasoning 42:159–173.
8. Kim W, Choi B, Hong E-K, Kim S-K (2003) A Taxonomy of Dirty Data, Data Mining and Knowledge Discovery 7: 81–99.
9. Kruskal JB (1964) Nonmetric multidimensional scaling: a numerical method. Psychometrika 29:115–129.
10. Nuhfer E, Knipp D (2006) The use of a knowledge survey as an indicator of student learning in an introductory Biology course, CBE life sciences education 5: 313–316.
11. Mazza R, Milani C (2005) Exploring usage analysis in learning systems: Gaining insights from visualisations. In: Proceedings of the Workshop on usage analysis in learning systems at 12th International Conference on Artificial Intelligence in Education, New York, USA 1–6.
12. Palacios A, Sánchez L, Couso I (2010) Diagnosis of dyslexia with low quality data with genetic fuzzy systems, International Journal on Approximate Reasoning 51: 993-1009.
13. Sánchez L, Couso I, Otero J, Palacios A (2010) Assessing the evolution of learning capabilities and disorders with a graphical exploratory analysis of surveys containing missing and conflicting answers. Neural Network World 20: 825–838.

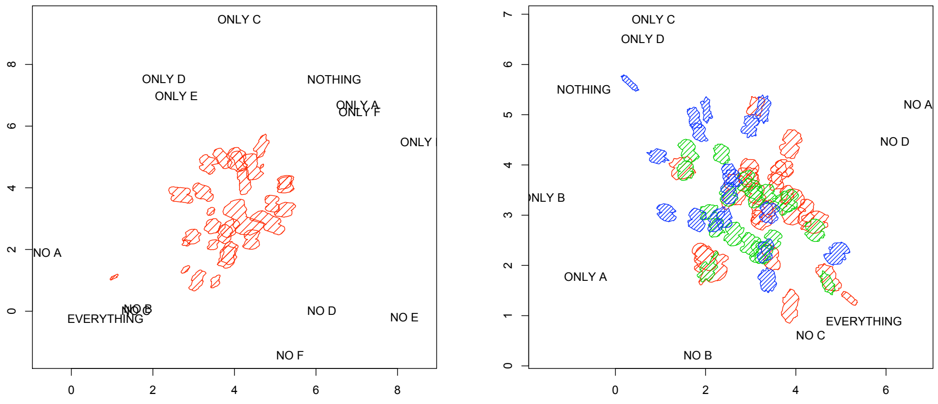

Fig. 3 Left part: Differences in knowledge of Statistics for students in Ingenieria Telematica. Right part: Differences in knowledge about Computer Science between the students of Ingenieria Tecnica Industrial specialized in Chemistry, Electricity and Mechanics.

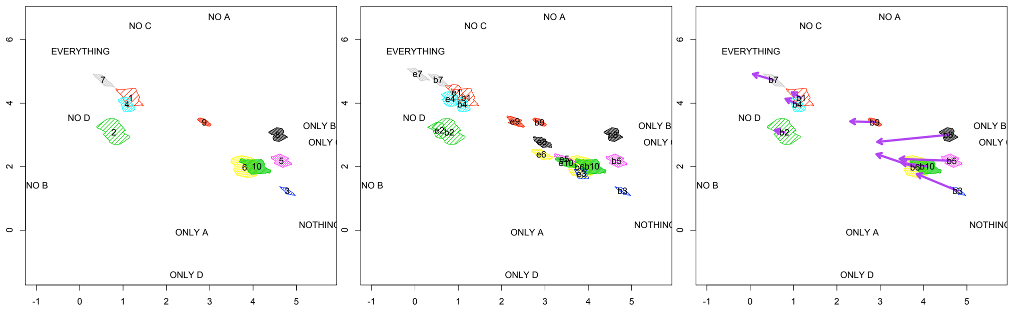

Fig. 4 Evolution of the learning of pre-doctoral students. Left part: Initial survey. Center: superposition of initial and final maps. Right part: The displacement has been shown by arrows.