The Role of KDD Support-Interpretation Tools in the Conceptualization of Medical Profiles: An Application to Neurorehabilitation

Karina Gibert1,2, Alejandro García-Rudolph2, Gustavo Rodríguez-Silva1
Department of Statistics and Operation Research, Universitat Politècnica de Catalunya, Barcelona, Spain1
Knowledge Engineering and Machine Learning group, Universitat Politècnica de Catalunya, Barcelona, Spain2
Institut Guttmann-Hospital de Neuorrehabilitació, Badalona, Spain1

SUMMARY
Purpose: The work presents the usefulness of Traffic lights panel for assisting the interpretation of clustering results and includes an application to a real case of discovering response patterns to a rehabilitation treatment for brain damage patients. Work Method: A KDD framework where first, descriptive statistics of every variable was done, data cleaning and selection of relevant variables. Then data was mined using Exogen based on rules (CIBR), a hybrid AI and Statistics technique which combines inductive learning (AI) and clustering (Statistics). A prior Knowledge Base (KB) is considered to properly bias the clustering; semantic constraints implied by the KB hold in final clusters, guaranteeing interpretability of the results. Class panel graph, previously used for interpretation is abstracted and transformed to a Traffic lights panel to assist the expert in a final process of conceptualizing the obtained classes. Work Results: A set of 4 classes was recommended by the system and interpretation permitted profiles labeling. From the medical point of view, composition of classes is well corresponding with different patterns of increasing level of response to rehabilitation treatments. The Traffic lights panel is confirmed as a very useful tool to approach the results of the clustering to the expert, making the final interpretation easier Discussion: All the patients initially assessable conform a single group. Severe impaired patients are subdivided in three profiles which clearly conform a single group. Severe impaired patients are subdivided in three profiles which clearly distinct response patterns. Particularly interesting the partial response profile, where patient could not improve executive functions. Traffic lights panel is clearly representing the profiles, so the expert can very quickly label them. Conclusions: Meaningful classes were obtained and, from a semantics point of view, the results were sensibly improved regarding classical clustering, according to our opinion that hybrid AI&Stats techniques are more powerful for KDD than pure ones. Interpreting the results upon the Traffic lights panel is much easier than presenting the CPG directly to the expert
Keywords: decision support and knowledge management, rehabilitation, clinical test, TBI, knowledge discovery, interpretation-oriented tools, class panel graphs, traffic lights panel, exogenous clustering based on rules, knowledge-based applications in medicine.

1. INTRODUCTION
Neuropsychological rehabilitation seeks to reduce cognitive disability after acquired brain injury. However, there is not enough data yet to allow neuropsychological rehabilitation based on scientific evidence type I. Although there is a considerable amount of comparative studies aimed to show strongest efficacy of rehabilitation versus other interventions, most of them remain inconclusive. It is accepted by the international scientific community the difficulties for using standard methods as in other therapies or clinical trials in this field, mainly because the studied populations are highly heterogeneous and there is an important lack of knowledge about the natural evolution of the process for different patients, as well as for the “active” components of the treatments to be controlled. Nowadays it is well known that Knowledge Discovery (KDD) approach provides a good framework to analyze complex phenomena as the one presented here for getting novel and valid knowledge that can improve the background corpus doctrinae
[1-2]. In 1989, the first Workshop on Knowledge Discovery from Data (KDD) was held. Seven years later, in the proceedings of the first International Conference on KDD, Fayyad gave one of the most well known definitions of what is termed Knowledge Discovery from Data:
“The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (16)
KDD quickly gained strength as an interdisciplinary research field where a combination of advanced techniques from Statistics, Artificial Intelligence, Information Systems, Visualization and new algorithms are used to face the knowledge acquisition from huge data bases. The term Knowledge Discovery from Data appeared in 1989 referring to high level applications which include particular methods of Data Mining. Fayyad’s proposal marked the beginning of a new paradigm in KDD research:
“Most previous work on KDD has focussed on [...] DM step. However, the other steps are of considerable importance for the successful application of KDD in practice”.
The approach included prior and posterior analysis tasks as well as the application of DM algorithms. These may in fact require great effort when dealing with real applications. Data cleaning, variable selection, transformation, selection of DM techniques and optimization of parameters (if required) are critical for a proper analysis and are often time consuming and difficult, mainly because the processes should be tailored to each specific application, and human interaction is required. Interpretation of results
is also often time consuming and requires much human guidance [17].

In this work we are focussing on the process of post-processing the results of a clustering method as an integral part of a whole KDD process. Traditionally, the clustering results are expressed as a partition of the set of elements to be clustered (patients in this case). So several groups of patients are listed as final result. After that, the analyst is the responsible helping the expert to identify the particularities of every group and to discover the underlying clustering criteria that permits a semanantics interpretation of the results, obtaining effective valid and useful knowledge to increase the corpus doctrinae. Very often, basic statistics per group and some basic inference tests between groups are performed manualy to understand the nature of the discovered profiles. When either the number of discovered classes and/or the number of variables to be considered is big, this is a very complex process and it becomes difficult and tedious to analyze the results, obtaining effective valid and useful knowledge to increase the corpus doctrinae. Very often, basic statistics per group and some basic inference tests

process is the following:

1. Perform the Class panel graph (CPG) of all the variables of the considered assessment scales. On the other hand, measurements were available before and after the treatment and the differences between pre and post treatment variables where calculated. It was decided to use measurements after the treatment and differences to perform an Exogenous Clustering based on rules (ECIBR, [18]) analysis. Measures before the treatment and other descriptive information (patient's age, gender, severity of lesion (Glasgow scale), length of post-traumatic amnesy (in days) and evolution time of the lesion (in months) will only be used to enrich final interpretation.

Several previous experiences from real applications [9, 3, 10-13] show that using CIBR-like [5] methods use to be better than using any statistical clustering method by itself, owing to the property of those methods based on holding, in the final clusters, the semantic constraints implied by a KB which expresses the prior expert knowledge; what guarantees interpretability of the resulting classes. Also, it uses to be better than pure inductive learning methods, since it reduces the effects of missing some implicit knowledge in the KB. In [14] this was confirmed again for a previous sample of the present research and it was seen that classical clustering performed badly with this kind of data. Thus, in this work, classical hierarchical clustering was disregarded and only the EclBR (see [4-5] for ClBR details and [14] for ECIBR) was performed using some prior knowledge from the expert (see details in [14]), the Ward's criteria and Gibert's mixed metrics [6]. Runnings were performed using the software KLASS v7. [15]. The result is a partition of the data that has to be interpreted. The use of Traffic lights panels is introduced for better supporting the process of class conceptualization.

Interpretation-oriented tools: Originally, we approached the interpretation of the classes on the basis of basic statistics per class, Class panel graph (CPG) [6], which is presented bellow, and significance tests assessing relevance of differences among classes (ANOVA, Kruskall-Wallis or chi-2 independence test, depending on the item). The aim is to extract qualitative information from the CPG, to obtain a meaningful description of classes by identifying which variables indicate particularities of every class regarding the others. In this work, the Traffic light panel has been introduced as an abstraction of the CPG providing information closer to the expert knowledge and making even easier the process of recognizing the concepts represented by the profiles. Thus, the final interpretation process is the following:

1. Perform the Class panel graph (CPG) of all the variables versus the discovered classes
2. Calculate the basic statistics per class
3. Using 1.) and 2.) identify variables or combination of variables with specific ranges of values in a class, that distinguish the class from the others.
4. Assign qualitative levels to the variables implied in step 3.
5. Perform a Traffic lights panel for the variables,
using the qualitative values assigned in 4.)

6. Show the Traffic lights panel to the expert and ask him to select a label for the class. The expert is conceptualizing the class in this step, on the basis of the Traffic lights panel.

7. Perform significance tests assessing relevance of differences for the variables implied in step 7) (ANOVA, Kruskall-Wallis or chi-2 independence test, depending on the item).

Class panel graph: It is a graphical representation in form of panel (see figure 1), containing variables in the columns and classes in the rows. Conditional distributions of the variables through the classes is shown per column. They can be represented either via multiple histograms or via multiple boxplots [6]. It shows in a single panel the information of a number of variables though the classes. In a very compact way it provides perspective of the variables behaving particularly for a certain class and supports the interpretation process.

Traffic lights panel: It was introduced after realizing that the histograms or boxplots represented in the class panel graph could be perceived as too technical by some non-technical experts as environmental experts or medical experts, even if they appeared conveniently marked by the analyst with the relevant values of the variables. The Traffic lights panel displays the qualitative dominant levels of a set of variables through a set of classes. It corresponds, in fact, to an abstraction of the information provided by the class panel graph, providing a more symbolic representation, which is closer to the interpretation codes of a non-technical expert, like a physician. The analyst have to read the class panel graph, mark the characteristic values of the variables for the different classes and assign qualitative levels to those values.

The dominant level of a variable in a given class can be determined in two ways:

a) Identify the qualitative level of the mean or median of the variable in the class

b) Identify the qualitative level of the mode of the variable in the class

When the variables have only 3 qualitative levels assigned, it is very interesting to assign the colors of a traffic light to those levels (red for the bad or negative value, yellow for the medium or neutral value and green for the good or positive value), being the bad values the higher or lower values of the variable depending on the variable's semantics. This is explaining the name of the graph. If more than 3 qualitative levels should be considered, intermediate colors should be included. But, it is important to keep red and green as non-verbal codes for bad and good values to connect with the expert's implicit codes for interpreting. An important property of the Traffic lights panel is that, as the classes are well-constructed, they must be distinguishable, and they must represent different profiles. Thus, there should not be two rows of the graph with the same combination of colors.

Experimental procedure: The target sample includes 71 patients with TBI between 16 and 50 years, receiving neurorehabilitation treatment at the Institut Guttmann-Hospital de Neurorehabilitació in the year 2007. Patients with language area affected were excluded from the sample, since in [14] it was confirmed that they follow a treatment only based in logopedics and they conform a very particular profile, already known by doctors. All patients were administered the neuropsychological assessment at admission. Same evaluation was also performed at the end of the rehabilitation. Differences between pre- and post-treatment test scores were used to measure particular patient's improvements in the domains of attention, memory and executive functions. After the initial evaluation all the patients initiated a two months rehabilitation program with a personalized intervention, where patients worked in each one of the specific cognitive domains, considering the degree of the deficit and the residual functional capacity.

Tests Instruments: Neuropsychological assessment was based on the scales ordinary measured at the Institut Guttmann and covered the major cognitive domains [7]:

• Measures of attention included Digit Span Forward, Trail Making Test-A, and Stroop Test (word-colour condition).

• Memory and Learning was assessed with Digit Span Backward, Rey verbal auditory learning tests (RAVLT) and Letter-Number sequencing (LyN)

• Executive functions was assessed with the Wisconsin Card Sorting Test, Trail Making Test-B, Stroop Test (interference condition), and Fluency verbal test (PRM).

3. RESULTS

Forty-seven persons with TBI (60 men, 11 women) participated in the study. The mean age at time of injury was 29.225 years (SD=12 years). Their initial clinical severity of TBI according to Glasgow Coma Scale (GCS) was estimated as severe (GCS between 3 and 8) in 70.3 % of the patients and moderate (GCS between 9 and 12) in 29.6% cases. A final set of 12 variables was considered to be relevant for the target problem: 3 from attention scales, 4 from memory and learning scales and 5 from executive functions scales. As said before, the measures after the treatment and the differences respect the initial evaluation were considered for clustering. Using ECIBR, 4 classes with different kind of responses to treatment were found.

Figure 1 shows the CPG of relevant variables measured before (PRE) and after (POST) the treatment and also the differences between both scores (DIFE). The column Classe indicates the class and the column nc indicates the number of patients classified in that class. The basic statistics per class were also calculated for all the variables. The analyst was then marking the characteristic values of the variables in the different classes and assigned qualitative levels and colors to every value. Instead of presenting the marked CPG to the expert, the Traffic lights panel was build. Figure 3 shows in a graphical way how the Traffic lights panel
The Role of KDD Support-Interpretation Tools in the Conceptualization of Medical Profiles: An Application to Neurorehabilitation

The experts’ point of view, looking at different variables [9]. The conceptualization process performed by the experts upon the Traffic lights panel allowed to assign conceptual lables to every class and reconfirmed the profiles found in [14] with a previous samples with a slightly different set of tests (see [14] for details on description: a single group of people already able to perform the tests before the treatment, which improve after treatment Assessable (C49), and three groups of people with severe impairment before the treatment and different responses: Global Improvement (C47), DysExecutive (C46, still problems with executive functions after the treatment), Resistant (40, still severe impairment after the treatment, they do not respond to the treatment).

4. DISCUSSION

Clustering techniques allow detecting different groups of patients using their neuropsychological evaluations. Facing such a complicated phenomenon as rehabilitation, concerned with a lack of clear patterns and difficulties for establishing relationships between exercises and patient improvements, requires indeed to take into account as much prior expert knowledge as possible, even if it is a partial description of the phenomenon. The additional knowledge provided by experts is expressed by means of logical rules and concerned some associations between some sets of tests. The ECIBR is a hybrid technique which can use this prior knowledge to guarantee interpretability of results [9] by integrating clinical knowledge inside the analysis. Finally, a set of 4 classes was recommended by the system, reconfirming the profiles found in [14], which from the medical point of view is well corresponding with different patterns of response to rehabilitation treatments. In this paper, particular focus to the pro-

Figure 1. Class Panel Graph for the whole set of tests

Figure 2. Complete Traffic lights panel for the whole set of tests considered in the analysis

Figure 3. Transformation of a Class panel graph in a Traffic lights panel (memory tests)

for the memory tests was constructed. For every variable it has been analyzed whether the high scores represented a good performance of the tests by the patient or a bad one and colors were assigned accordingly.

The same procedure was used for the whole set of variables and Figure 2. shows the complete Traffic lights panel presented to the expert. Upon those panel, a clear conceptual interpretation is possible from
cess of interpreting the results was made: First, Class panel graphs of the whole set of variables was build (figure 1) and the analyst transformed it into a Traffic lights panel (figure 2). Regarding the memory tests, for example, the Traffic lights panel tell the expert that at the beginning of the treatment, all the groups except c49 were so much affected that they could not perform at all any of the memory tests. At the end of the treatment c40, which is not responding to treatment at all, is still not performing any of the memory tests; all the other patients do. The group c46 shows the best improvement in memory functions (green color for differences in scores) and so on. Only looking at the Traffic lights panel, the information provided is quickly suggesting concepts to the expert and class labeling becomes much more easier than using the original class panel graphs. It is not so evident to extract all this information from the Class panel graph directly. From this experience and some other ones performed on the environmental sciences field, we can say that the introduction of the Traffics light panel approaches the results to the expert knowledge and makes the interpretation process more easy.

5. CONCLUSIONS AND FUTURE WORK
ECIBR was used on a sample of 71 patients to reconfirm the profiles already discovered in [14] for patients with no Language area impairments. ECIBR provided a framework for an easy expert prior knowledge acquisition which could properly be considered in the data mining step. Using this knowledge, even being a partial description of the domain, meaningful classes were obtained from a semantics point of view and the results were sensibly improved regarding classical clustering, according to our opinion that hybrid techniques that combine AI and Statistics are more powerful for KDD than pure ones. The use of CPG provided a good support for results interpretation. But introducing the Traffic lights panel approaches even more the clustering results to the expert background knowledge and makes the interpretation even easier. The main contribution of the Traffic lights panel is that the technical reading of the class panel graph, understanding which conditional distributions differ from one class to another and assessing it by statistical tests becomes clearly separated from the conceptualization process performed by the expert, which, upon the elaboration of the analysts over the class panel graph, is able to induce concepts from the classes structure, and can propose conceptual labelling for every profile.

In this work, the Traffic lights panel has been manually induced from the CPG and the basic statistics per group table, but a process to induce it at least semiautomatically is currently in progress. This work highlights the importance of correct postprocessing the results of a Data Mining process to properly obtain useful and understandable knowledge upon Fayyad’s recommendations for KDD [16]. The profiles identified in this work helped the experts to identify some clinical hypothesis to be tested in future analysis and currently the relationship between the type of response to the treatment and the concrete rehabilitation program followed by the patient is analyzed.

REFERENCES

Corresponding author: Karina Gilbert Oliveres, PhD. Statistics and Operation Research. Universitat Politècnica de Catalunya. Edifici C5 Campus Nord. C Jordi Girona, 1-3 08034 Barcelona Tel (+34) 934 017 323 Fax: (+34) 934 015 855 E-mail: karina.gilbert@upc.edu