CLASSIFICATION BASED ON RULES AND THYROIDS DYSFUNCTIONS

KARINA GIBERT1,*.† AND ZDENKO SONICKI2

1Department of Statistics and Operation Research, Universitat Politècnica de Catalunya, C. Pau Gargallo, 5, Barcelona, 08028 Spain
2Andrija Stampar School of Public Health, Medical School, University of Zagreb, Croatia

SUMMARY

Classification in ill-structured domains (ISD) is a difficult problem for the actual statistical and artificial intelligence techniques, because of the intrinsic characteristics of those domains. Classification based on rules is our proposal to overcome the limitations of Statistics and Artificial Intelligence techniques referred to in this particular context. In this paper, an application of the classification based on rules to a set of real data is presented. Data base is about thyroid function and data was provided by a hospital from Zagreb (Croatia) covering a period of two years. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS: clustering; knowledge base; thyroid dysfunction

1. INTRODUCTION

Common characteristics of ill-structured domains [4] are:

(i) Heterogeneous data matrices: The variables used to describe the objects may be either quantitative or qualitative. The qualitative variables used have a great number of modalities, according to the expertize of the user.
(ii) Additional knowledge of the domain structure is available. It usually is qualitative knowledge relative to the overall structure of the domain (relationships among variables, classification goals, etc).
(iii) Partial and non-homogeneous knowledge: Experts use to deal with great amounts of implicit knowledge. That is why only partial knowledge is available. On the other hand, this knowledge used to be non-homogeneous, i.e. with different degrees of specificity.

*Correspondence to: Karina Gibert, Department of Statistics and Operation Research, Universitat Politècnica de Catalunya, C. Pau Gargallo, 5, Barcelona, 08028 Spain.
†E-mail: karina@eio.upc.es
Although additional expert knowledge on the domain structure is always available, construction of complete knowledge bases — to be used in diagnostic-oriented systems — is almost unreachable for ill-structured domains due to the complexity of these kinds of domains.

On the other hand, statistical clustering of ill-structured domains (based on distances, which are, in fact, syntactic criteria) usually has a rather random behaviour. Actually, quantitative and qualitative variables coexist in ill-structured domains, but standard statistical techniques were not specifically designed, neither for simultaneous treatment of numerical and categorical variables nor for treatment of great quantities of qualitative information.

As said before, classification based on rules is our proposal for successfully classifying ill-structured domains. From a technical point of view, in Reference [3] it is shown that, the classification based on rules generally has a beneficial effect on the cost of the algorithm. KLASS is a clustering tool that implements this methodology. It can use semantic information (in the form of logic rules) to guide the classification process. One of its most important features is the use of both qualitative and quantitative variables in the object descriptions. The symbolic representation of the qualitative variables is maintained, and unnatural codifications of quantitative variables are also avoided. The definition of a class representative for qualitative variables is needed as well as a new measure to evaluate distances between individuals [1].

2. THE METHODOLOGY

Classification based on rules consists of a mixed classification strategy that:

1. Incorporates partial and/or non-homogeneous knowledge that the expert has on the domain. Formalizing it by means of logic rules provides maximum expressiveness and flexibility (complete knowledge is not necessary anymore). Build Knowledge Base including expert rules.
2. Expert rules are used to build an initial induced partition on the domain. This is equivalent to pack knowledge pieces in functional units, introducing semantics into the system.
3. A chained reciprocal neighbours classification (of quadratic cost) may be performed in each rules-induced class.
4. In a final step, an integration process, based on the conceptual description of the rules-induced classes, generates one unique hierarchy for all the individuals.
   A great advantage of this methodology is the possibility of using rules that involve variables defined as transformations of the observed ones. In this point, the mechanism is especially powerful, providing the possibility to study the data simultaneously in different co-ordinate systems.
5. Some automatic tools help the user to interpret the results and to suggest modifications on the Knowledge Base used to find the rules-induced classes. If so, repeat all the steps with the new Knowledge Base.

As can be seen, an iterative working methodology is proposed: starting with observational data and expert knowledge, the process described above is integrated with some other tools oriented to the interpretation of the classes. Using them, the user may realize the existence of rules, acquired by experience, that he unconsciously applies in spite of being unable to formalize them from the beginning. At the end, a satisfactory classification can be obtained, according to the expert goals. This methodology solves the expert difficulties when formalizing its knowledge in a complete and
precise way in ill-structured domains. In Reference [2] a formal description of this methodology can be found.

3. APPLICATION TO THE DIAGNOSIS OF THYROID FUNCTION

In this paper, the application of clustering based on rules to a real data base is presented. It is about medical environment.

Data base comprises results of the routine assays performed at Clinical Hospital Setre Milosrdnice, Zagreb (Croatia), and it has been collected for a period of two years.

A sample of 1002 patients was described by 12 laboratory tests and three factors, relevant to the outcome of diagnosis. Results presented in this paper were obtained using a subset of six variables:

1. total triiiodothyronine ($T^3$)
2. total thyroxine ($T^4$)
3. thyroid stimulating hormone (TSH)
4. gender (male, female)
5. age
6. drug therapy (thyrosuppresion, thyroid hormone, without therapy)

in order to allow future comparison with previous classifications performed by other methods [5] on the same sample (and the same subset of variables).

According to physical examination, experienced physicians decided on diagnosis of thyroid function state of each patient:

a. euthyreosis (842 cases),
b. hyperthyreosis (104 cases) or
c. hypothyreosis (56 cases).

Predictions are based on laboratory tests results, gender, age and information about possible drug therapy. Laboratory tests are described in detail in Reference [5].

The variable Diagnosis could act, in some sense, as the response variable. So, the important thing in this application is not discovery of classes, but characterization of these classes and the possibility to establishing a Knowledge Base to identify the diagnosis of a new case. This is one of the applications for which classification based on rules can perform as some kind of supervised learning method.

This is a set of data quite difficult to be analyzed. Among other reasons for that, there are a great number of missing data in all the variables. On the other hand, in the data set some patients have similar or same descriptions. One possibility is to avoid repeated descriptions, so reducing the size of the input data file. Management of weighted objects had to be included in the system.

Owing to the delay between physical examination and the obtention of laboratory test results, there are some pairs of patients with identical descriptions and different diagnoses. This fact has also to be taken into account in the generation of the final Knowledge Base. Fortunately, this occurs in a small percentage of cases.

Laboratory methods changed during the two years of the data collection phase. So, numerical measures are not always in the same scale for a given variable. Ranges of normal values changed
with the laboratory method, or even just with the change of a reactant. So, data are not directly comparable in their original numerical format. That is why those variables were encoded in three categories (low, normal and high), attending to medical criteria.

First of all, a hierarchical clustering was performed on the data, without taking advantage of the additional knowledge that the expert could provide. Using reciprocal neighbors algorithm a cut with 5 classes gives a 90 per cent predictive accuracy. Figure 1 (left) shows the dendrogramme generated in this case. In spite of having high predictive accuracy, it is not clear why individuals with euthyreosis are split into two classes; the same occurs with patients with hyperthyreosis.

Our objective in this application is, therefore, to improve the conceptual interpretation of the classes and their characterization, and not only to improve the predictive capacity, which is already high enough.

The expert provided some rules to improve initial results like:

\[
\text{If } T3 = \text{Normal} \land T4 = \text{Normal} \land TSH = \text{Normal} \rightarrow \text{Euthyreosis (not-ill patients)}
\]

\[
\text{If } T3 = \text{High} \land T4 = \text{High} \land TSH = \text{low} \rightarrow \text{Hyperthyreosis}
\]

Classification based on rules was used with this knowledge base on the sample of 1002 patients. Finally, a set of eight classes was obtained. The description of those classes is shown in Table I (frequency distribution of each variable in each class is shown). In this case, a clear conceptual interpretation of the classes is available.

To comment on some of these classes:

1. \( C_1 \), for example, can be interpreted as the class of the \textit{euthyreosis} people (not ill);
2. \( C_3 \) old patients without illness;
3. \( C_5 \) people with \textit{hyper} on thyrerostatic therapy;
4. \( C_6 \) people with \textit{hyper} on hormone therapy;
5. \( C_7 \) young women without illness or
6. \( C_8 \) as the class of patients with \textit{hypothyreosis}.

The final accuracy was increased to 92.5 per cent which means about 75 misclassified objects over 1002. Part of this misclassification cannot be avoided at present, since there are several identical cases with contradictory diagnostics in the sample, owing to the existent delay between
the moment when the physician decides the diagnostic and the arrival of some relevant tests results.

Figure 1 (left) shows the dendrogramme obtained by performing clustering without using rules, while Figure 1 (right) shows the one obtained including the expert rules into the clustering. As can be seen, the structure of the tree is rather different.

4. CONCLUSIONS

The domain of thyroid diagnostics is a good real example of an ill-structured domain.

For the presented application it can be seen that classification based on rules can improve the quality of the results, even when a small set of very simple rules is used. In this case, interpretability of classes improved on combining the knowledge base with the clustering process. From other applications, it has been also seen that the introduction of semantics information in the form of rules into the clustering process, generally produces clusters easy to be interpreted by the user. This methodology gives especially good performance in ISD.

This particular application shows how clustering based on rules can act as a supervised learning method when there exists a response variable to be used as a reference for the classification. Of course, this variable is not used for clustering the data, but to be compared, after clustering, with the results provided by the method.

At present, other sets of rules are being studied with this data. The next phase of this research, actually in progress, will include the definition of a base of decision rules using automatic techniques on the generated clusters.

Comparison with other techniques [5] is also being done at the moment.

The data show a strong structure from the beginning and improvement due to the Knowledge Base is only perceived with reference to the interpretability of the classes. As a future work, it is interesting to see if this strong structure can be found again by using the variables in its original
form (without grouping values of laboratory tests in high, normal and low levels). Part of this research, actually in progress, is oriented to study the effect of using the Knowledge Base to bias the clustering process when variables are taken in its natural form (numerical one for laboratory tests, categorical one for therapy and gender).

REFERENCES