# Graphical exploratory analysis of educational knowledge surveys with missing and conflictive answers using evolutionary techniques

Luciano Sánchez*, Inés Couso, and José Otero

Computer Science and Statistics Departments,
Universidad de Oviedo,
Campus de Viesques s/n Gijon (SPAIN)
{luciano,couso,jotero}@uniovi.es

**Abstract.** Analyzing the data that is collected in a knowledge survey serves the teacher for determining the student's learning needs at the beginning of the course and for finding a relationship between these needs and the capacities acquired during the course. In this paper we propose using graphical exploratory analysis for projecting all the data in a map, where each student will be placed depending on his/her knowledge profile, allowing the teacher to identify groups with similar background problems, segment heterogeneous groups and perceive the evolution of the abilities acquired during the course.
The main innovation of our approach consists in regarding the answers of the tests as imprecise data. We will consider that either a missing or unknown answer, or a set of conflictive answers to a survey, is best represented by an interval or a fuzzy set. This representation causes that each individual in the map is no longer a point but a figure, whose shape and size determine the coherence of the answers and whose position with respect to its neighbors determine the similarities and differences between the students.

**Key words:** Knowledge Surveys, Graphical Exploratory Analysis, Multidimensional Scaling, Fuzzy Fitness-based Genetic Algorithms

## 1 Introduction

Knowledge surveys comprise short questions that students can answer writing a single line, or choosing between several alternatives in a printed or web-based questionnaire [5]. These surveys can be used for assessing the quality of the learning and they are also meaningful from a didactical point of view. On the one hand, they allow students to perceive the whole content of the course [6]. On the other hand, teachers can use these surveys for deciding the best starting level for the lectures, specially in Master or pre-doctoral lectures [10], where the

* This work was funded by Spanish M. of Education, under the grant TIN2008-06681-C06-04.

profiles of the students attending the same course are much different. Recently this has also been applied to teacher education and certification [11]. When the survey is done at the end of the course, the effectivity of the teaching methodology along with the attitude and dedication of the students is measured. There is certain consensus in the literature in that the relationship between methodology/dedication and scoring is weak [2]. Because of this, a survey (different than an exam, designed to score the students) is needed. Finally, elaborating the survey serves by itself to establish the course contents, timeline and the teaching methodology [9].

In this context, this paper is about graphically analyzing the data that is collected in a knowledge survey. We intend to determine the student's learning needs at the beginning of the course and also to find a relationship between these needs and the capacities acquired during the course. To this end, we propose projecting the data in a map, where each student will be placed according to his/her knowledge profile, allowing the teacher to identify groups with similar background problems, segmenting heterogeneous groups and showing the evolution of the abilities acquired during the course.

This is not a new technique by itself, since these statistical methods (and, generally speaking, intelligent techniques) for analyzing questionnaires and surveys are part of the common knowledge [9]. Moreover, the proliferation of free data mining software (see, for instance [1]) has driven many advances in the application of Artificial Intelligence in educational contexts [7]. Indeed, there exist some tools that can generate views of the aforementioned data for easily drawing conclusions and making predictions about the course effectivity. The innovation of our approach is not in the use of graphical techniques but in extending them to data that is possibly incomplete or imprecise. This is rarely done when analyzing surveys and as a matter of fact the extension of graphical exploratory analysis to low quality data-based problems is very recent [3, 4]. As far as we know, these last techniques have not yet been applied in an educational context.

Notwithstanding, we believe that their use will make possible to solve better two frequent problems: the situation where the student does not answer some questions of the survey and the cases where there are incompatible answers that might have been carelessly answered. Within our approach, we will consider that a missing or unknown answer in the survey is best represented by an interval. For instance, if the answer is a number between 0 and 10, an unanswered question will be associated with the interval [0,10]. We will not try to make up a coherent answer for the incomplete test, but we will carry the imprecision in all the calculations. In turn, an incoherent set of answers will also be represented by an interval. For instance, assume that the same question is formulated in three different ways (this can be done for detecting random answers to the tests) and the student answers incoherent results. Let $\{6, 2, 4\}$ be the different answers to the question. With our methodology, instead of replacing this triplet by its mean, we will say that the answer is an unknown number in the range $[2, 6]$ (the minimum and the maximum of the answers).

Using intervals for representing unknown values produces that each individual in the map is no longer a point but a figure, whose shape and size determine the coherence of the answers and whose relative position determines the similarities between it and the other students. In this paper we will explain how this map can be generated with the help of interval (or fuzzy) valued fitness function-driven genetic algorithms. We will also show the results of this new analysis in three actual surveys, answered by Spanish engineering and pre-doctorate students.

The structure of this paper is as follows: in Section 2 we introduce Graphical Exploratory Analysis for vague data and its relation with knowledge surveys. In the same section we explain an evolutionary algorithm for computing these maps, and in Section 3 we show the results of this method in three real-world cases. The paper concludes in Section 4.

## 2 Graphical exploratory statistics

There are many different techniques for performing graphical exploratory analysis of data: Sammon maps, Principal Component Analysis (PCA), Multidimensional Scaling (MDS) , self-organized maps (SOM), etc. [3]. These methods project the instances as points in a low dimensional Euclidean space so that their proximity reflects the similarity of their variables. However, we have mentioned that the surveys can be incomplete or they possibly contain conflicting answers, and also that an incomplete survey can be taken as the set of all surveys with any valid value in place of the missing answer. Observe that, in that case, the projection will not be a point but a shape whose size will be larger as the more incomplete or imprecise the survey is.

This extension from a map of points to a map of shapes has already been done for some of the techniques mentioned before. For instance, Fuzzy MDS, as described in [3, 4], extends MDS to the case where the distance matrix comprises intervals or fuzzy numbers, as happens in our problem. Crisp MDS consists in finding a low-dimensional cloud of points that minimizes an stress function. That function measures the difference between the matrix of distances between the data and the matrix of distances between this last cloud. The interval (or fuzzy) extension of this algorithm defines an interval (fuzzy) valued stress function that bounds the difference between the imprecisely known matrix of distances between the objects and the interval (fuzzy) valued distance matrix between a set of shapes in the low-dimensional projection.

Let us assume for the time being that the distance between two surveys is an interval. For two imprecisely measured multivariate values $x_i = [x_{i1}^-, x_{i1}^+] \times \ldots \times [x_{if}^-, x_{if}^+]$ and $x_j = [x_{j1}^-, x_{j1}^+] \times \ldots \times [x_{jf}^-, x_{jf}^+]$, with $f$ features each, the set of distances between their possible values is the interval

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

Some authors have used a distance similar to this before [4], and further assumed that the shape of projection of an imprecise case is a circle. We have

found that, in our problem, this last is a too restrictive hypothesis. Instead, 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.

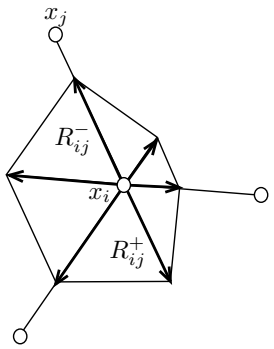


**Fig. 1.** The projected data are polygons defined by the distances $R_{ij}$ in the directions that pairwise join the examples.

For a multivariate set of imprecise data $\{x_1, \ldots, x_N\}$, let $\overline{x}_i$ be the crisp centerpoint of the imprecise value $x_i$ (the center of gravity, if an interval, or the modal point, if fuzzy), and let $\{(z_{11}, \ldots, z_{1r}), \ldots, (z_{N1}, \ldots, z_{Nr})\}$ be a crisp projection, with dimension $r$, of that set. We propose that the radii $R_{ij}^+$ and $R_{ij}^-$ depend on the distance between $x_i$ and $\overline{x}_j$ (see Figure 2 for a graphical explanation) as follows

$$R_{ij}^+ = d_{ij}\left(\frac{\delta_{ij}^+}{\delta_{ij}} - 1\right) \quad R_{ij}^- = d_{ij}\left(\frac{\delta_{ij}}{\delta_{ij}^-} - 1\right) \tag{2}$$

where $d_{ij} = \sqrt{\sum_{k=1}^{r}(z_{ik} - z_{jk})^2}$, $\delta_{ij} = \{D(\overline{x}_i, \overline{x}_j)\}$, $\delta_{ij}^+ = \max\{D(x_i, \overline{x}_j)\}$, and $\delta_{ij}^- = \min\{D(x_i, \overline{x}_j)\}$. We also propose that the value of the stress function our map has to minimize is

$$\sum_{i=1}^{N}\sum_{j=i+1}^{N} d_H(D_{ij}, [d_{ij} - R_{ij}^- - R_{ji}^-, d_{ij} + R_{ij}^+ + R_{ji}]^+)^2 \tag{3}$$

where $d_H$ is the Haussdorff distance between intervals.

### 2.1 Characteristic points

We also propose adding several prototypic surveys (we will call them "characteristic points") corresponding to a survey without mistakes, a completely wrong survey, one section well answered but the remaining ones wrong, etc. With the help of these points, the map can be used for evaluating the capacities of a student by comparing it with its closest characteristic point.

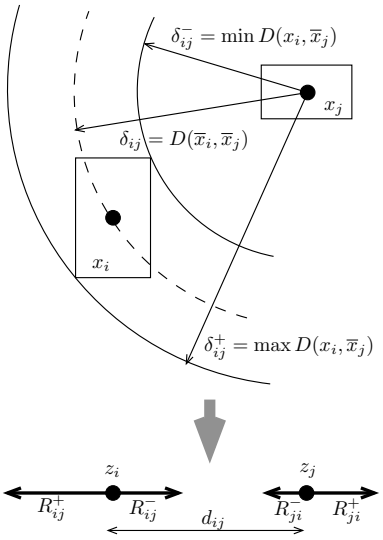


**Fig. 2.** The distance between the projections of $x_i$ and $x_j$ is between $d_{ij} - R_{ij}^- - R_{ji}^-$ and $d_{ij} + R_{ij}^+ - R_{ji}^+$

### 2.2 Evolutionary algorithm

An evolutionary algorithm is used for optimizing the stress function and searching the best map. In previous works we have shown that interval and fuzzy fitness functions can be optimized with extensions of multiobjective genetic algorithms. In this paper we have used the extended NGSA-II defined in [8], whose main components are summarized in the following paragraphs.

**Coding scheme** Each individual of the population represents a set of coordinates in the plane, thus each chromosome consists of the concatenation of so many pairs of numbers as students, plus one pair for each characteristic point (i.e. "Everything", "Nothing", "Only Subject X", "Every Subject but X", etc). The chromosome is fixed-length, and real coding is used.

**Objective Function** The fitness function was defined in eq. (3).

**Evolutionary Scheme** A generational approach with the multiobjective NSGA-II replacement strategy is considered. Binary tournament selection based on the crowding distance in the objective function space is used. The precedence operator derives from the bayesian coherent inference with an imprecise prior, the dominated sorting is based on the product of the lower probabilities of precedence, and the crowding in based on the Hausdorff distance, as described in [8].

**Genetic Operators** Arithmetic crossover is used for combining two chains. The mutation operator consists in performing crossover with a randomly generated chain.

## 3 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.

### 3.1 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.

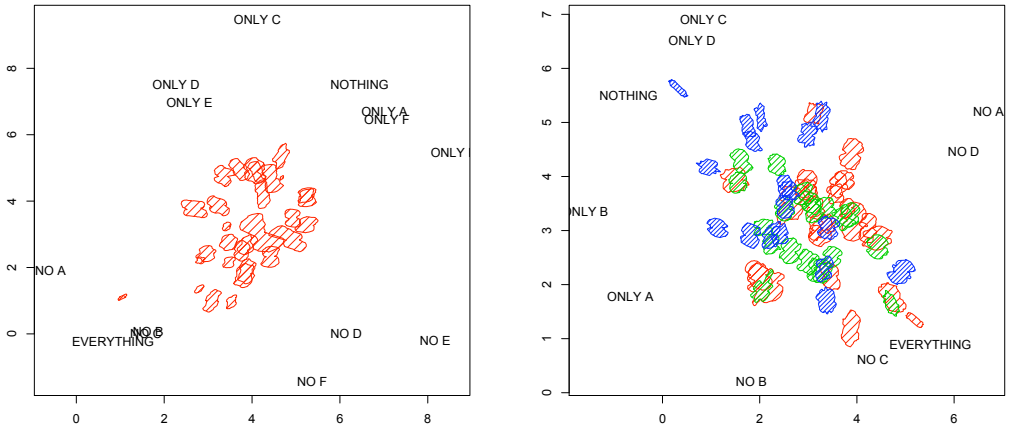


**Fig. 3.** Left part: Differences in knowledge of Statistics for students in Ingenieria Telematica. Right part: Differences in knowledge about Computer Science between the students of Ingenieria Tecnica Industrial specialized in Chemistry, Electricity and Mechanics.

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).

### 3.2 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).

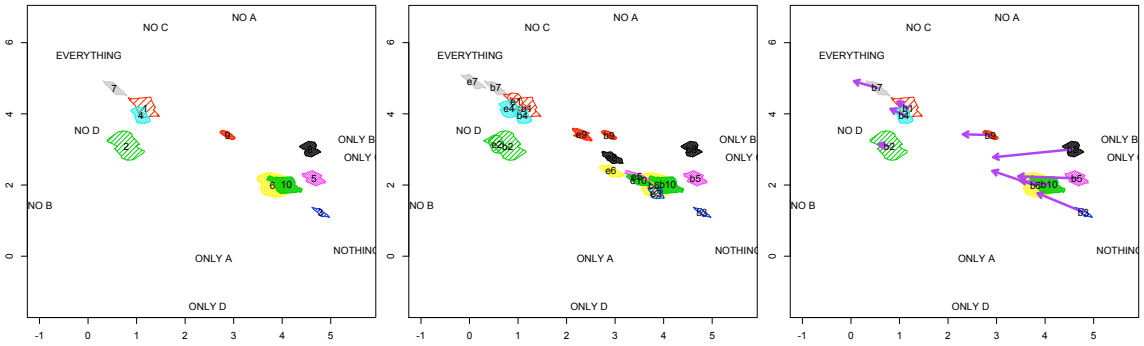


**Fig. 4.** Evolution of the learning of pre-doctoral students. Left part: Initial survey. Center: superposition of initial and final maps. Right part: The displacement has been shown by arrows.

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.

## 4 Conclusions

In this work we have extended with the help of a fuzzy fitness-driven genetic algorithm the Multidimensional Scaling to imprecise data, and exploited the new capabilities of the algorithm for producing a method able to process incomplete or carelessly filled surveys that include conflictive answers. The map of a group of students consists on several shapes, whose volume measures the degree to which a survey lacks consistency. We have shown that these maps can help detecting heterogeneous groups and can also be used for assessing the results of a course.

## References

1. Alcala-Fdez, L. et al. KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems. Soft Computing 13:3 (2009) 307-318.
2. Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. Review of Educational Research, 51 (3) 281-309.
3. Denoeux, T., Masson, M.-H., Multidimensional scaling of interval-valued dissimilarity data. Pattern Recognition Lett. 21, 83-92. 2000
4. Hebert, P. A., Masson, M. H., Denoeux, T. Fuzzy multidimensional scaling. Computational Statistics and Data Analysis 51. pp 335-359. 2006.
5. Knipp, D., 2001, Knowledge surveys: What do students bring to and take from a class?: United States Air Force Academy Educator, Spring, 2001.
6. Nuhfer, E. (1993) "Bottom-Line Disclosure and Assessment," Teaching Professor, Vol. 7, n. 7, 8-16
7. Romero, C., Ventura,S., Garca, E. Data mining in course management systems: Moodle case study and tutorial Computers & Education, Volume 51, Issue 1, August 2008, Pages 368-384
8. Sanchez, L., Couso, I., Casillas, J. Modeling vague data with genetic fuzzy systems under a combination of crisp and imprecise criteria. MCDM 2007. Honolulu, Hawaii, USA, (2007).
9. Wirth, K. and Perkins, D. (2005) Knowledge Surveys: The ultimate course design and assessment tool for faculty and students. Proceedings: Innovations in the Scholarship of Teaching and Learning Conference, St. Olaf College/Carleton College, April 1-3, 2005, 19p
10. Nagel, L., Kotz, T. (2009) Supersizing e-learning: What a CoI survey reveals about teaching presence in a large online class. The Internet and Higher Education.
11. Zeki Saka, A., Hitting two birds with a stone: Assessment of an effective approach in science teaching and improving professional skills of student teachers. Social and Behavioral Sciences, Volume 1, Issue 1, 2009, Pages 1533-1544