Knowledge Discovery with Clustering Based on Rules. Interpreting Results

Karina Gibert[1], Tomàs Aluja[1] Ulises Cortés[2]

[1] Dep. of Statistics and Operation Research. {karina, aluja}@eio.upc.es
[2] Departament of Software. ia@lsi.upc.es
Universitat Politècnica de Catalunya***
Pau Gargallo, 5. Barcelona. 08028. SPAIN.

Abstract. It is clear that nowadays analysis of complex systems is an important handicap in Statistics, Artificial Intelligence, Information Systems, Data visualization, and other fields.
Describing the structure or obtaining knowledge of complex systems is known as a difficult task. The combination of Data Analysis techniques (including clustering) , Inductive Learning (knowledge-based systems), Management of Data Bases and Multidimensional Graphical Representation must produce benefits on this field.
Clustering based on rules (*CBR*) is a methodology developed with the aim of finding the structure of complex domains, which performs better than traditional clustering algorithms or knowledge based systems approaches. In our proposal, a combination of clustering and inductive learning is focussed to the problem of finding and interpreting special patterns (or concepts) from large data bases, in order to extract useful knowledge to represent real-world domains. This methodology and its behaviour as a *Knowledge Discovery* has been, in fact, presented in previous papers ([3], [5], [2]...).
The aim of this paper is to emphasize the *reporting* phase. Some tools oriented to the interpretation of the clusters are presented; automatic rules generation is presented and applied to a real research. Actually, in a *KD* system, data preparation and interpretation of the results is as important as the analysis itself. In this paper, missing data treatment is analysed; a statistical test, based on non parametric techniques, for comparing several classifications is presented. Also, a method for finding characteristic values of the classes is presented; this is based on the prototype of each class. Finally, these characterizations allow automatic generation of decision rules, as a predictive tool for future items.

Keywords: Combining many methods in one system, statistical tests in *KDD* applications, medicine: diagnosis and prognosis, from concept learning to concept discovery, Prior domain knowledge and use of discovered knowledge.

*** This research has been partially financed by the project *TIC'96-0878.*

1 Introduction

The formation of and distinguishing between different classes of objects (clustering) has been in use for very long. This process has been studied from the point of view of Statistics, AI, and other areas. The classes may be interpreted as diagnoses, predictions, etc. From a Machine Learning point of view, clustering is useful for the automated generation of classification rules, which is extremely interesting in knowledge-based environments, in particular the diagnosis oriented ones.

However, in real applications, it is usual to work with very complex domains [2], such as mental disorders, sea sponges[4]..., where data bases with both qualitative and quantitative variables appear; and expert(s) have some prior knowledge (usually partial) of the structure of the domain — which is hardly taken into account by clustering methods.

Clustering based on rules is a methodology developed in [4] with the aim of improving the process of finding the structure of an *ill-structured domain.* A combination of clustering and inductive learning is focussed to the problem of finding and interpreting special patterns (or concepts) from large data bases, in order to extract useful knowledge to represent real-world domains. Actually, *CBR* can be seen as a process of building a knowledge model for a given domain. That is why it is also connected with Knowledge Discovery of Data (*KDD*) and Data Mining (*DM*) [1]. In fact, we agree with the idea that a number of real applications in *KDD* either require a clustering process or can be reduced to it [8]. From this point of view, clustering techniques are an important pile for what is known as *KDD.*

CBR is a new way to perform classifications of any heterogeneous data matrix, as well as building a set of rules to describe the knowledge contained in the domain. Especially good results are obtained when analyzing data from complex domains. If we consider that Fayyad defines a *KDD* process as the "*overall process of finding and interpreting patterns from data, typically interactive and iterative, involving repeated application of specific data mining methods or algorithms and the interpretation of the patterns generated by these algorithms*"[1], it is clear that *clustering based on rules* fits this definition very closely. Following Fayyad, two important key points of *KDD* are: *i*) using domain knowledge and *ii*) domain characterization. It can be seen that these elements play an important role in the methodology presented here.

2 The methodology: Clustering based on rules

In this section, *clustering based on rules* (or rule-based clustering, *CBR*) is described. As most *KDD* systems, it combines, integrated in the same system, acquisition and management of prior knowledge from the expert with a Data Mining method (in this case, automatic clustering) [2] and some interpretation tools to analyze the results. It is an *iterative* and *interactive* process, structured in two major phases which finally organize the set of objects into a set of classes

that are presumed to be *semantically* interpretable: initially, there is a process of acquisition of the available background knowledge, followed by the clustering process *strictu sensu*.

On the one hand, this methodology helps the user to make explicit his prior knowledge relevant to the problem. The main idea is to allow the user to introduce *constraints*, which may be based on *semantic* arguments..., on the formation of classes; the expert provides them in form of *logic–rules*. It is important to note that no complete definition of the domain may be provided; this implies a great difference with classical Knowledge-Based systems, where the completeness of the Knowledge Base determines the predictive capacity.

The use of rules in the clustering process contributes (acting as a semantic bias) to increasing the classification quality (and to decreasing the computational cost). In fact, the rules act as selectors that cluster objects which could be considered similar in the expert's experience. So the resulting classes (and therefore their prototypes) tend to be more meaningful to the expert's eye.

After a previous phase of detecting and solving the rule conflicts, the conditions imposed by the expert are used to induce a sort of *super-structure* on the domain, the *classes induced by the rules*. A *residual* class is formed with all the objects for which no prior knowledge is given. Clustering will be performed *within* each class induced by the rules and prototypes are built for each one of them. Another important feature is that different kind of variables are considered. So, mixed distances are needed for clustering in order to evaluate distances between objects partially described by numerical variables and partially described by qualitative ones. In [3] details are given of a new family of *mixed-metrics* used in our applications. This distance has given good results until now, but others could also be used [6].

Finally, the elements of the residual class and the prototypes of the rules–induced classes are clustered together and a a global structure is found. Hierarchical clustering is especially suited to our purposes, mainly because the expert can provide heterogeneous knowledge, *i.e.* very specific knowledge of small parts of the domain, together with more general knowledge about other parts and the prototypes will join global hierarchy at different levels.

The idea is to obtain *cooperation* between a knowledge-based process and a clustering one so as to analyze a complex domain. The part not described by the Knowledge Base is analyzed by the Data Mining method. Final description of the domain is the result of combining the clustering with the Knowledge Base. Several real applications show how including rules improves the quality of the results, in the sense that it produces classes with clear meaning.

At the end of this process, the system has *acquired* the knowledge needed to organize the domain, and the expert has succeeded in making explicit his knowledge in a relatively friendly way (see [2] for a complete and detailed description of this methodology). To some extent, the system can act in a similar way to a supervised learning method.

After that, interpretation–oriented tools help the expert to *understand* which clusters were formed and *why*. Automatic generation of final reports is available;

prototype generation is the basis for describing the classes. Our idea is to provide human interpretable descriptions of the classes.

There is a last step oriented to the consolidation of the discovered knowledge related to automatic rules generation. This is useful for later predictive goals.

3 About definition and preprocessing the data

Firs of all, we would like to point out that the system is able to store *metadata*, such as the modalities of each categorical variable, or the range of definition of certain numerical ones. This will make the interpretation of the results, and the data cleaning process easier.

During data cleaning, an important aspect is *missing* data treatment. At the begining, imputation was carries out, as usual, *before* processing data. It was observed that this introduced some incoherencies in the classes induced by the rules, and other possibilities were considered. In this paper, results imputing missing data *after* rule evaluation are presented in §5.

The system is prepared to deal with missing values during the rules evaluation: rules are evaluated only on the present information. Then, the knowledge provided by the expert can be taken into account to substitute for the missing values. Involving the Knowledge Base in inputation is, of course, better that carrying out the imputation under absolute lack of knowledge, as happens when imputation is a previous step to the analysis. Indeed, improvement in the quality of the final clusters appears, in the sense that more compact classes are obtained (see §5). Several real applications show the same behaviour referring this point.

4 Interpreting tools

Actually, given a partition (classification) of a large set of objects it seems necessary to introduce tools for assisting the user in the interpretation tasks, in order to establish the *meaning* of the resulting classes. Often it is not enough for the user to automatically obtain the classes, but to understand *why* those classes where detected. This is also another important point of a *KDD* system and this section presents some ideas concerning our own approach to this topic.

4.1 Class characterization and automatic rules generation

Some statistical packages include several tools to orient the interpretation of a given classification, such as the *contribution* of a certain variable to the formation of a given class, but finally, the interpretation itself must be done by the user in a non-systematic way. We provide a system, based on the use of the representative of each class (see [3]), to find a characterization of a given class, automatically. The idea is to identify, if possible, the variable X_k and the values $\mathcal{D}_k^C$ that allow identification of each class ($i \in C \Leftrightarrow x_{ik} \in \mathcal{D}_k^C$). Sometimes, pairs of variables are needed to distinguish a certain class from the others. Intersection analysis is

required. Even three or more variables may be needed to characterize a class. In this case, the analysis has a combinatory complexity. Our proposal is to introduce negative information and a recursive method based on consecutive conditioning of descriptions. Characterizations are given in logic terms and the *semantics* of the classes is then obvious to the expert.

With these characteristic descriptions, a method for automatic generation of classification rules is designed. The resulting Knowledge Base can be used as a predictive tool for new items.

To provide details on these techniques is not the purpose of this paper. In the application (§5) an example of them can be found.

4.2 Comparing two classifications

The index: Sometimes it is interesting to compare two classifications $\mathcal{P}_1 = \{C_i^1, i = 1 : n_1\}, \mathcal{P}_2 = \{C_j^2, j = 1 : n_2\}$ (including the case $n_1 < n_2$) of the same set of n objects. When several methods are used in parallel for processing the same dataset, it is interesting to evaluate whether results differ widely or not. Also, if two prior Knowledge Bases are provided by different experts, it is interesting to quantify if the degree of discordance is important or not.

Statistical literature provides some tests relative to the independence of two classifications. The most known is, may be, the χ^2 independence test; others are the test of Akaike.... But we are not really interested in testing the independence of two classifications, but in assessing when differences between two classifications may be disregarded. One natural measure for that is the index $\delta(\mathcal{P}_1, \mathcal{P}_2) \in [0, 1]$ (from now on, it will be noted as δ for short). If $n_{ij} = card\ C_i^1 \cap C_j^2$:

$$\delta(\mathcal{P}_1, \mathcal{P}_2) = 1 - \frac{\sum_{(ij)\in\mathcal{N}} n_{ij}}{n}, \qquad \mathcal{N} = \{(ij) : n_{ij} = \max_i n_{ij} = \max_j n_{ij}\} \quad (1)$$

Grosso modo, it can be interpreted as the percentage of cases not equally classified by $\mathcal{P}_1$ and $\mathcal{P}_2$. If $\mathcal{P}_1$ is a reference partition of the objects — provided by the expert or some other source —, then $1-\delta$ may also act as a quality coefficient. In [2], formal definition and details on that index are provided. For short, it can be said that δ is adimensional (which allows comparisons), $0 \leq \delta \leq 1$, if $\mathcal{P}_1 = \mathcal{P}_2 \Leftrightarrow \delta = 0$.

The test: Anyway, in real applications, the index by itself is not enough to decide if small differences between two partitions of a given large set of objects can be dismissed or not. A significance test on that index isconstructed to decide when two partitions can be considered statistically equal or not. The test is

$H_0 : \mathcal{P}_1$ *different from* $\mathcal{P}_2$ **The decision rule:** *if* $P_{H_0}(\delta < d_0)$ *too small* $\Rightarrow$ *reject* H_0
$H_1 : \mathcal{P}_1$ *equal* $\mathcal{P}_2$

To built a statistical test, the reference distribution for the statistic δ supposing H_0 true must be known (p_δ). This distribution can be calculated theoretically, but it involves a combinatory probability problem; it suggests the use

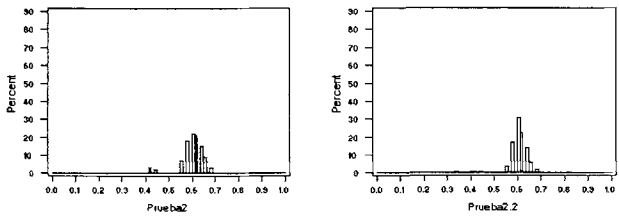
of a non parametric technique to estimate it. A prove is designed on the basis of the permutation test proposed by Fisher. The main idea is to simulate a sample of δ values for building its empirical distribution ($\hat{p}_\delta$) as an estimate of p_δ.

To achieve this, a random sample of k pairs of classifications (which will be represented in the form of rectangular tables) is built and δ is calculated upon each of them. Tables are generated by permuting the n objects through the different classes. However, some precautions are needed for these permutations, since p_δ changes depend on several conditions. In order to obtain comparable values for δ, factors n, n_1, n_2 and marginal distributions will be fixed for permutations. Given the estimate of p_δ, an estimation of the p-value for the observed value of δ, namely d_0, can be calculated and the test can be solved. Even a 95% confidence interval for that p-value can be calculated: $p-value \pm 1.96\sqrt{\frac{p-value(1-p-value)}{k}}$.

Figure 1 illustrates how permutations are well designed: two different tables with equal characteristics (*i.e.* $n, n_1 \ldots$) give, indeed, the *same* $\hat{p}_\delta$.

$\mathcal{P}_1 \backslash \mathcal{P}_2$	1	2	3	4	
1	28	0	7	0	35
2	0	9	3	3	15
	28	9	10	3	50

$\mathcal{P}_1 \backslash \mathcal{P}_2$	1	2	3	4	
1	20	6	7	2	35
2	8	3	3	1	15
	28	9	10	3	50



$n = 50$, $n_1 = 4, n_2 = 2$
marginal of $\mathcal{P}_1$: (35,15)
marginal of $\mathcal{P}_2$:(28,9,10,3)

Fig. 1. Good fit of p_δ.

The analysis of the test: An additional experiment was designed in order to obtain more information on the behaviour of the p_δ. The experimental conditions include the factors and levels listed in table 1. Among the 3125 possible tables, 40 were randomly generated for our study.

n *(sample size)*	$n_1 (= card\ \mathcal{P}_1)$ $n_2 (= card\ \mathcal{P}_2)$	*Matrix type*	*Type of marginals (of $\mathcal{P}_1$ and $\mathcal{P}_2$)*
25	2	*Independent-like*	*uniform*
50	4	*Diagonal*	*uniform-like*
100	7		*two modalities greater than the rest*
250	11		*one modality greater than the rest*
1000	15		*one modality much greater than the rest*

Table 1. Experimental factors considered in the experiment.

For each table, the proposed non parametric technique was used to obtain a reference distribution for δ under the hypothesis H_0. Some characteristics of this reference distribution were recorded: mean, standard deviation, minimum, maximum, symmetry and kurtosis. The observed value for δ, (d_0) and the $p - value(d_0)$ which allows resolution of the test are also calculated.

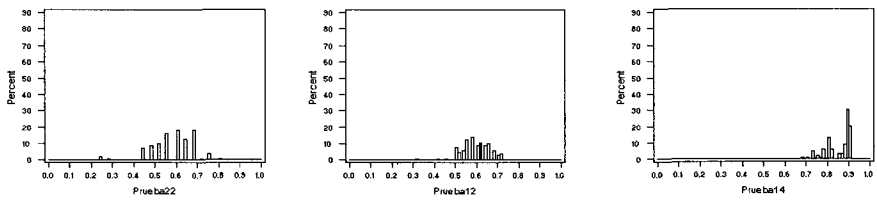


Fig. 2. Different forms of $\hat{p}_\delta$.

Resulting reference distributions have very different forms (see figure 2). A global analysis of the relationships among variables was made using Multivariate Analysis Techniques. The conclusions are the following:

- Changes in factor values $(n, n_1, \ldots)$ change the form of p_δ.
  - Variance decrease with n, since δ is defined as a proportion $\delta \propto \frac{1}{n}$.
  - Kurtosis decrease with n.
  - Symmetry increase with n.
  - In some cases (some rectangular tables) n-modality can be presented.
  - Big n_1, n_2 and small n produce uniform distributions for p_δ.
- In general, as n increases, p_δ becomes symmetric, less pointed, more continnuous: it tends to a normal distribution.
- The form of p_δ is orthogonal to d_0: reference distribution of δ is found independently of the observed table.
- Localization of p_δ depends on d_0. This is a consequence of maintaining the marginal distributions and sample size for permutations constant. Non comparable elements will be found otherwise.
- Diagonal tables give null p-values: When the two partitions are similar, significant p-values are found, as wanted.

5 An application in medicine: Thyroid dysfunctions

Among other real applications, an application to medicine was selected for this paper. The data base comprises results of the routine assays performed at Clinical Hospital *Setre Milosrdnice*, Zagreb (Croatia), and it has been collected over a period of two years.

A sample of 1002 patients was described by 12 laboratory tests and 3 factors, relevant to the outcome of diagnosis. The study took place using a subset of 6 variables — total triiodothyronine ($T3$), total thyroxine ($T4$), thyroid stimulating hormone (TSH), gender (male, female), age and drug therapy (thyrosuppresion, thyroid hormone, without therapy) —, in order to allow future comparison with previous classifications performed by other methods [9] on the same sample (and the same variables).

By means of physical examination, experienced physicians decided on diagnosis of thyroid function state of each patient: euthyreosis (842 cases), hyperthyreosis (104 cases) or 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 [9].

The variable *Diagnosis* could act, in some sense, as the *response* variable. In consequence, it was not considered in the clustering process. So, the key in this application is not discovery of classes, but their *characterization* and the possibility of establishing a Knowledge Base to identify the diagnosis of a new case. This is one of the applications in which *CBR* can perform, to some extent, as a supervised learning method.

A first study was presented in [5] where interest of including rules was shown. *CBR* was used with a simple set of rules provided by the expert:

$If\ T3 = \text{Normal} \wedge T4 = \text{Normal} \wedge TSH = \text{Normal} \longrightarrow Euthyreosis$ (non-ill patients)
$If\ T3 = \text{High} \wedge T4 = \text{High} \wedge TSH = \text{Low} \longrightarrow Hyperthyreosis$

Final classes obtained by using the rules were semantically interpretable, while classes obtained by classical hierarchical clustering were difficult for the expert to interpret (see tree in figure 3 (*left*), details in [5]). An automatic characterization for the partition was found and interpretation of the classes was clear:

$$(\mathcal{C}_1, (Therapy = No) \wedge (Age \in [30, 60]) \wedge (Gender = Female) \wedge (T3 = Normal)),$$
$$(\mathcal{C}_2, (T3 = High)), (\mathcal{C}_3, (Age > 60)), (\mathcal{C}_4, (Age \in [16, 30])),$$
$$(\mathcal{C}_5, (Therapy = Thyrosuppression)), (\mathcal{C}_6, (Therapy = Thyroidhormone)),$$
$$(\mathcal{C}_7, (Gender = Male)), (\mathcal{C}_8, (T3 = Low))$$

It can be seen that the cluster criteria is a combination of the variables *therapy, age, gender* and *levels of T3* and *T4*, and it is obvious that classification rules can be derived for this characterization.

In a second phase of this research, missing data treatment was considered. Instead of carrying out the missing data imputation as a previous step *before* evaluating the rules, it was done, as indicated in §3, *after* building the rules induced partition. Figure 3 (*right*) shows the structure of the resulting tree. Characterization of classes is, in this case, directly related to diagnosis (see expression 2) and comparison between the resulting partition and variable Diagnosis produces table 2. The accuracy was 92%, which means about 75 misclassified objects out of 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. The observed value of δ, $d_0 = 0.0928$

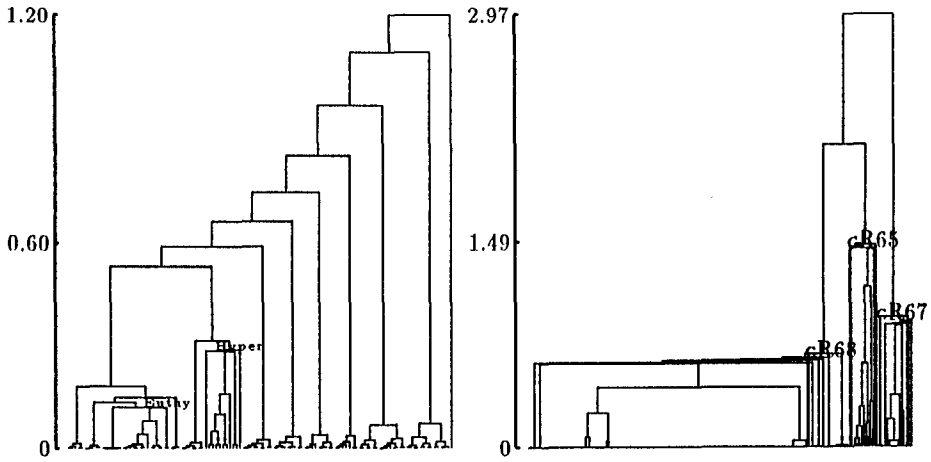


Fig. 3. *(left)* Hierarchical tree obtained with global missing data imputation; *(right)* Hierarchical tree obtained with local missing imputation.

and the permutation test presented in §4.2 produces a $p - value(0.0928) = 0$: partitions produced by the *CBR* method and variable *Diagnosis*, provided by the physicians, are significantly equal, though few discordances are observed.

$$\big((\mathcal{C}_1, (TSH = Normal)), (\mathcal{C}_2, (TSH = High)), (\mathcal{C}_3, (TSH = Low))\big) \qquad (2)$$

The other interesting feature to comment is that it seems to be possible to predict diagnosis using only the level of hormone TSH. This will have to be confirmed whith a more extensive study, actually in progress.

6 Conclusions and future work

CBR successfully combines AI techniques with Statistical Methods, for finding the structure of complex domains (see §2). In this paper, its properties as a *KD* system are presented: taking prior knowledge into account, applying a repeated

Class / Diagnosis	1	2	3	
Hiperthyreosis	31	3	**70**	104
Hypothyreosis	10	**42**	4	56
Euthyreosis	**797**	26	19	842
	838	71	93	1002

$d_0 = 0.0928$
$p - value(0.0928) = 0.000$

Table 2. Relationship between Diagnostic and results of *CBR* with missing inputation *after* rules evaluation.

DM technique (in this case, clustering), including some interpretation-oriented tools to help the user to find the *meaning* of the classes... are some of the common features of a *KDD* process and *CBR*.

Particular aspects were the object of previous papers. In this one, interpretation oriented tools are focussed on:*i*) *Missing* inputation can disturb results if it is done at the very beginning of the study. It seems better to do it after processing prior knowledge. *ii*) Automatic techniques for *characterization* of the generated clusters are presented; prototypes of classes are used to determine their characteristic values. This contributes to an easy interpretation of the classes. *iii*) From this point, research on *automatic generation* of a base of *decision rules* is carried out at present. Preliminary results are presented here. *iv*) Finally, an index to *compare classifications* is presented. When a reference partition exists, the system behaves somewhat like a supervised machine learning system and this index acts as a quality coefficient. Distances between "*expert classifications*" and "*automatic classifications*" can be calculated. *v*) A non parametric *test* assesses the significance of this index (equality of classifications). An experiment was designed to study the form of p_δ in different situations. Conclusions are reported in §4.2.

In the last section an application to the domain of thyroid diagnostics is presented. For this application, *CBR* inputing missing data after rules evaluation can improve the quality of the results, even when a small set of very simple rules is used. From other applications, it has also been seen that the introduction of *semantic* information in the form of rules into the clustering process, generally produces clusters which are easy for the user to interpret.

Acknowledgements: To Dr. Zdenko Sonicki for providing data and for his collaboration. To Juan Carlos Martín, to Juan José Márquez for implementing part of the system.

References

1. Fayyad, U., *et al. From Data Mining to Knowledge Discovery: An overview* **Advances in KD and DM**, Fayyad, U.,*et. al.* R. AAAI/MIT, 1996.
2. Gibert, K, Cortés, U. (98) Clustering based on rules and knowledge discovery in ill-structured domains, Computación y Sistemas, México, 1998. (in press).
3. — Weighing quantitative and qualitative variables in clustering methods, *MATHWARE* **10**(4), January 1997.
4. — Combining a knowledge based system with a clustering method for an inductive construction of models in: P. Cheeseman *et al.* (Eds.), *Selecting Models from Data: AI and Statistics IV*, LNS n° 89 (Springer-Verlag, New York, 1994) 351 - 360.
5. Gibert, K., Sonicki, Z. (97) Classification based on rules and medical research. Proc Applied Stochastic Models and Data Analysis. Ed. Lauro *et al.*. Napoli. pp 181–186.
6. Gower,J. C., A general coefficient for similarity, *Biometrics*, (27) 857–872.
7. Lebart, L *et al.* **Traitement statistique des données**. Dunod, Paris.
8. Nakhaeizadeh, G. *Classification as a subtask of of Data Mining experiences form some industrial projects.* In *IFCS'96*. Kobe, Japan (*in press*). pp. 17–20
9. Sonicki, Z. *et al.* (93) The use of induction in routine laboratory diagnostics of thyroid, *LIJECNICKI VJESNIK* **115**, pp 306–309 (in Croatian).