Unsupervised Information-Based Feature Selection for Speech Therapy Optimization by Data Mining Techniques

Mirela Danubianu, Valentin Popa
Faculty of Electrical Engineering and Computer Science
"Stefan cel Mare" University of Suceava
Suceava, Romania
e-mail: mdanub@eed.usv.ro, valentin@eed.usv.ro

Iolanda Tobolcea
Faculty of Psychology and Education Science
"Alexandru Ioan Cuza" University of Iasi
Iasi, Romania
e-mail: itobolcea@yahoo.com

***Abstract*— Data mining was proven to be an efficient way to find new and useful knowledge in data. Since data dimensionality has major implications on the performance of the algorithms used, one of the data pre-processing operations refers to reducing the number of features. One way to do that is feature selection based on their relevance and redundancy analysis. This paper presents a feature selection method which is applied on data provided by TERAPERS – a computer-based speech therapy system for Romanian children suffering of dyslalia.**

Keywords-data mining, feature selection, feature relevance, feature redundancy, speech disorder therapy

I. INTRODUCTION

The development of the informational society, which led to the increased use of the information technology in the most diverse areas of life, has allowed collecting and storing a huge amount of data. For this reason, over the last years we have witnessed the development of a research area designed to analyze large volumes of data in order to discover valuable and unexpected information, called Knowledge Discovery in Databases (KDD).

Defined as the process of identifying "valid, novel useful and understandable patterns from large data sets" [1], KDD can be viewed as a sequence of several steps. A symbolic representation of KDD process is presented in Figure1 [2].

It starts with a business analysis for determining the KDD goals. Then, there is a data understanding stage which aims to collect and describe data and to verify data quality, followed by the data preparation stage. The core of KDD process is the data mining stage. Data mining involves the analysis of large volumes of data using algorithms which, at acceptable efficiency of calculation, produce a particular enumeration of patterns from such data. As an exploration and analysis technique applied on large amounts of data in order to detect patterns or rules with a specific meaning, data mining may facilitate the discovery, from apparently unrelated data, of relationships that are likely to anticipate future problems or might solve the problems under study. It involves the choice of the appropriate data mining task, and, taking into account specific conditions, the choice and the implementation of the proper data mining algorithm. For the next stage, the mined models are evaluated against the goals defined in the first stage. The last stage of the process uses the knowledge discovered in order to simply generate a report or to deploy a repeatable data mining process.

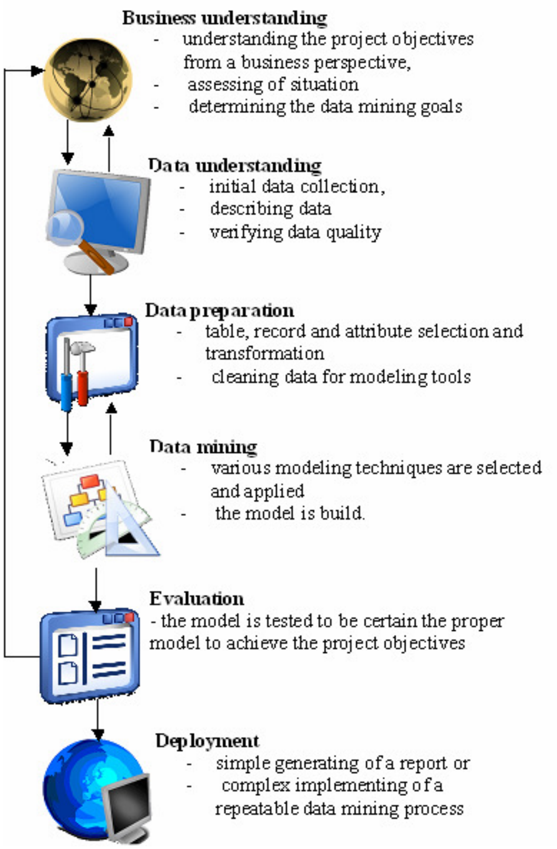

Business understanding
- understanding the project objectives from a business perspective,
- assessing of situation
- determining the data mining goals
Data understanding
- initial data collection,
- describing data
- verifying data quality
Data preparation
- table, record and attribute selection and transformation
- cleaning data for modeling tools
Data mining
- various modeling techniques are selected and applied
- the model is build.
Evaluation
- the model is tested to be certain the proper model to achieve the project objectives
Deployment
- simple generating of a report or
- complex implementing of a repeatable data mining process

Figure 1. Overview of KDD process

Although the stage of applying data mining algorithms is considered the key element of the KDD process, it must be noted that the results provided in this phase are strongly conditioned by several factors such as: data quality and their organization. It is known that in data collected from various primary sources one can find missing values, distortions misrecording or inadequate sampling. Therefore, it is very important to carefully examine the data before carrying out further analyses. Moreover, as one of the most critical operations in the KDD process, the proper preparation and transformation of the initial data set are essential in order to produce useful features for the selected data mining methods.

Data preparation is focused mainly on two issues: firstly, the data must be organized into a standard processing form by data mining algorithms, and, secondly, the data sets used must lead to the best performance and quality for the data mining stage.

II. Dimensional Data Reduction for Data mining

Nowadays, huge amount of data are easily collected and stored. The dimensions of a data set are determined both by the number of cases and by the number of features considered for each case. Most data mining techniques may not be effective for high-dimensionality data, so the solution consists in data dimensionality reduction. To analyze the opportunity of data reduction we need to know what are the gains and losses, and therefore, we must compare computing times and predictive or descriptive accuracy for the model built for the whole dataset with those built for reduced data sets.

In order to reduce the number of cases, sampling or filtering can be used. By filtering, the cases that do not satisfy an imposed condition can be removed from the analyzed data set; by sampling, a subset of cases with a similar behavior to the whole population can be built. In the last case, a sampling error always occurs: it decreases with the increase in the size of subset, and it becomes zero when the complete data set is considered. The size of a suitable subset is calculated by taking into account the computation cost, the accuracy of the estimator and some data characteristics.

On the other side, feature reduction may be achieved either by feature selection or by feature composition. These methods should produce fewer features, so the algorithms can learn faster and even the accuracy of the built models could be improved [3].

Feature composition involves data transformations that can improve the results and performances of data mining operations. Feature selection aims to detect a subset of features having data mining performances comparable to the full set of features, but with significantly reduced computational costs. This is possible using either feature ranking or minimum set algorithms.

Feature ranking algorithms provide ranked lists of features, ordered according to specific evaluation criteria such as: data accuracy and consistency, information content or statistical dependencies between features. They provide information on the relevance of a feature compared to the relevance of other features, without showing the desirable minimum set of the features. On the other hand, minimum subset algorithms consider that all features have the same relevance and return a minimal set to be used for further analyses.

Feature selection depends on the overall processing goal and its performance evaluation criteria, on the existing data set and the type of model targeted, on the original set of pattern features and on the defined feature selection criterion.

Data dimensionality reduction affects all phases of a data mining process. It must be started in the data preparation stage. In many cases, feature reduction is part of the data mining algorithm and it can also be applied in the evaluation stage for a better evaluation and consolidation of the results obtained.

We can therefore conclude that, by means of the data dimensionality reduction, we aim to improve the performance of the data mining operation, as well as that of the resulted models, to reduce the model dimensionality without affecting its quality, and last but not least, to allow the user to visualize results in fewer dimensions in order to improve the decision making process.

III. Feature Selection Based on Relevance and Redundancy Analysis

Practice has demonstrated that irrelevant input features induce great computational costs for the data mining process and may lead to overfitting. To avoid these drawbacks, feature selection research has focused on the choice of relevant features from the whole data set [4]. Some results have also revealed the existence and the negative effect of feature redundancy [3] [5] [6]. The conclusion was that it is necessary to reduce the number of redundant features to a minimum level in order not to affect the accuracy of the model built. In [7] it is stated that "features are relevant if their values vary systematically with category membership". This means that a feature is relevant if it is correlated with the class. This was formally defined in [3] as follows: a feature F_i is relevant iff there are f_i and c for which $P(F_i=f_i) > 0$, so that

$$P(C=c|F_i=f_i) \neq P(C=c) \qquad (1)$$

A complete definition of feature relevance takes into account the existence of three disjoint categories of features named: *strongly relevant, weakly relevant* and *irrelevant features* [8].

Let F be the original set of features, F_i a feature, $S_i = F-\{F_i\}$ and C the class associated.

It can be said that:

- F_i is strongly relevant if
$$P(C|F_i,Si) \neq P(C|S_i) \qquad (2)$$
- F_i is weakly relevant if
$$P(C|F_i,S_i) = P(C|S_i) \text{ and}$$
$$\exists\, S'_i \subset S_i, \text{ so that } P(C|F_i,S'_i) \neq P(C|S'_i) \qquad (3)$$
and, finally,
- F_i is irrelevant if

$$\forall\, S'_i \subset S_i,\ P(C|F_i,S'_i) = P(C|S'_i) \qquad (4)$$

A feature with strong relevance is always necessary for an optimal subset and it cannot be removed without affecting the original conditional class distribution. A weakly relevant feature is not always necessary but in certain condition it may become necessary, whereas an irrelevant feature is not necessary at all.

Feature redundancy can be expressed using the feature correlation property, since it is accepted that two features are redundant to each other if they are completely correlated. In order to define the redundancy of features, it is useful to define the feature's Markov blanket [5].

Let us consider the notation mentioned above, and let be $M_i \subset F$ ($F_i \notin M_i$). M_i is said to be a Markov blanket for F_i if

$$P(F\text{-}M_i\text{-}\{F_i\}, C|F_i, M_i) = P(F\text{-}M_i\text{-}\{F_i\},C|M_i) \quad (5)$$

The condition above requires that M_i contains both the information that F_i has about C and about all the other features.

Finally, we could say that a feature F_i is redundant and it should be removed from F if and only if it is weakly relevant and it has a Markow blanket M_i within F.

A short look over a whole set of features reveals that it may contain four disjoint parts. These are: irrelevant features, redundant features as part of weakly relevant features, weakly non-redundant relevant features and strongly relevant features. An optimal subset must contain all relevant features and the weakly relevant but non-redundant ones.

Relevance is usually defined in terms of correlation or mutual information, so the mutual information on the data can be used as a feature selection criterion. In order to define mutual information for two variables (or features) we start from the concept of entropy, as a measure of random variable uncertainty. For a variable X, the entropy is defined as:

$$E(X) = -\sum_i P(x_i)\log_2(P(x_i)) \quad (6)$$

The entropy of a variable X, after observing the values of another variable Y, is defined as:

$$E(X \mid Y) = -\sum_j P(y_i)\sum_i P(x_i \mid y_i)\log_2(P(x_i \mid y_i)) \quad (7)$$

where $P(x_i)$ is the prior probability for all values of X, and $P(x_i|y_i)$ is the posterior probabilities of X given the value of Y. The value by which the entropy of X decreases, estimates additional information about X provided by Y. It is called information gain [9] and it is calculated using the following expression:

$$I(X,Y) = E(X) - E(X|Y) \quad (8)$$

We take into account that for the discrete random variable, the joint probability mass function is

$$P(x_i|y_j) = P(x_i,y_j) / P(y_j) \quad (9)$$

and the marginal probability function p(*x*) is:

$$P(x_i) = \sum_j P(x_i, y_j) = \sum_j P(x_i \mid y_j)p(y_j) \quad (10)$$

where $p(x,y)$ is joint probability distribution function of *X* and *Y*, and $p(x_i)$ and $p(y_j)$ are the marginal probability distribution functions of *X*, respectively *Y***.** Since these are probabilities, we have

$$\sum_i \sum_j p(x_i, y_j) = 1 \quad (11)$$

Finally, for two discrete random variables *X* and *Y*, information gain is formally defined as:

$$I(X,Y) = \sum_j \sum_i p(x_i, y_j)\log(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}) \quad (12)$$

According to this expression, we could state that a feature Y is more correlated to feature X than feature Z if:

$$I(X,Y) > I(Z,Y) \quad (13)$$

It can be observed that information gain favors features with more values, so it should be normalized. In order to compensate its bias and to restrict its values to range [0,1], it is preferable to use the symmetrical uncertainty [10], defined as:

$$SU(X,Y) = 2\left[\frac{I(X,Y)}{E(X)+E(Y)}\right] \quad (14\ )$$

A value of "1" for symmetrical uncertainty means that knowing the values of either feature completely predicts the value of the other, whereas a value of "0" implies that X and Y are independent.

There are many feature selection methods that consider the subset evaluation approach. In these cases, feature relevance and features redundancy are handled.

In the traditional framework for feature selection using subset evaluation [11], candidate feature subsets based on a certain search strategy are produced. Each of the candidate subsets is evaluated by a certain measure and it is compared with the previous best one with respect to this measure. If the new subset is found to be better, it replaces the previous best subset. These two stages are repeated until a stopping criterion is satisfied. This method poses difficulties due to the searching through the feature subsets.

A new framework proposed in [12] avoids implicitly handling features redundancy and allows an efficient elimination of redundant features by explicitly handling the features redundancy. This framework, presented in Figure 2, consists of two steps: firstly, the relevance analysis is carried out and the irrelevant features are removed; secondly, a redundancy analysis provides the final subset by eliminating the redundant features from the relevant ones. The advantage of this method consists in the decoupling relevance and the redundancy analyses that lead to an efficient way to find a subset that approximates an optimal subset.

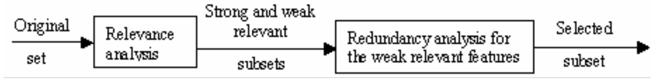


Figure 2. Feature selection through relevance and redundancy analysis

Let us use SU(X,Y) as a correlation measure for both the relevance and redundancy analysis. Such a correlation between any feature F_i and the class C is called C-correlation ($SU(F_i,C)$) and the correlation between any pair of features F_i and F_j ($i \neq j$) is called F-correlation [13].

As we have noted above, the optimal features subset contains those feature which are strongly correlated with the class but are not correlated with each other, and are not redundant. In order to achieve that, C-correlation for each feature must be calculated. Once a relevance threshold γ is established experimentally by the user, one can assume that a feature F_i is relevant if $SU(F_i,C) > \gamma$. After relevant features are selected, they are subject of redundancy analysis. In a natural approach, one could evaluate the correlation between individual features for redundancy analysis, but there are two drawbacks. Firstly, if two features are not completely correlated, it is difficult to determine feature redundancy and which one should be removed. Secondly, this involves calculating the F-correlation for a great number of pairs which it is inefficient for high-dimensional data sets. To avoid these problems it is indicated to approximately determine feature redundancy by approximating Markov blankets for the relevant features found in the previous stage. The basic idea is that a feature with a greater C-correlation value offers more information about the class than a feature with a smaller one. Consequently, when $SU(F_j,C) \geq SU(F_i,C)$, it is necessary to evaluate if F_j can form an approximate Markov blanket for F_i in order to keep more information about the class. For two relevant features F_i and F_j ($i \neq j$), we could say that F_j forms an approximate Markov blanket for F_i if [13] :

$$SU(F_j,C) \geq SU(F_i,C) \tag{15}$$

and

$$SU(F_i,F_j) \geq SU(F_i,C) \tag{16}$$

In (15), $SU(F_i,C)$ is heuristically used as a threshold to establish if the F-correlation $SU(F_i,F_j)$ is a strong one.

So, in order to find the appropriate feature subset, those for which there are Markov blankets, which are redundant, should be eliminated from the relevant feature set.

The whole process is presented in Figure 3.

```
Input:  {F,C} ; F={F_1, F_2, … F_n}
        γ
Output: S_opt

 1. S= ϕ
 2.   for i=1 to n do begin
 3.     calculate SU(F_i,C)
 4.       if SU(F_i,C) ≥ γ
 5.         S=S ∪ {F_i}
 6.   end                  // S contain all relevant features
 7. order  S descending on SU(F_i,C)        // this aims to
    make easier the comparison  between SU(F_i,C) and
    SU(F_j,C) for i≠j
 8. F_j= FirstElement(S)
 9.   do  begin
10.     F_i =NextElem(S,F_j)
11.       if  F_i is not null
12.         do begin
13.           if  SU(F_i, F_j) ≥ SU(F_i,C)
14.             S = S-{F_i}
15.           F_i=NextElement(S,F_i)
16.         until F_i is not null
17.     F_j = NextElement(S, F_j)
18.   until F_j is not null
19. S_opt = S
```

Figure 3. Feature selection method

As it can be observed in the first phase (lines 1-6), one obtains the relevant feature set S. These features are decreasingly ordered (line 7) according to their SU values. Then, the ordered list S is processed (lines 8-19) in order to select the optimum feature subset. This means that the features are filtered based on the presence or the absence of approximate Markov blankets.

IV. EXPERIMENTAL RESULTS

A. *Data Set Description*

We have tested this method on data collected by the TERAPERS system. This is a system which aims to assist the personalized therapy of dyslalia (an articulation speech disorder) and to track how the patients respond to various personalized therapy programs. Implemented in March 2008, the system is currently used by the therapists from Regional Speech Therapy Center of Suceava.

An important aspect of assisted therapy refers to its ability to adapt to the patients' characteristics and evolution. In order to adapt the therapy programs, the therapist must carry out a complex examination of children, through recording relevant data related to personal and family anamnesis. Anamnesis data can provide information on the various causes that may negatively influence the normal development of language. It contains historical data and data provided by the cognitive and personality examination.

The data provided for the personalized therapy programs includes the number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. In addition, the report downloaded from a mobile device collects data on the efforts of child self-employment. The data refers to the exercises done, the number of repetitions for each of these exercises and the results obtained. The tracking of child's progress materializes data indicating the assessing time, and the child's status at that moment. All this data is stored in a relational database, composed of 60 tables.

The data stored in the TERAPERS's database together with the data from other sources (e.g. demographic data, medical or psychological research) compose the set of raw data that can constitute the subject of data mining process. It might be useful, because as it was shown in [14], one can use classifications in order to distribute the people with different speech impairments in predefined classes (if attribute diagnosis contains class label, we can predict a diagnosis based on the information contained in various predictor variables), clustering can be used to group people with speech disorders on the basis of features similarity and to help therapists to understand who are their patients; also, one can use association rules to determine why a specific therapy program has been successful on a segment of patients with speech disorders, while it was ineffective on another segment of patients.

For our experiments, a data set consisting of 102 features with numeric and descriptive values and 400 cases was considered. This is anamnesis data or data derived from complex examinations, based on which classification models will be built, in order to suggest the diagnosis for future

ISBN: 978-1-61208-202-8

cases. Firstly, we have eliminated the features that obviously are not relevant for the objective set (e.g. parents' name and work place) and we obtained 71 features. The feature selection method described above was applied on this data set, and we have compared the performances of the model built on the reduced set of features with those obtained for the model built on the whole data set.

Shown in Figure 4, this experiment is designed and implemented in WEKA [15].

The attribute "*diagnosis*" was considered as class label, and three patients' classification processes were built. There are identical in terms of models, but they differ because the same operators are applied on different datasets.

The first process is carried out on the whole set of features, the second one uses a reduced set which contains all the relevant features, while the third one is applied on the data set formed only by the relevant and non-redundant features.

An experiment containing three processes, each of them using another classification model was carried out (Figure 4). Two rules classification models and a decision tree model (J48) have been considered.

Relevant features are obtained by the C-correlation estimation. As it can be noticed in Figure 5, an ordered list of features is obtained and those for which $SU(F_i,C) = 0$ are removed. The result consists of 52 relevant features. The final feature subset, obtained by removing those for which the expression (16) is respected, (lines 13-14 in Figure 3), contains 10 features.

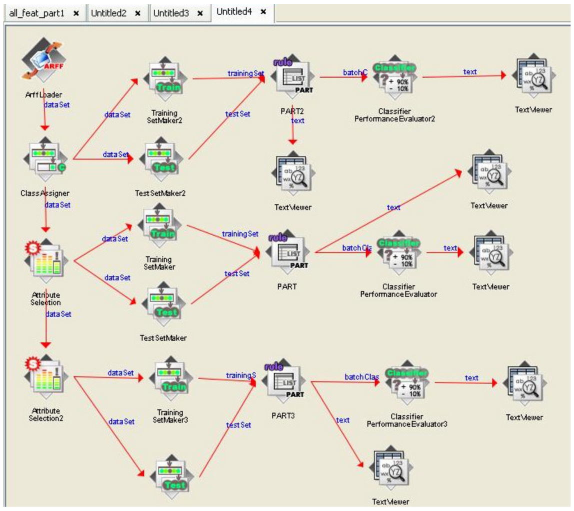
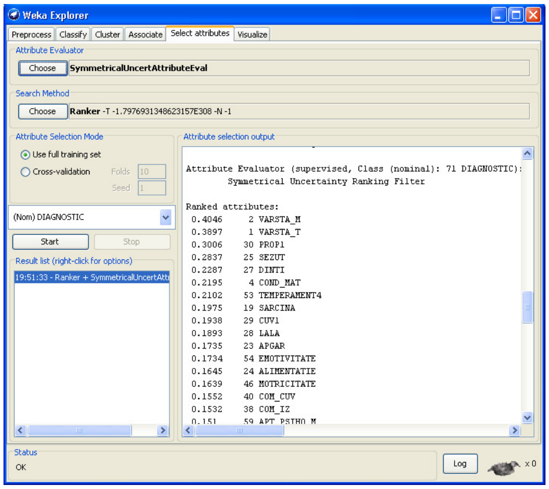


Figure 4. WEKA Knowledge flow for the classification experiment

An analysis of the performances of the three processes, in terms of percent-correct classified cases is shown in Figure 6, and a visual comparison between these performances is presented in Figure 7.

As it can be observed, there are little differences between the percent-correct classified cases for the same classifier applied on the three data sets. For the methods studied, the best results are obtained for the subset of relevant and non-redundant features subset.



Attribute Evaluator: SymmetricalUncertAttributeEval

Search Method: Ranker -T -1.7976931348623157E308 -N -1

Attribute Selection Mode: ● Use full training set ○ Cross-validation Folds 10 Seed 1

(Nom) DIAGNOSTIC

Result list (right-click for options): 19:51:33 - Ranker + SymmetricalUncertAttr

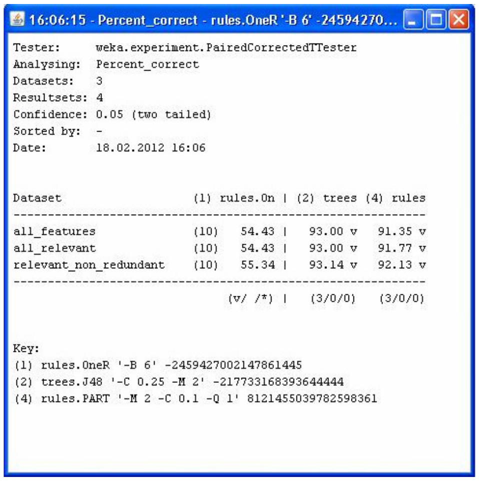
Attribute selection output

```
Attribute Evaluator (supervised, Class (nominal): 71 DIAGNOSTIC):
        Symmetrical Uncertainty Ranking Filter

Ranked attributes:
 0.4046     2 VARSTA_M
 0.3897     1 VARSTA_T
 0.3006    30 PROP1
 0.2837    25 SEZUT
 0.2287    27 DINTI
 0.2195     4 COND_MAT
 0.2102    53 TEMPERAMENT4
 0.1975    19 SARCINA
 0.1938    29 CUV1
 0.1893    28 LALA
 0.1735    23 APGAR
 0.1734    54 EMOTIVITATE
 0.1645    24 ALIMENTATIE
 0.1639    46 MOTRICITATE
 0.1552    40 COM_CUV
 0.1532    38 COM_IZ
 0.151     59 APT_PSIHO_M
```

Status: OK

Figure 5. Partial list of relevant features



```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   3
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       18.02.2012 16:06


Dataset                  (1) rules.On | (2) trees (4) rules
-------------------------------------------------------------
all_features             (10)   54.43 |   93.00 v   91.35 v
all_relevant             (10)   54.43 |   93.00 v   91.77 v
relevant_non_redundant   (10)   55.34 |   93.14 v   92.13 v
-------------------------------------------------------------
                              (v/ /*) |   (3/0/0)   (3/0/0)


Key:
(1) rules.OneR '-B 6' -2459427002147861445
(2) trees.J48 '-C 0.25 -M 2' -217733168393644444
(4) rules.PART '-M 2 -C 0.1 -Q 1' 8121455039782598361
```

Figure 6. Percents of corrected classified cases for the three data sets

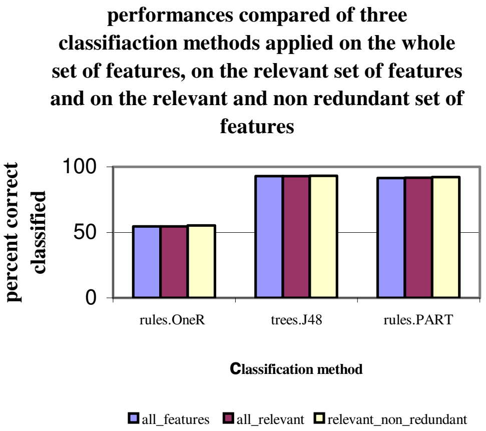


Figure 7. Percents comparison of the corrected classified cases

Significant differences between the three processes have been obtained for the elapsed training time. These results are presented in Figure 8. As it is shown for all the three methods, the best times are achieved for the subset consisting of relevant and non-redundant features.

Practically, for the dataset consisting in 400 cases described above, for the least efficient method (rules.oneR), the training process for the whole set of features lasted 0.37 sec, while for the feature subset containing only relevant and non-redundant features this process it lasted 0.05 sec; for the most efficient method (tree.J48), the elapsed training time for the whole set of features was 0.06 sec and for the relevant and non-redundant features this time was 0.01 sec.

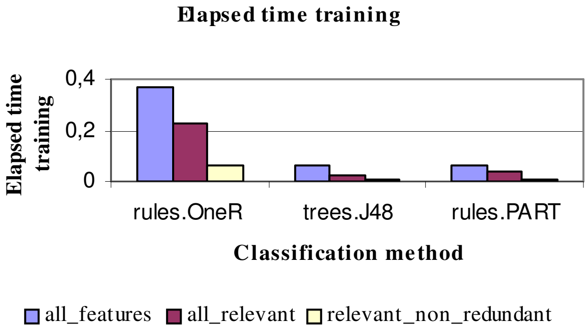


Figure 8. Comparison of elapsed time training

V. CONCLUSIONS AND FUTURE WORK

This work is part of the research that aims to implement a data mining system that will allow the optimization of personalized therapy of speech disorders for children with dyslalia. Combining the feature selection methods with the data mining algorithms is a good practice; therefore, in this paper, we have studied such a method based on the features relevance and redundancy analysis.

This method, applied on the anamnesis data provided by TERAPERS, has shown that both the percent of correctly classified cases and that of the elapsed time for training are better if the considered data mining algorithms are applied on data containing a reduced subset of features.

It must be noted that these results cannot be generalized for all data mining methods and algorithms. This is why we intend to study the impact of feature reduction on clustering and association rules mining.

ACKNOWLEDGMENTS

This paper was supported by the project "Progress and development through post-doctoral research and innovation in engineering and applied sciences– PRiDE - Contract no. POSDRU/89/1.5/S/57083", project co-funded from European Social Fund through Sectorial Operational Program Human Resources 2007-2013.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Cambridge

[2] Danubianu M., Pentiuc S.G., Tobolcea I., Schipor O.A. (2010). Advanced Information Technology - Support of Improved Personalized Therapy of Speech Disorders, *International Journal of Computers Communications & Control*, ISSN 1841-9836, 5(5), pp: 684-692

[3] Kohavi R., John G.(1997). Wrappers for feature subset selection. *Artificial Intelligence, special issue on relevance*, 97(1-2), pp. 273-324

[4] Peng H. Long F., Ding C. (2005). Feature Selection based on mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 27, No. 8,

[5] Koler D., Sahami M. (1996). Towards optimal feature selection. *In Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning.* Morgan Kaufman

[6] Hall M.A.(2000). Correlation-based feature selection for discrete and numeric class machine learning, *In proceedings of the Seventeenth International Conference on Machine Learning*, p. 359-366, 2000

[7] Gennari J.H., Langley P., Fisher D.(1989). Models of incremental concept formation. *Artificial Intelligence*, (40), p. 11-16

[8] John G.H., Kohavi R., Pfleger P.(1994). Irrelevant features and the subset selection problem. *In Machine Learning: Proceedings of the Eleventh International Conference*, Morgan Kaufman

[9] Quinlan J.R.(1993). C4.5: Programs for Machine Learning, Morgan Kaufmann,

[10] Press, W.H., Teukolsky, S.A., Vetterling. W.T., Flannery. B.P.(1988) Numerical Recipes in C, Cambridge Univerity Press, Cambridge

[11] Liu, H., Motoda, H.(1998). Feature Selection for Knowledge Discovery and Data Mining, Boston Kluwer Academic Publishers, ISBN 0-7923-8198-X

[12] Yu, L., Liu, H.(2004). Redundancy based feature selection for microarray data, *Proc of the Tenth ACM SIGMOD Conference on Knowledge Discovery and Data Mining*, pp. 737-742

[13] Yu, L., Liu, H.(2004). Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, 5, pp. 1205-1224

[14] Danubianu M., Pentiuc St. Gh., Socaciu T. (2009). Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques, *The Fourth International Multi Conference on Computing in the Global Information Technology ICCGI 2009*, Vol: CD, 23-29 August, Cannes - La Bocca, France

[15] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Vol. 11, issue 1.