INT J COMPUT COMMUN, ISSN 1841-9836
7(5):824-831, December, 2012.

Data Dimensionality Reduction for Data Mining: A Combined Filter-Wrapper Framework

M. Danubianu, S.G. Pentiuc, D.M. Danubianu

Mirela Danubianu, Stefan Gheorghe Pentiuc
Dragos Mircea Danubianu
"Stefan cel Mare" University of Suceava
Romania, 720229 Suceava, 1 Universitatii
E-mail: mdanub@eed.usv.ro, pentiuc@eed.usv.ro
dragosdanubianu@yahoo.com

Abstract:
Knowledge Discovery in Databases aims to extract new, interesting and potential useful patterns from large amounts of data. It is a complex process whose central point is data mining, which effectively builds models from data. Data type, quality and dimensionality are some factors which affect performance of data mining task. Since the high dimensionality of data can cause some troubles, as data overload, a possible solution could be its reduction. Sampling and filtering reduce the number of cases in a dataset, whereas features reduction can be achieved by feature selection. This paper aims to present a combined method for feature selection, where a filter based on correlation is applied on whole features set to find the relevant ones, and then, on these features a wrapper is applied in order to find the best features subset for a specified predictor. It is also presented a case study for a data set provided by TERAPERS a personalized speech therapy system.
Keywords: data mining, feature selection, filters, wrappers.

1 Introduction

As an efficient way to find new and useful knowledge in data, knowledge discovery in databases (KDD) process, and implicitly data mining as its main step, have been the subject of extensive research. The main issue relates to model building process performances, and these performances are affected by factors such as type, quality and dimensionality of available data. As most data mining techniques may not be effective for high-dimensionality data, the solution consists in its reduction. In order to reduce the number of cases, one can use sampling or filtering, whereas feature reduction may be achieved by feature selection or feature composition. Feature selection aims to identify and to remove as many irrelevant and redundant features as possible with respect to the task to be executed, and it can be made using two approaches: filters and wrappers.

Filters do not consider the effect of selected features on the performance of the whole process of knowledge discovery, since the used feature selection criterion does not require a predictor evaluation for reduced data sets. Wrappers take into account the feed-back related to the performance of the selected set of features in the KDD process. So, it is used as criteria for feature selection the predictor performance. Wrappers often give better results than filters, because feature selection is optimized for the specific learning algorithm used, but if the computational complexity and execution time are considered, they are too expensive for large dimensional datasets since each selected feature set must be evaluated with the predictive algorithm used.

In these circumstances we aim to study if an apriori filtering of features based on their relevance related to the class may improve a closed loop feature selection process. Section 2 presents some theoretically aspects related to data mining and the influence of data dimensionality on its performances. There are also enumerated some data dimensionality reduction methods. Section 3 refers some aspects regarding feature selection whereas Section 4 provides a comparison

between filters and wrappers. Section 5 presents a framework, which proposes a combination filter-wrapper and Section 6 offers some experimental results obtained by applying the proposed method over a dataset collected by TERAPERS - a computer-based speech therapy developed within the Center for Computer Research in the Stefan cel Mare University of Suceava, and used by the therapists from Regional Speech Therapy Center of Suceava from March 2008.

2 Data Mining and Data Dimensionality

Defined as the process of exploring and analyzing large volumes of data, in order to find new relationships within data or new patterns, data mining is a step in knowledge discovery in database (KDD). Its task is to analyze large volumes of data in order to extract previously unknown, interesting and potential useful patterns, and its performances are affected by factors such as: type, quality and dimensionality of data. Hypothetically, having more data, results are more precise, but practical experience with data mining algorithms has shown that this is not always true. On the one hand, the high dimensionality of data can cause data overload, and on the other hand 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. [1] This make some data mining algorithms non applicable. The solution for these problems is the reduction of data dimensions.

The size of a data set is determined both by the number of cases and by the number of features considered for each case. In order to reduce number of cases one can use sampling or filtering. 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. Sometimes, even the accuracy of built models could be improved. [2] Methods used for feature selection, can be classified as: filters or open loop methods, and wrappers or closed loop methods.

3 Feature Selection

Feature selection aims to identify and to remove as much irrelevant and redundant features as possible with respect to the task to be executed. It has the potential to be a fully automatic process, and brings some benefits for data mining, such as: an improved predictive accuracy, more compact and easily understood learned knowledge and reduced execution time for algorithms.

Feature selection methods are divided in two broad categories, filters and wrappers, and within these categories algorithms can be further individualized by the nature of their evaluation function and by the means the space of feature subsets is explored. Typically, feature selection algorithms perform a search through the space of feature subsets, and must solve four problems which affect such search:

- to select a point in the feature subsets space from which to start the search. A first choice, called forward selection, supposes to begin with no features and successively add attributes, whereas a second one, backward selection, begins with all features and successively remove them;

- even heuristic search strategies not guarantee finding the optimal subset, such strategies can give good results, and are more feasible than exhaustive search strategies which are prohibitive just for a small initial number of features;

- the most important factor which makes difference among feature selection algorithms is evaluation strategy. There are feature selection methods which operate independent of any learning algorithm, and irrelevant features are filtered before learning begins, based on general characteristics of the data to evaluate. Other methods use an induction algorithm combined with a statistical re-sample technique to estimate the final accuracy of feature subsets;

- each feature selection process must solve the problem regarding stop searching through the space of feature subsets. One might stop adding or removing features when none of the alternative improves upon the gain of current feature subset, or one might continue to alter the feature subset as long as the gain does not degrade.

4 Filters vs. Wrappers

The earliest and simplest approaches to feature selection were filters, called also open loop feature selection methods. Based on selecting features through class separability criteria, filters do not consider the effect of selected features on the performance of the whole process of knowledge discovery, as is presented in Figure 1(a). They provide usually a ranked list of features that are ordered according a specific evaluation criterion such as: accuracy and consistency of data, information content or statistical dependencies between features. [3] They give also information about the relevance of a feature compared with the relevance of other features, and do not tell to the analyst what is the desirable minimum set of the features. [4]

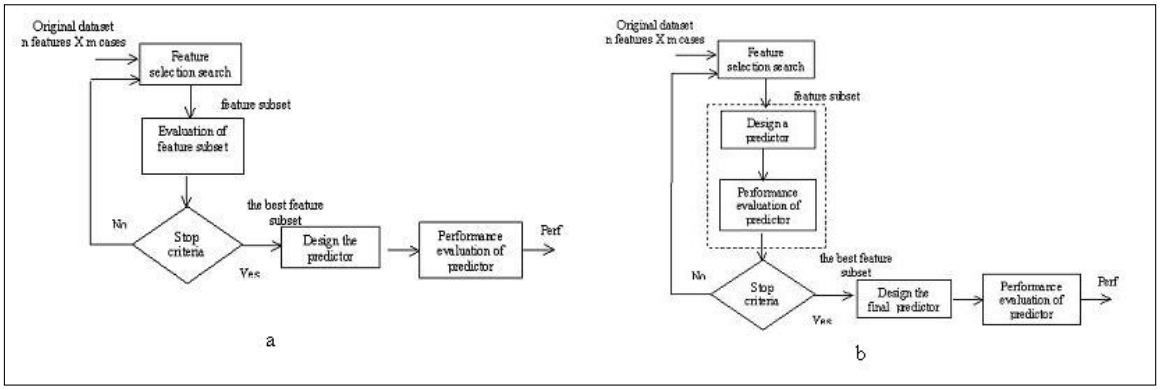

Figure 1: Open loop feature selection method (a), and closed loop feature selection method (b)

Wrappers, known also as closed loop feature selection methods take into account the feedback related to the performance of the selected set of features for the complete KDD process. They use the prediction performance as selection criteria, and evaluate the quality of selected features by comparing the performances for prediction algorithms applied on the reduced set of features and on the original one. Figure 1(b) presents a closed loop feature selection method. [2]

Regarding the final predictive accuracy of a learning algorithm, wrappers often give better results than filters, because feature selection is optimized for the specific learning algorithm used. But if the computational complexity and execution time are considered, wrappers are too expensive for large dimensional datasets, since each selected feature set must be evaluated with the predictive algorithm used. Additional, since the feature selection is closely coupled with a learning algorithm, wrappers are less general than filters and they must be run every time when one switch from one learning algorithm to another.

5 A Combined Approach Filter-Wrapper for Feature Selection

Studying the advantages and limitations for the two general feature selection methods we can conclude that, if improved performance for a specific learning algorithm is required, a filter can provide a reduced initial feature subset for a wrapper which contains only relevant features, as shown in Figure 2. This approach could produce shorter and faster search for the wrapper.

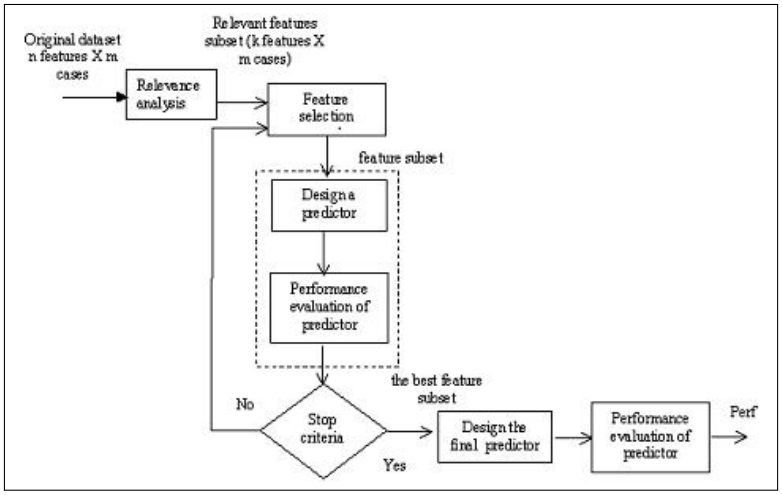

Figure 2: A framework which combines feature relevance analysis with closed loop feature selection

Since practice has demonstrated that irrelevant input features lead to great computational cost for data mining process and may cause overfitting, more feature selection researches have focused on extraction of relevant features from the whole data set in order to apply data mining algorithms upon these data. [5] [6] But how can we establish if a feature is relevant or not? In [7] is stated that features are relevant if their values vary systematically with category membership. That means that a feature is relevant if it is correlated with the class. Formally this was defined in [2] as follows:

Definition 1. A feature F_i is relevant iff there exists f_i and c for which $p(F_i = f_i) > 0$ such that

$$p(C = c|F_i = f_i) \neq p(C = c) \tag{1}$$

Relevance is usually defined in terms of correlation or mutual information. In order to define mutual information for two features we start from the concept of entropy, as a measure of uncertainty of a random variable. For a variable X the entropy is defined as:

$$E(X) = -\Sigma p(x_i) log_2(p(x_i)) \tag{2}$$

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

$$E(X|Y) = -\Sigma p(y_i) \Sigma p(x_i, y_i) log_2(p(x_i|y_i)) \tag{3}$$

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 [8] and is calculated using the following expression:

$$I(X, Y) = E(X) - E(X|Y) \tag{4}$$

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

$$p(x_i|y_j) = p(x_i, y_j)/p(y_j) \tag{5}$$

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

$$p(x_i) = \Sigma p(x_i, y_j) = \Sigma p(x_i|y_j)p(y_i) \tag{6}$$

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

$$I(X,Y) = \sum_j \sum_i p(x_i, y_j) log \frac{p(x_i, y_j)}{((p(x_i)p(y_j))} \tag{7}$$

According to this expression, one says that a feature Y is more correlated to feature X than feature Z if:

$$I(X,Y) > I(Z,Y) \tag{8}$$

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 be used symmetrical uncertainty, defined as:

$$SU(X,Y) = 2\frac{I(X,Y)}{E(X) + E(Y)} \tag{9}$$

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. Starting from these considerations, in the proposed framework, first a relevance analysis is made using the symmetrical uncertainty $SU(F_i, C)$ between each feature F_i and the class C. [9] Based on this analysis one removes the irrelevant features, and one obtains a features subset containing only the relevant features. Then, on this dataset one applies closed loop feature selection methods, using as search strategy both forward and backward selection, and using a decision tree as predictor, both for feature selection and for the performance evaluation.

6 Experimental Results

We have applied the framework described above for a real dataset collected by 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 its adaptation according to individual patients' characteristics and evolution, for which therapist must perform complex examination of children, materialized in recording of relevant data relating to personal and family anamnesis. These collected data may provide information relative to various causes that may negatively influence the normal development of the language. Further, one provide to the personalized therapy programs data such as number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. The tracking of child progress materializes data which indicate the moment of assessing the child and his status at that time. All these data are stored in a relational database, composed of 60 tables.

Data stored in the TERAPERS database is the set of raw data that can be the subject of data mining process. It might be useful, because as it was shown in [10] one can use classification in order to places the people with different speech impairments in predefined classes (if attribute

diagnosis contain the class label one can predict a diagnosis based on information contained in various predictor variables), one can use clustering to group people with speech disorders on the basis of similarity of different features and to help therapists to understand who are they patients, or one can use association rules to determine why a specific therapy program has been successful on a segment of patients with speech disorders and on the other was ineffective. For our experiments one considers a data set consisting of 72 features with numeric and descriptive values and 312 cases. These are anamnesis data or data derived from complex examination on which one intend to build a classification model to predict, in order to suggest to therapist the diagnosis for future cases. On this data set one have applied the feature selection method described above. Shown in Figure 3, such experiment is designed and implemented in WEKA. [11]

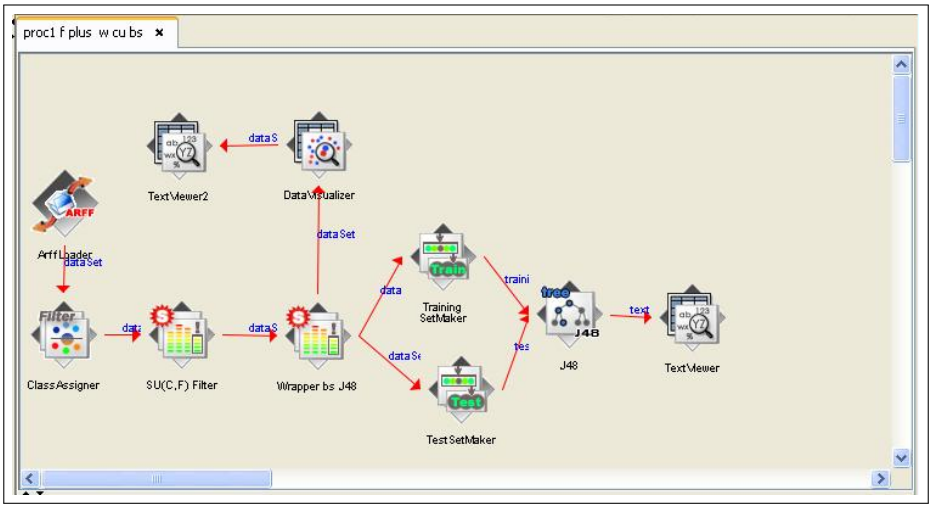

Figure 3: WEKA knowledge flow for the proposed framework

One has considered the attribute *diagnosis* as class label, and we have built a feature selection process in two steps.In the first stage one have applied a unsupervised filter against the whole set of features. Based on values for symmetrical uncertainty $SU(F_i, C)$ there were retained 52 relevant features. In the second stage over this feature subset one have applied a wrapper which uses as estimation predictor a decision tree classifier (J48). For this wrapper one applies alternatively the two search strategies forward and backward search.

To analyze the influence of data reduction we need to know what we gain or what we lose so, we must compare computing times and the accuracy for the model built for reduced data sets obtained using the described approaches. Figure 4 presents the performances of the classifiers built on feature subset produced by wrapper using a backward selection strategy and on features obtained from the same wrapper which use, this time, a forward selection strategy.

As one can see in Figure 4 for the subset obtained by backward selection, the predictions' performance, measured in correctly classified instances, are better than those for the feature subset obtained by forward selection. In Figure 5 a comparison of the execution time for forward and backward selection in both cases, with filter before wrapper, and without filter before wrapper is made. One see that for backward selection, which provides better performance for prediction, the process which considers as input for wrapper the data subset provided by filter is three times faster than the other process.

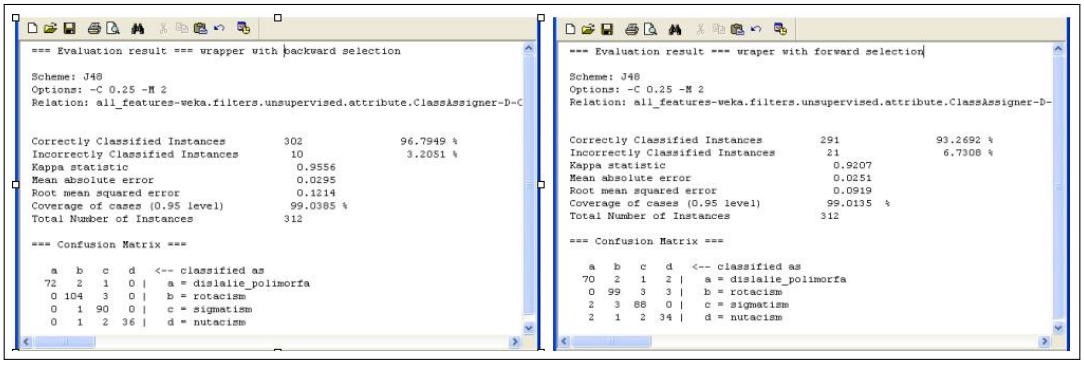

Figure 4: Performances of classifier built on feature subsets obtained by proposed method

	Execution time for (sec):	
	forward selection strategy	backward selection strategy
without filter before wrapper	23	3208
with filter before wrapper	58	1110

Figure 5: Performances of classifier built on feature subsets obtained by proposed method

7 Conclusions and Future Works

As a possibility to reduce the number of feature considered by data mining algorithms, in order to make them more efficient, this paper presents a method which uses a combination filter-wrapper. We have used a correlation based filter on the whole set of features, then on relevant subset of features we have applied a wrapper which uses a decision tree classifier for prediction. As a case study we have applied this method on data collected by TERAPERS a system which aims to assist speech therapists on personalized therapy of dyslalia. The process was designed and implemented in WEKA. We have compared the performances obtained both for feature selection by the described method, and for feature selection using only the same wrapper as in first case. We have achieved clearly superior performances for execution time, when we have used for feature selection the combined approach and backward selection as search strategy for wrapper. The positive results obtained for the considered data encourage us to continue our work. We will try to improve these execution times by parallelization of feature selection operations.

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.

Bibliography

[1] Danubianu M., Pentiuc S.G., Tobolcea I., Schipor O.A., Advanced Information Technology - Support of Improved Personalized Therapy of Speech Disorders, *INT J COMPUT COMMUN*, ISSN 1841-9836, 5(5): 684-692, 2010.

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

[3] Hall, M., Correlation-based feature selection for discrete and numeric class machine learning, *Proc. of International Conference on Machine Learning*, 359-365, Morgan Kaufmann, 2000.

[4] Douik A., Abdellaoui M., Cereal Grain Classification by Optimal Features and Intelligent Classifiers, *INT J COMPUT COMMUN*, ISSN 1841-9836, 5(4):506-516, 2010.

[5] Peng H. Long F., Ding C., Feature Selection based on mutual Information: Criteria of Max-Dependency, Max-Relevance and Min-Redundancy, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(8):1226 - 1238, 2005.

[6] John G.H., Kohavi R., Pfleger P., Irrelevant features and the subset selection problem, *Machine Learning: Proceedings of the Eleventh International Conference*, 121-129, Morgan Kaufman, 1994.

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

[8] Quinlan J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufman, 1993.

[9] Yu, L., Liu, H., Efficient Feature Selection via Analysis of Relevance and Redundancy, *Journal of Machine Learning Research*, 5:1205-1224, 2005 .

[10] Danubianu M., Pentiuc St. Gh., Socaciu T., 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, 2009.

[11] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, 11(1):10-18, 2009.