Graphical exploratory analysis of vague data in the early diagnosis of dyslexia

Luciano Sánchez
Univ. Oviedo
Dpt. Informática
33071 Gijón, Asturias
luciano@uniovi.es

Ana Palacios
Univ. Oviedo
Dpt. Informática
33071 Gijón, Asturias
anapalacios@gmail.com

M. R. Suárez
Univ. Oviedo
Dpt. Informática
33071 Gijón, Asturias
mrsuarez@uniovi.es

Inés Couso
Univ. Oviedo
Dpt. Estadística
33071 Oviedo, Asturias
couso@uniovi.es

Abstract

We are in the initial stages of the design of a fuzzy rule-based testing system that can be used by unqualified personnel while screening the children for dyslexia. The main novelty of our work is the exploitation of low quality data (incomplete items, intervals, lists, subjective values, etc.) A sample of infants has been obtained, and some tests have been applied to them. In addition, a psychologist has examined and diagnosed each child. We want to relate the responses to the tests with the expert judgement of the professional, and highlight those factors that are involved in the early development of the dyslexia during the preschool age. However, this data being imprecise, we lack tools to assess the quality of each test and its influence in the diagnosis. We have used different graphical visualization techniques, to detect the most useful sets of factors, but have found that none of the approaches that we are aware of is able to show all the relevant information in the sample. Therefore, we propose a new Multidimensional Scaling algorithm, that gains a better insight into the spatial properties of the data and also into the amount of vagueness in the pieces of information comprising it.

1 Introduction

Dyslexia is a learning disability in people with normal intellectual coefficient, and without further physical or psychological problems that can explain such disability. It has been estimated that between 4% and 5% of schoolchildren have dislexia, with reading and writing problems [1]. The average number of children in a Spanish classroom is 25, therefore most of them have dyslexic children. Dyslexia may become apparent in early childhood, with difficulty putting together sentences and a family history. Recognition of the problem is very important in order to give the infant an appropriate teaching.

Using Soft Computing techniques for diagnosing dyslexia seems to us a natural choice, because of the properties of our data (linguistic terms, and vague measurements). As a matter of fact, there are many references where fuzzy techniques were used to learn medical diagnosis models from data (see, for instance, [18][14][2][12]). In particular, in [5] and [10], fuzzy techniques have been used in the diagnosis of disabilities in language, and in [17] some different dyslexia related tests were assessed with soft computing-based methods. However, in all of the preceding works, the data was crisp or categorical. Instead, most of our measurements are not crisp. Some of our responses are linguistic ("low", "high"), others are subjective (for example, the "squareness" of a hand-drawn shape) or interval valued (f.e. a dyslexia degree "between 2 and 4"). Lastly, a high percentage of cases have missing values. None of the preceding ap-

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1417–1424
Torremolinos (Málaga), June 22–27, 2008

proaches are directly applicable to the problem at hand.

In previous works [16], we have proposed some learning algorithms that input a vague training set and produce a fuzzy rule-based regression model, as we will explain in Section 2. Nevertheless, before we can apply these new algorithms, we must carry an exploratory analysis for determining a small subset of inputs (we use 413 different tests) that is informative enough to reproduce the criteria of the psychologist, or else the system will not have practical use.

The problem of feature selection in regression models with linear dependences between the variables is customary solved with Principal Component Analysis (PCA) or Factor Analysis. Both techniques have been generalized to certain types of fuzzy data, see for instance [6][13]. In addition, nonlinear dependences between vague data have also been studied. On the one hand, certain fuzzy feature selection algorithms that were designed for fuzzy classifiers can also be applied to regression problems with fuzzy data [15]. Moreover, modern approaches like Independent Component Analysis (ICA) and Self Organized Maps (SOM) have fuzzy extensions, but they are intended to improve the robustness when working with crisp data [8][3]. Other nonlinear extensions of PCA, like Curvilinear Component Analysis (CCA) [11] have not yet been extended to the fuzzy case. However, these advanced nonlinear techniques are closely related to a technique widely used in psychology, Multi-Dimensional Scaling (MDS) [9], that has been recently generalized to the fuzzy case [7]. We will use and extend this last technique.

The structure of this work is as follows: In Section 2 we will review the tests used in this research, measuring verbal understanding, logic reasoning, memory, and sensory-motor ability. In Section 3 we introduce the new graphical model. In Section 4 the results of applying the new model to our data are shown. Section 5 concludes the paper.

2 Symptoms and detection of dyslexia

The dyslexia is a disability that is diagnosed with the help of some tests, by a psychologist or specialist dyslexia teacher. In many Spanish schools these tests are routinely applied to children. However, attendance to school is not mandatory under the age of 6, thus dyslexic children might not be examined early enough, unless their parents suspect a problem. We intend to provide the parents with an automated tool that can screen for certain symptoms, suggesting that a professional is contacted for further diagnosis, if needed. We want to find those children possibly affected by dyslexia and also those that, without having dyslexia, have learning problems or are prone to have them in the future.

According to [1], the characteristic signs of the dyslexia depend on the age of the child. We are mostly interested in preschool education (between ages 4 and 6). These children are being initiated in reading and writing, but they can not properly read yet. Therefore, we can detect a tendency to the dyslexia, but the symptoms of the disability are not self-evident and the tests are designed to detect them. The most significant symptoms have to see with a slow acquisition of language skills (failing to remember lists of names, numbers, alphabet, days of the week, shapes, colours, etc.), mismatching words with similar pronunciation, limited vocabulary, attention deficit, hyperactivity and natural ability with technical toys, i.e., greater manual than linguistic ability, which typically will show in I.Q. tests.

From ages from 6 to 9 the symptoms begin to be conspicuous. Reading, writing and calculus skills are not acquired at the proper rate. In this case, the tests have to detect the symptoms, as before, but they are also intended to separate those children which actually suffer from dyslexia from those others for whom the problem can be related to other causes.

All the tests and evaluation criteria that have been used in this research are currently being used in Spanish schools for detecting dyslexia. In Figure 1 we have included an example of

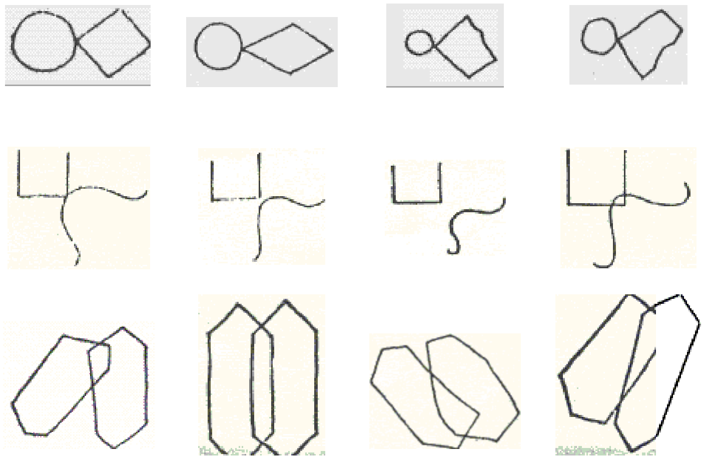

Figure 1: Example of some of Bender's tests for detecting dyslexia. Upper part: The angles of the shape in the right are qualified by a list of adjectives that can contain the words "right," "incoherent," "acceptable," "regular" and "extra." Middle and lower part: The relative position between the figures can be "right and separated," "right and touching," "intersecting", etc.

one of the tasks that the children have to solve in these tests: copying some geometric drawings. When a child is being evaluated, the expert has to decide whether the angles, relative position and other geometrical properties have been accurately copied or not, choosing between a given set of adjectives. Other tests produce numbers, or linguistic labels as "low", or "very high". Lastly, we allow the use to express indifference between different responses by means of intervals, as in "lower than 3" or "between 2 and 4". There are 13 categories of tests, that expand to a total of 413 numerical, categorical and interval-valued variables.

In this research we have selected a sample of 65 infants between 5 and 8 years old, in urban schools of Asturias (Spain), and collected their responses to the tests mentioned before. Afterwards, the same children were examined by a psychologist, who assigned each one of them a subjective score, which is an interval of values between 0 (normal child) and 4 (high degree of dyslexia). We remark that we do not intend to design a classifier, but an interval-valued regression model, as we can not assume that the output variable is categorical neither crisp, i.e., certain children are not assigned numbers by the psychologist, but ranges of values. Consequently, the desired output of our model is an interval of numbers, between 0 and 4, describing the degree of dyslexia. The width of the interval codifies the confidence in the prediction, and will be related to the vagueness of the input: the more vague the input is, the less specific the output of the model should be. The objective of this paper is to use soft computing techniques to relate these measured variables with the expert judgement of the professional, and highlight those factors that are involved in the early development of the dyslexia during the preschool age.

3 Graphical exploratory statistics

In the problem at hand, the data is vague and there is also a high proportion of missing values. We need to detect whether the vagueness of each instance is too high - thus that instance should be removed from the training set. Measures of vagueness are affected by the scaling of the data. For instance, one of the tests is assigned a value between 0 and 40, while others produce binary values. If the data is not scaled, the imprecision in the first test will be much more noticeable than the imprecision in the second. In crisp data, we solve the scaling problem by applying PCA to the correlation matrix instead of the distance matrix, but there is not an standard procedure when the data is imprecise.

In this respect, PCA finds projections of maximal variability, because the first k principal components span a subspace containing the best k-dimensional view of the data. Therefore, PCA minimizes the sum of the squared distances between the points and their projections. Multidimensional scaling (MDS) [9] generalizes this property, as it projects the instances in a low dimensional Euclidean space so that their proximity reflects the similarity of their variables. Fuzzy MDS, as described in [4][7], in turn generalizes MDS to the case where the distance matrix comprises intervals or fuzzy numbers. In this sense, the MDS map coordinates are the best approximation we have to the PCA when the data is vague and there are missing values.

3.1 A new fuzzy-MDS algorithm

MDS also depends on the scaling of the data. But, having into account that our ultimate objective is to obtain a fuzzy rule based-model, this problem can be circumvented. The main innovation in our algorithm is related to this problem: we will map the *activation space* instead of the input space. That is to say, we do not consider that two examples are similar when their coordinates are near in the space. Instead, we will consider that two examples are similar when they fire the same rules of the knowledge base.

Therefore, we propose to modify the Fuzzy MDS algorithm in [7] and use a new, non-Euclidean distance measure. This distance takes into account not only the vagueness, but also the granularity of the linguistic discretization used in the fuzzy knowledge base, and in this sense does not depend on the scaling of the data.

3.1.1 Representation of an instance in the activation space

We will assume, as most researchers in this field do, that those fuzzy sets defining the meaning of each linguistic variable form a Ruspini's partition. This means that the memberships of any crisp value to the elements of such a partition are conditional probability distributions. In other words, every precise observation of a numerical variable can be matched with a precise probability distribution over its corresponding universe of linguistic labels. For example, if a linguistic variable has three terms "L","M" and "H", then we will map every numerical value x_0 to a triplet of membership values. This triplet can be named $(\mathrm{p}(\mathrm{L}|x_0), \mathrm{p}(\mathrm{M}|x_0), \mathrm{p}(\mathrm{H}|x_0))$ and $\mathrm{p}(\mathrm{L}|x)+\mathrm{p}(\mathrm{M}|x)+\mathrm{p}(\mathrm{H}|x) = 1$ for all x.

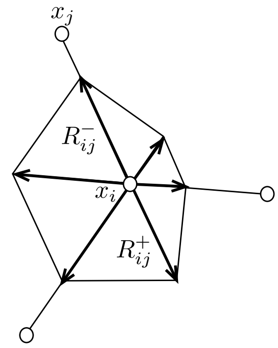

Figure 2: The projected data are polygons defined by the distances R_{ij} in the directions that pairwise join the examples.

We propose that interval-valued observations are mapped to sets of probabilities. For example, an interval $[x_1, x_2]$ will be mapped to a set $\{(\mathrm{p}(\mathrm{L}|x), \mathrm{p}(\mathrm{M}|x), \mathrm{p}(\mathrm{H}|x)) \mid x \in [x_1, x_2]\}$, which we will enclose in the imprecise probability distribution given by the triplet of upper probabilities $(\mathrm{p}^*(\mathrm{L}|x), \mathrm{p}^*(\mathrm{M}|x), \mathrm{p}^*(\mathrm{H}|x))=$ $(\max_{[x_1,x_2]} \mathrm{p}(\mathrm{L}|x), \max_{[x_1,x_2]} \mathrm{p}(\mathrm{M}|x), \max_{[x_1,x_2]} \mathrm{p}(\mathrm{H}|x))$, with $\mathrm{p}^*(\mathrm{L}|x)+\mathrm{p}^*(\mathrm{M}|x)+\mathrm{p}^*(\mathrm{H}|x) \geq 1$ for all x. Finally, we propose too that a fuzzy set with normal membership function $X(x)$ is represented by the tuple of upper probabilities $(\max_x X(x) \cdot \mathrm{p}(\mathrm{L}|x), \max_x X(x) \cdot \mathrm{p}(\mathrm{M}|x), \max_x X(x) \cdot \mathrm{p}(\mathrm{H}|x))$. Notice that our representation produces an interval-valued distance either with interval valued or fuzzy data, because it depends on the extrema of a set of distances that are induced by a set of probability distributions in both cases.

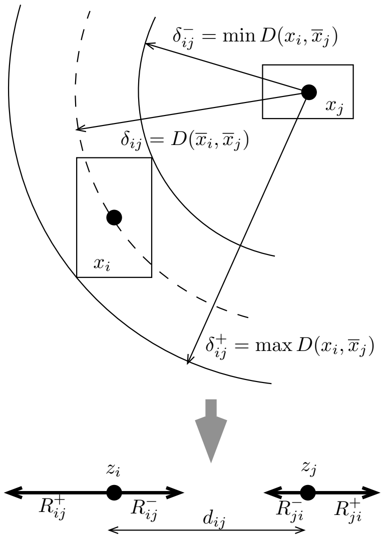

Figure 3: The distance between the projections of x_i and x_j is between $d_{ij} - R_{ij}^- - R_{ji}^-$ and $d_{ij} + R_{ij}^+ - R_{ji}^+$

3.1.2 Distance between cases

According to our representation, we define the distance between two imprecisely measured multivariate values $x_i = (x_{i1}, \ldots, x_{if})$ and $x_j = (x_{j1}, \ldots, x_{jf})$, with f features each, where the value of the feature x_{ik} is represented by the imprecise probabilities $(p^*_{ik1}, \ldots, p^*_{ikn})$, as the set of distances

$$\begin{gathered} D_{ij} = \{ \sqrt{\textstyle\sum_{k=1}^{f} \sum_{l=1}^{n} (p_{ikl} - p_{jkl})^2} : \\ \textstyle\sum_l p_{ikl} = \sum_l p_{jkl} = 1 \text{ for all } k \\ \text{and } p_{ikl} \le p^*_{ikl}, p_{jkl} \le p^*_{jkl} \text{ for all } k, l \}. \end{gathered} \tag{1}$$

The distance between two imprecise objects (either interval or fuzzy) is an interval, that can be easily obtained by quadratic programming or approximated by Montecarlo simulation.

3.1.3 Stress function

In [7] the projection of an imprecise case is a circle. We have found that, in our problem, this is a too restrictive hypothesis. Instead, we propose to approximate the shape of the projections by a polygon (see Figure 2) whose radii R_{ij}^+ and R_{ij}^- are not free variables, but depend on the distances between the cases.

Let $\{x_1, \ldots, x_N\}$ be a set of multivariate imprecise data, let $\overline{x}_i$ be the crisp centerpoint of the imprecise value x_i 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 (see Figure 3 for a graphical explanation):

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

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 $\delta_{ij}^- = \min\{D(x_i, \overline{x}_j)\}$.

Consequently, we propose that the value of the stress function 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.

4 Experiments

In [4, 7] the Haussdorff distance was not used, but the average of the differences between the extrema of the intervals, thus the stress function can be optimized by quadratic programming (QP). However, the stress function is not convex (notice, for example, that it is invariant under translations and rotations) and QP is not guaranteed to find a good solution. Our stress function function contains the "max" operator and it is not differentiable. Instead of using QP, the experiments in this section were obtained with a real-coded genetic algorithm, where each chromosome contains all the coordinates z_i of a map. Tournament selection, and a mix of uniform and arithmetic crossover were used. The population size was 100 and the number of generations 1000.

In Figure 4, we have applied some standard exploratory statistics to the data. Each interval value has been replaced by its centerpoint, and categorical data was represented by numbers. Principal Component Analysis, Independent Component Analysis and Multidimensional Scaling were evaluated. For scaling the data, the MDS algorithm was applied to the correlation matrix (therefore the map is different than that of PCA). The numbers are

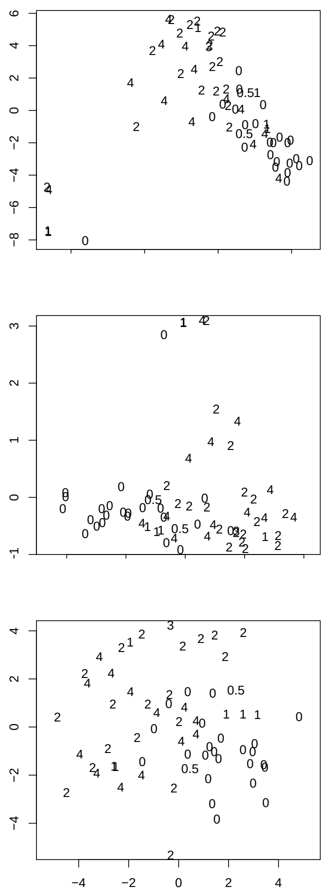

Figure 4: From top to bottom: PCA, ICA and MDS analysis of the data. Each interval has been replaced by its centerpoint. The numbers are the mean degree of dyslexia of the children, from 0 (no dyslexia) to 4.

the mean degree of dyslexia of the children, from 0 (no dyslexia) to 4. In the three diagrams, there is an area where normal children are concentrated, but there is not a clear dependence between the position of a point and the dyslexia degree that it represents. It is also emphasized that these maps do not consider the vagueness of the data.

In Figure 5 we have applied both the algorithm in [4, 7] and the proposal in this paper to our data. The first algorithm represents the data as circles. Since the scores of the tests have different ranges, this map does not show too useful information. By contrast, the MDS that we propose here is based on distances in the activation space (the memberships of the data to the linguistic partitions). Observe that it gains much more insight into the structure of this data.

In Figure 6, three subsets of 8 features have been examined. One of them has been obtained by means of a feature selection algorithm based on our own definition of fuzzy mutual information [15]. A second set has been obtained with classical techniques (analysis of variance), replacing the missing data and the imprecise measurements by crisp numbers. In the last place, an expert has chosen 8 variables that she considered relevant.

The second map, which is based in classical techniques, did not take into account the missing values neither the imprecision. As a consequence of this, highly imprecise variables were not avoided and most cases overlap in the map. Unexpectedly, the set of variables chosen by the expert has a good separability, but it does not produce a clear distinction between children with and without dyslexia. Only the first map, using fuzzy Mutual Information, has produced reasonable results.

Lastly, in Figure 7, we have studied the influence of the number of labels in the linguistic variables in the separability of this last case. Granularities 3, 5 and 7 were compared over that feature set obtained by MI. The set of granularity 5 seems to be a sensible choice, as the projection of the set of granularity 7 does not significantly improve the separability of the data, and 3 labels are not enough for separating the cases.

5 Concluding remarks

When tackling real-world problems involving imprecise data, we lack many standard procedures in the work cycle “exploratory analysis - preprocessing - learning -validation.” In this work, we have adapted an exploratory technique to the design of fuzzy rule-based systems. A new measure of distance, that allows us to analyze the effect of different granularities, was introduced, along with a more flexible representation based on polygons.

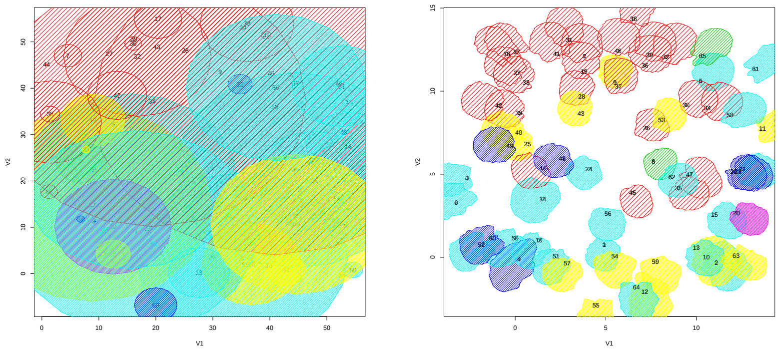

Figure 5: Left: MDS in input space. Right: MDS in activation space. The use of the activation space instead of the input space gains more insight into the structure of the data, and does not require scaling. The meaning of the colors is: Red: 0. Green: lower than 1. Blue: 1. Cyan: 2. Magenta: 3. Yellow: 4

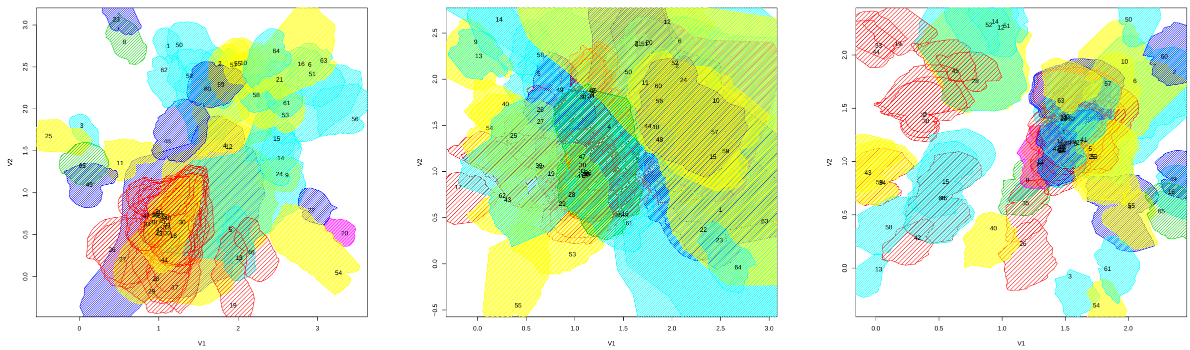

Figure 6: Three subsets of 8 variables have been examined. Left: Fuzzy Mutual Information [15]. Center: Analysis of Variance of linear models, with crisp data. Right: Selection of the most relevant tests, according to the human expert. The crisp selection did not take into account the vagueness. The set of variables chosen by the expert does not produce a clear distinction between children with and without dyslexia. The use of fuzzy IM produced reasonable results.

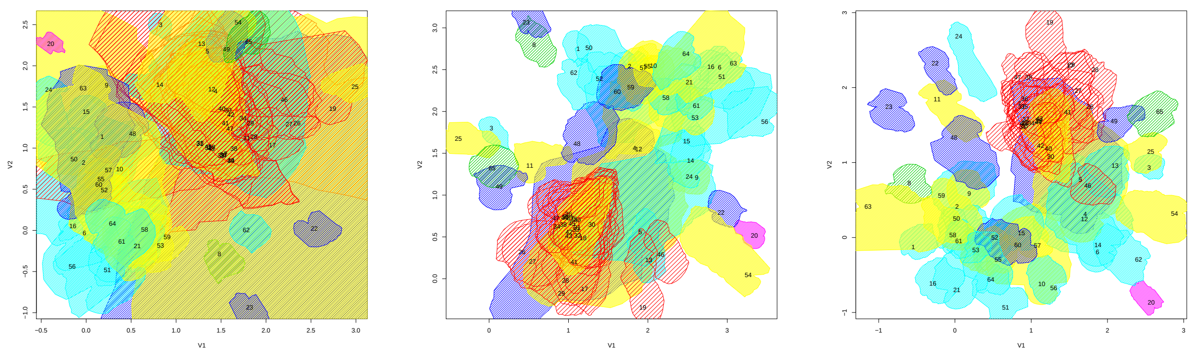

Figure 7: Interval MDS in activation space, with granularities 3, 5 and 7, and the feature set obtained by IM. The projection of the set of granularity 7 does not significantly improve the separability of the data and 3 labels are not enough for separating the cases, thus the best granularity is 5.

The next problem we have to solve is the selection of the most relevant variables. We have shown that the classical techniques, that discard the imprecision, are not adequate for this problem. In future works, we intend to extend the fuzzy MI based algorithm that we have mentioned in the paper, and derive new wrapper-type feature selection techniques for regression models with imprecise data.

Acknowledgements

This work was funded by Spanish M. of Science and Technology and by FEDER funds, within the grant TIN2005-08036-C05-05, and by Plan de Ciencia, Tecnologia e Innovacion 2006-2009 (PCTI) of Asturian regional Governement.

References

[1] Ajuriaguerra, J. Manual de psiquiatría infantil (in Spanish). Toray-Masson. 1976

[2] Akbarzadeh, M. R., Moshtagh, M. A hierarchical fuzzy rule-based approach to aphasia diagnosis. Journal of Biomedical Informatics 40, pp 465-475. 2007

[3] Bezdek, J. C., Tsao, E. C. K. and Pal, N. R. Fuzzy Kohonen clustering networks. In Proc. IEEE Int. Conf. on Fuzzy Systems, pp- 1035-1043, 1992

[4] Denoeux, T., Masson, M.-H., Multidimensional scaling of interval-valued dissimilarity data. Pattern Recognition Lett. 21, 83-92. 2000

[5] Georgopulos, V. A fuzzy cognitive map to differential diagnosis of specific language impairment. Artificial intelligence in Medicine 29. pp 261-278. 2003

[6] Giordani, P., Kiers, H., Principal Component Analysis of symmetric fuzzy data. Computational Statistics and Data Analysis 45(3). pp 519-548. 2004.

[7] Hebert, P. A., Masson, M. H., Denoeux, T. Fuzzy multidimensional scaling. Computational Statistics and Data Analysis 51. pp 335-359. 2006.

[8] Honda, K., Ichihashi, H. Fuzzy local independent component analysis with external criteria and its application to knowledge discovery in databases. International Journal of Approximate Reasoning 42(3), pp. 159-173. 2006.

[9] Kruskal, J.B., Nonmetric multidimensional scaling: a numerical method. Psychometrika 29, 115-129. 1964.

[10] Lakov, D. V. Soft Computing Agent Approach to Remote Learning of Disables. 2nd. IEEE Intl. Conf. Intelligent Systems. pp 250-255.2004

[11] Lee, J. A., Lendasse, A., Verleysen, M., Curvilinear Distance Analysis versus Isomap. European Symposium on Artificial Neural Networks ESANN'2002. pp 185-192. 2002.

[12] Lurie, A., Marsala, C., Hartley, S., Bouchon-Meunier, B., Dusser, D. Patients' perception of asthma severity. Respiratory Medicine, pp 2145-2152. 2007

[13] Nakamori, Y., Watada, J. Factor analysis for fuzzy data. Sixth Intl. Conf. Fuzzy Systems, pp. 1115-1120. 1997.

[14] Di Nuovo, A., Catania, V., Di Nuovo, S., Buono, S., Psychology with soft computing: An integrated approach and its applications. Applied Soft Computing 8, pp 829-837, 2007

[15] Sanchez, L., Suarez, M. R., Villar, J. R., Couso, I. Some results about mutual information based feature selection and fuzzy discretization of vague data. Intl. Conf. Fuzzy Systems FUZZ-IEEE 2007, pp. 1-6. 2007.

[16] Sanchez, L., Otero, J., Villar, J. R. Learning fuzzy linguistic models from low quality data by genetic algorithms. Intl. Conf. Fuzzy Systems FUZZ-IEEE 2007, London. 2007.

[17] da Silva, F., Munzlinger, E., Jannette, L., Riveros, M. Validacao de um sistema especialista para o pre-diagnostico da dislexia (in Portuguese). XXV Congresso da Sociedade Brasileira de Computacao (CSBC). 1206-1209. 2005

[18] Stylios, C., Georgopoulos, V. C., Fuzzy Cognitive Structure for Medical Decision Support Systems. Forging New Frontiers: Fuzzy Pioneers II. 218, pp 151-174. 2007