Data Dimensionality Reduction Framework for Data Mining

M. Danubianu[1], St. Gh. Pentiuc[1]
[1]*Faculty of Electric Engineering and Computer Science, 'Stefan cel Mare' University of Suceava, Universitatii, 1, Suceava, Romania*
mdanub@eed.usv.ro

***Abstract*— The database built by TERAPERS project contains a considerable volume of data about the personal or familial anamnesis, and regarding the process of personalized therapy of dyslalia. This data can be the starting point of data mining processes that could provide useful information for the design and adaptation of different therapies to obtain the maximum efficiency. Because data dimensionality affects the performances of data mining tasks, this paper presents two supervised feature selection methods to be used in the frame of an information system. These methods were validated by experiments in the classification of Romanian patients with speech disorders. Obtained results will be used to implement Logo-DM, which is intended to be a data mining system aiming to optimize the personalized therapy of dyslalia.**

***Index Terms*—Data mining, data pre-processing, feature selection.**

I. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a complex process which aims to extract new, interesting and potential useful patterns from large amounts of data. Its central point is data mining, which effectively builds models from data. Many factors such as type and quality of data affect the success and performance of data mining tasks.

Theoretically, having more data, results are more precise. However, practical experience with data mining algorithms has shown that this is not always true. Feature subset selection, as a process of identification and removing as much irrelevant and redundant information as possible, reduce data dimensionality, allow building patterns faster and more effectively and sometimes improves accuracy of future classification.

Using TERAPERS system in order to assist the therapy of spech disorders allows specialists to collect a considerable volume of data about the personal or familial anamnesis, and regarding the process of personalized therapy. This data can be the foundation of data mining processes that could provide information for designing and adaptation of different therapies in order to obtain the best results at maximum efficiency.

This paper aims to make a study on feature selection opportunity and on proper methods to be used in Logo-DM system. This is a data mining system that we will develop in order to help specialists which use computer-based speech therapy systems to optimize their personalized therapeutically path.

II. DATA DIMENSIONALITY AND KNOWLEDGE DISCOVERY IN DATABASES

Data mining involves the application of algorithms able to detect patterns or rules with a specific means from large amounts of data, and represents one step in knowledge discovery in database process.

There are more techniques for extracting patterns from data, but the most commonly used are: classification, clustering and association rules. It can be stated that data mining is intended to be a specific form of automatic learning, where the environment is seen through a database [1] [2].

KDD definition refers to the fact that large data sets are used. The size of such data set is determined by the number of cases analyzed and the number of features considered for each case. It was found that if there are a lot of features, it is possible that the number of cases in data set to be insufficient for data mining operations.

The high dimensionality of data can cause also data overload, and make some data mining algorithms non applicable. The solution for these problems is data dimensionality reduction as is shown in Fig. 1.

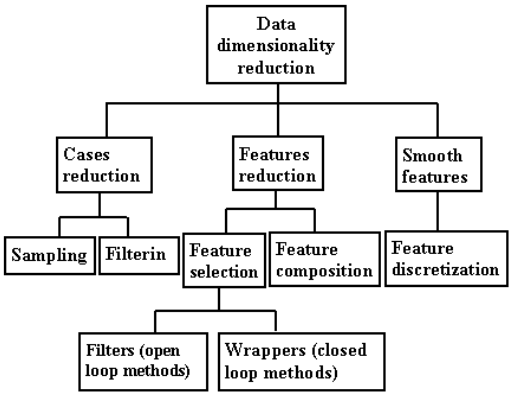

Fig. 1. Data dimensionality reduction methods.

Manuscript received July 18, 2012; accepted February 20, 2013.

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.

There were proposed many ways to data reduction, but the most used are those that attempt to preserve the character of the original data by deleting data that are not essential.

The most critical dimension in the original data set is often the large number of cases. This can be reduced either by sampling or by filtering. Filtering refers to the removal from analyzed data set of those cases that do not satisfy an imposed condition. Sampling aims to build a subset of cases which has a similar behavior with the whole population. The size of a suitable subset is calculated by taking into account the cost of computation, the accuracy of the estimator and some characteristics of data.

Most data mining techniques were not designed to cope with large amounts of features, so dealing only with relevant features is effective and efficient.

The feature reduction process should produce fewer features and less data so the algorithms can learn faster. At the same time it might provide higher accuracy for the resulted models and improve the comprehensibility of extracted knowledge [3].

In the real-world applications, two standard classes of tasks are used to produce a reduced set of features: feature selection and feature composition. Feature composition depends on knowledge of the application, and consists in various data transformation that can positively influence the performance of data mining operations.

Feature selection aims to choose a subset S_x of the complete set of input features $X=\{x_1, x_2, \dots x_k\}$ so that this subset will predict the output Y with an accuracy comparable to that obtained if the whole set X is used, but with a significantly reduced computational cost. Taking into account the feature selection criteria used, current methods are divided in two classes [3]: open loop methods or filters and closed loop methods or wrappers.

Open loop methods (Fig. 2) are based on selecting features through the use of separability between class criteria. These approaches do not consider the effect of selected features on the performance of the whole process of knowledge discovery since the used criteria does not require a predictor evaluation for reduced data sets.

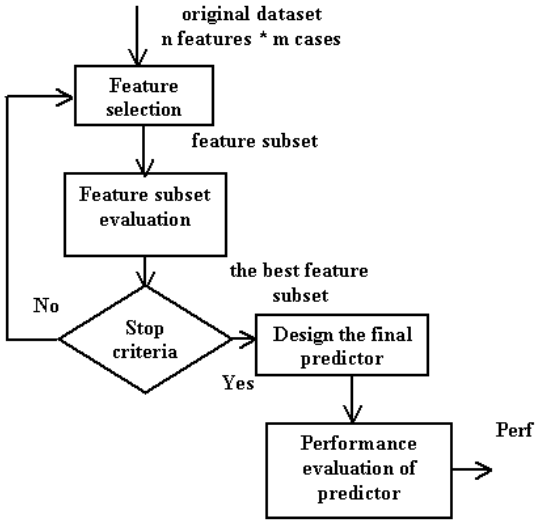

Fig. 2. Open loop feature selection method.

Closed loop methods (Fig. 3) use as criteria for feature selection the predictor performance. The quality of a selected feature subset is evaluated using as criteria

$$\text{Perf}_{\text{feature}} = \text{Perf}_{\text{predictor}}, \quad (1)$$

where $Perf_{predictor}$ is the performance evaluated for a whole prediction algorithm applied on reduced data set.

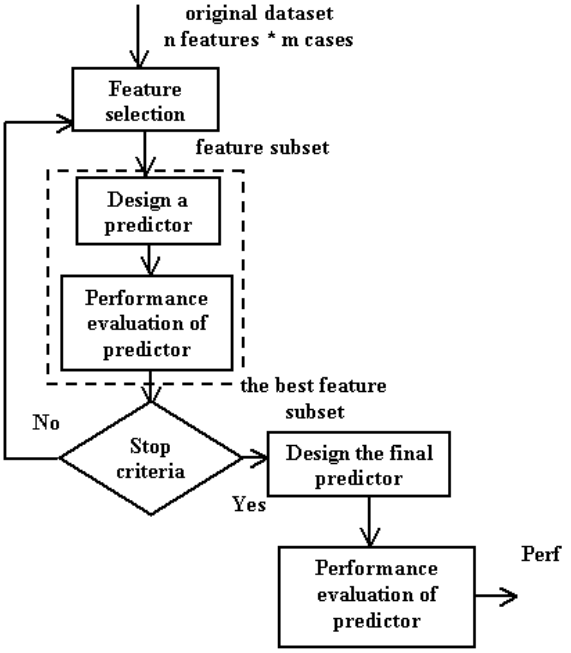

Fig. 3. Closed loop feature selection method.

Although the open loop methods are computationally less expensive, closed loop methods provide better selection for feature subsets since they allow best prediction.

There is a large number of search strategies, but in the following will be considered the sequential ones, which add or remove features sequentially. The most representative methods of this type are forward and backward feature selection. Forward selection search [4] starts from an empty subset S_x. In each step all possible sets of features built by adding a new feature x_i to the previous selected features are evaluated, and finally is chosen the set that provide the maximum increase of the performance criteria *Perf* considered. This process continues until the best features subset $\boldsymbol{S_x}$ is found, as in the following algorithm.

Let $\boldsymbol{E}$ be the input set of n cases described by the feature set $\boldsymbol{X}=\{x_1, x_2, \dots x_k\}$, and ε a relative threshold value

1. $\boldsymbol{S_x}=\emptyset$; $Perf_{old}=0$
2. $j=1$
3. $i_o = \arg\max \{Perf(\boldsymbol{S_x}\cup\{x_i\})/\forall\, x_i\in\boldsymbol{X}\setminus\boldsymbol{S_x}\}$
4. $Perf_{new} = Perf(\boldsymbol{S_x}\cup\{x_{io}\})$
5. if $(Perf_{new}-Perf_{old})/Perf_{new}<\varepsilon$ then stop
6. $\boldsymbol{S_x} = \boldsymbol{S_x} \cup \{x_{io}\}$
7. $Perf_{old} = Perf_{new}$
8. $j = j+1$
9. if $j\le k$ then go to step 4
10. stop

It may be noted that at the step 4, are evaluated k-j+1 candidate subsets, thus for j=1,..,m are evaluated $s=m(2k-m+1)/2$ subsets. But the number of all possible subsets of m

features from an original set of k features is $p=k!/(k-m)!/m!$. Because $s<p$ the solution provided by the forward feature selection algorithm is a sub-optimal one.

Backward selection [4] starts with the entire feature set and removes one feature at a time. Firstly the selection criteria is evaluated for the whole set of features. Then, all possible subsets with one feature discarded are formed and their performance according the *Perf* criteria are evaluated. At each step the feature which causes the smallest deprecation of performance is removed. This procedure continues until the best k-features subset is selected, as shown below.

```
    Let E be the input set of n cases
described by the feature set X={x1, x2, … ,
xk} and ε a threshold value
1.  Sx=E
2.  j=1
3.   Perfold = Perf(Sx)
4.   io=argmin{Perfold-Perf(Sx\{xi})/∀ xi∈Sx}
5.    if Δio>ε then stop
6.   Sx = Sx \ {xio}
7.   j = j +1
8.  if j<k then go to step 4
9.  stop
```

This approach provides also a suboptimal solution since for great values of k is impossible to examine all possible subsets of features. The exhaustive search of an optimal subset of features from an initial set of k features is $O(2^k)$, thus the optimal feature selection is a NP-hard problem, and from this reason are used the sub-optimal feature selection methods such as forward and backward selection.

III. Feature Selection for Logo-DM

The idea of trying to improve the quality of logopaedic therapy by applying some data mining techniques started from TERAPERS project developed within the Research Center for Computer Science in the University "Stefan cel Mare" of Suceava [5].

This project has developed a system which is able to assist speech therapists for personalized therapy of dislalya and to asses how the patients respond to various personalized therapy programs [6]. This system is currently used by the therapists from Regional Speech Therapy Center of Suceava.

In the context of the need for more efficient activities, it was shown that data mining methods, applied to data collected in TERAPERS, can provide useful knowledge for personalized therapy optimization.

Logo-DM is a data mining system, which is intended to use data collected in TERAPERS in order to answer the questions such as: what is the predicted final state for a child or what will be his/her state at the end of various stages of therapy, which the best exercises are for each case and how they can focus their effort to effectively solve these exercises or how the family receptivity, which is an important factor in the success of the therapy - is associated with other aspects of family and personal anamnesis [7].

The available data set consists of 60 relational tables. These data and data from other sources (eg demographic data, medical or psychological research) is the set of raw data that will be the subject of data mining [8]. Currently, the system contains data about 312 cases.

Creating target data set is accomplished through a join of tables containing useful features. Data set necessary to establish the profile of children with speech disorders, can be obtained by joining tables which contain general data about children, family and personal anamnesis, data on complex evaluation and diagnosis associated. The result is a set of 102 features.

We aim to realize a feature selection from data mentioned above in order to build an effective prediction model of diagnosis for new cases, based on anamnesis data of children who were diagnosed with various speech disorders.

In Fig. 4. the succession of the requested operations is presented.

We started from the original data set containing 312 cases and 102 features. For feature selection is used both *forward selection* and *backward selection* algorithm, and as predictor is considered a decision tree classifier. It is used the same predictor both for feature selection and future predictions.

In the first stage we have eliminated those features that obviously are not relevant for the proposed objective (e.g. name and work place of parents). The resulted data set, consisting in 96 features, is used in two directions. First it forms the basis for feature selection methods. Second it serves to build a model, whose performance is a reference that makes possible the assessment of the model built on reduced data set.

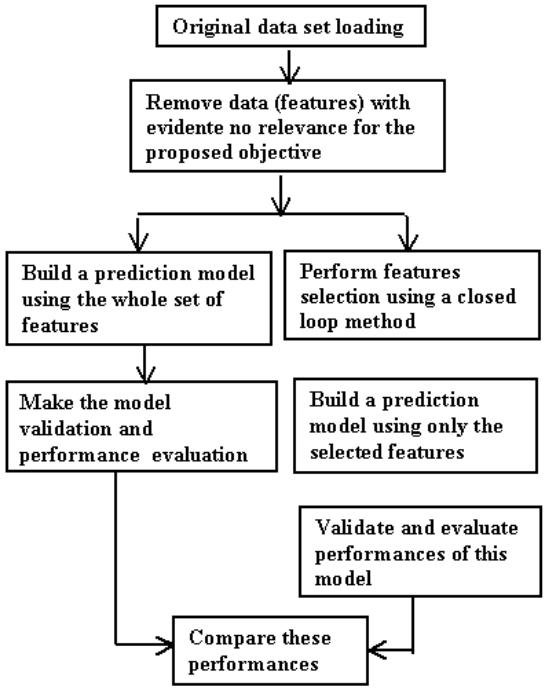

Fig. 4. The proposed feature selection framework.

The experiment was performed through a process designed and executed in RapidMiner5 [9].

Analyze of the results was focused on comparing the performance obtained by validating models built from subsets of features with the model trained on the original set.

Model accuracy and classification error have been considered according to different values for information gain used to split the decision trees [10].

It has been also compared the execution times for the modeling processes performed on the whole set of data and those performed on features subsets.

And last, but not least we have analyzed the models obtained in order to see which of them is more easily to understand and interpret.

In Fig. 5 is presented a comparison of classification accuracy of three different models: one built using the original set of features (acc_orig) and two others based on subsets of feature selected in closed loop processes using forward selection (acc_fw) and backward selection (acc_bw).

The variation of the execution times for the processes that involve feature selection operations is shown in Fig. 6.

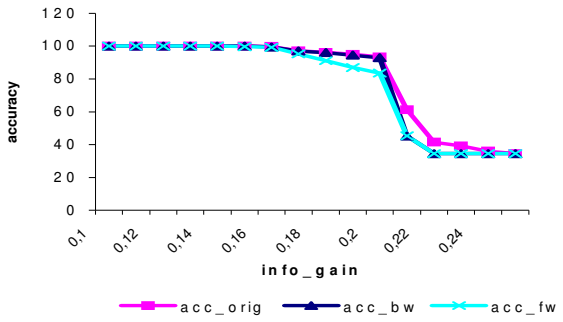

Fig. 5. Performance of models build on the original data set and on feature reduced data sets.

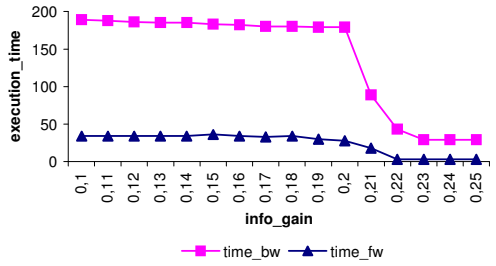

Fig. 6. Execution times for feature selection by backward and forward search.

Note that for an information gain less or equal to 0.17 all the three models have the same accuracy, then for information gain between 0,18 and 0,24 there are small differences among the values of accuracy obtained and final for information gain greater than 0,24 they are again equal. Forward selection algorithm led to a subset of four features for a good classification while the backward selection has built a subset that contains 70 features.

Although there are significant differences between the two subsets, in terms of number of features contained, accuracy varies insignificantly for different values of information gain. In this context it raises two problems: what subset will be chosen, and as a consequence what method will be used for proper feature selection. Since times needed for feature selection affect the cost of the entire process, and the measurements have shown that forward selection method is faster by an order of magnitude than backward selection, it is considered that for this case forward selection is preferable.

IV. CONCLUSIONS

The results presented in this paper are part of the research and experiments that aims to implement a data mining system that will allow the optimization of personalized therapy of speech disorders for children suffering of dyslalia. It was considered the pronunciation of phonemes characteristic for Romanian language.

It is started from the finding that the performances of the data mining operations and ultimately of the whole KDD process are strongly influenced by data dimensionality. This is determined by the number of cases considered and/or by the number of features that describes these cases. Many of these features are correlated as a result not all of them are absolutely necessary for building a valid model

They were studied two closed loop methods to reduce the set of features for a decision tree classifier. As evaluation method for classification performance it was used the accuracy, and as search strategy it was considered both forward selection and backward selection.

Data taken into account refers to personal and family anamnesis of children suffering of dyslalia, collected by the speech therapists from Regional Speech Therapy Center of Suceava via TERAPERS system. It was found that the accuracy of the model built on reduced set of features is at least as good as that obtained from original set. Forward selection method leads to a smaller number of features and to reduced execution times, so it will be used for Logo-DM implementation.

REFERENCES

[1] R. Butleris, A. Lopata, M. Ambraziunas, S. Gudas, "The Main Principles of Knowledge-Based Information Systems Engineering", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 4, pp. 99–102, 2012.

[2] C Cios, W. Pedrycz, R Swiniarski, L. Kurgan, *Data Mining A Knowledge Discovery Approach.* Spriger, Berlin, 2007, p. 606.

[3] R. Kohavi, G John, "Wrappers for feature subset selection", *Artificial Intelligence, special issue on relevance*, vol. 97, no. 1-2, pp.273–324, 1996

[4] J. Kittler, "Feature set search algorithms", *Pattern recognition and Signal Processing*, Sijhoff an Noordhoff, The Nederlands, 1978, pp. 41–60. [Online]. Available: http://dx.doi.org/10.1007/978-94-009-9941-1_3

[5] M. Danubianu, St. Gh. Pentiuc., O. Schipor, M. Nestor, I. Ungurean, D. M. Schipor, "TERAPERS - Intelligent Solution for Personalized Therapy of Speech Disorders", *International Journal on Advances in Life Science*, vol. 1, pp. 26–35, 2009.

[6] P. Kemesis, J. Ridzvanavicius, A. Stasiunas "Speech Perception Analyzer", *Elektronika ir Elektrotechnika (Electronics and Electrical Engineering),* no. 3, pp. 12–15, 1998.

[7] M. Danubianu, I. Tobolcea, S. G. Pentiuc, "Data Mining in Personalized Speech Disorders Therapy Optimization", *Knowledge-Oriented Applications in Data Mining,* Intech, Viena, 2011, pp. 321–338

[8] S. G. Pentiuc, I. Tobolcea, O.A. Schipor, M. Danubianu, D.M. Schipor, "Translation of the Speech Therapy Programs in the Logomon Assisted Therapy System", *Advances in Electrical and Computer Engineering*, vol 10, no. 2, pp. 48–52, 2010. [Online]. Available: http://dx.doi.org/10.4316/aece.2010.02008

[9] I. Mierswa, M. Wurst, R. Klingenberg, M. Scholz, T. Evler, "YALE: Rapid Prototyping for Complex Data Mining Tasks". in *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 2006, pp. 935–940 [Online]. Available: http://dx.doi.org/10.1145/1150402.1150531

[10] I. Kononenko, I. Bratko, "Information-based evaluation criterion for classifier's performance", *Machine Learning,* vol. 6, pp. 67–80, 1991. [Online]. Available: http://dx.doi.org/10.1007/BF00153760