Pattern Recognition Letters 69 (2016) 28–34

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A hybrid feature selection method based on instance learning and cooperative subset search☆

Afef Ben Brahim[a,*], Mohamed Limam[a,b]

[a] LARODEC, ISG, University of Tunis, Tunisia
[b] Dhofar University, Oman

ARTICLE INFO

Article history:
Received 31 January 2015
Available online 23 October 2015

Keywords:
Feature selection
Hybrid
Small sample size
Classification
Stability

ABSTRACT

The problem of selecting the most useful features from thousands of candidates in a low sample size data set arises in many areas of modern sciences. Feature subset selection is a key problem in such data mining classification tasks. In practice, it is very common to use filter methods. However, they ignore the correlations between genes which are prevalent in gene expression data. On the other hand, standard wrapper algorithms cannot be applied because of their complexity. Additionally, existing methods are not specially conceived to handle the small sample size of the data which is one of the main causes of feature selection instability. In order to deal with these issues, we propose a new hybrid, filter wrapper, approach based on instance learning. Its main challenge is that it converts the problem of the small sample size to a tool that allows choosing only a few subsets of features in a filter step. A cooperative subset search, CSS, is then proposed with a classifier algorithm to represent an evaluation system of wrappers. Our method is experimentally tested and compared with state-of-the-art algorithms based on several high-dimensional low sample size cancer datasets. Results show that our proposed approach outperforms other methods in terms of accuracy and stability of the selected subset.

1. Introduction

Today, high-throughput biotechnologies such as microarray and sequence methods can easily measure expression levels of thousands of genes. It is the case in cancer classification problems where a predictive model is built on the training data consisting of patients belonging to healthy or cancerous categories. The classification algorithm finds the relationship between the features which are gene expression profiles and the two class labels [1] . As there are thousands of gene expressions and only a few samples in a typical gene expression data set, serious problems occur with the application of many traditional statistical methods. Overfitting on the classifier is one of these problems. It leads to very good and often perfect classification performance on the training data, but this perfect performance does not translate to new unlabeled data resulting in a very limited classifier generalization. Kohane et al. [2] explained that the small sample size in genomic applications may be due to the high cost of the microarrays. Each sample involves the measurements of tens of thousands of variables corresponding to the expression of tens of thousands of genes measurable with microarray technology. The result is a large number of features compared to the number of samples. Kohane et al. [2] described this system as highly underdetermined system and explained that based on the relatively small number of observations, there is a large number of solutions in which the genes being measured could interact. Thus, due to the underdetermined nature of these systems, standard machine learning techniques do not hold up well because those techniques were developed under the assumption that the number of samples, m, is much larger than the features dimensionality d. Since it is difficult to increase m for the reasons explained above, dimension reduction is a solution for such problem. Reducing the number of genes will reduce the algorithm's variance. So, machine learning and more specially feature selection methods are useful to deal with high dimensional data sets.

In practice, for such high dimensional data, it is very common to use filter methods that measure the strength of relationship between each gene and the class label. However, Tolosi and Lengauer [3] demonstrated that filters ignore the correlations between genes, which are prevalent in gene expression data due to gene co-regulation. The consequence is that many redundant differentiated genes are included, meanwhile, useful but weakly differentiated genes may be omitted. On the other hand, Kohavi and John [4] showed that standard wrapper algorithms cannot be applied because of their high computational complexity due to the need to train a large number of classifiers. With tens of thousands of features, which

☆ This paper has been recommended for acceptance by A. Petrosino.

* Corresponding author. Tel.: +216 53824337.
E-mail address: afef.benbrahim@yahoo.fr (A. Ben Brahim).

http://dx.doi.org/10.1016/j.patrec.2015.10.005
0167-8655/

is the case in gene expression microarray data analysis, a hybrid approach could be adopted. It should follow a filter model in the search step selecting small number of candidate subsets of features. Then, a wrapper method is applied to the reduced subsets to achieve the best possible performance with a particular learning algorithm. Accordingly, the hybrid model is expected to be more efficient than filter and less expensive than wrapper.

By another hand, a great variety of feature selection methods have been developed with a focus on improving the predictive accuracy of learning models while reducing dimensionality but most of existing methods do not take into account the small sample size problem in their design. Nevertheless, learning in the small sample case is of practical interest. One reason for this is the difficulty in collecting data for each object. Yet, this data specificity produces some problems, not only for predictive performance of learning algorithms, but also results in the instability of feature selection results.

To deal with all these issues, we propose a hybrid feature selection approach which is specially designed for this type of data, in order to benefit from its small sample size. Our proposed method uses an instance based candidate feature subset selection in a filter step. The key idea is to decompose an arbitrarily complex problem into a set of locally ones through local learning of feature relevance, and then find relevant features globally. Each instance proposes a candidate subset of the most relevant features for this instance. Small sample size makes this process feasible with acceptable running time. Thus the high dimensionality of data is reduced to few subsets of features which number corresponds to the data sample size and this is when small sample size is of benefit to feature selection process. The candidate feature subsets are then integrated in a search procedure of the optimal feature subset, where CSS is used with a classifier algorithm as evaluation systems of wrappers. The main goal of our proposed methods is to reduce data set dimensionality while obtaining good performance in terms of accuracy, stability of feature selection and size of the obtained subset.

The remainder of the paper is organized as follows. Section 2 discusses the basic concepts of feature selection and their representative methods. In Section 3, we present our proposed hybrid feature selection approach. In Section 4 we evaluate the performance of our method with seven feature selection techniques based on seven high dimensional data sets and one large scale data set. We finally conclude this paper in Section 5.

2. Basic concepts of feature selection

In most common feature selection techniques, an evaluation function is used to assign scores to subsets of features and a search algorithm is used to search for a subset with a high score. The evaluation function can be based on some general relevance measure of the features to the prediction (filter model) or on the performance of a specific predictor (wrapper model). A third category of algorithms that fuse filters and wrappers are known as hybrid methods.

2.1. Filters

Filters [5] are not dependent of a specific type of predictive model, they only take characteristics of the data into consideration to select a best feature subset or to obtain a feature's ranking by assigning a score to each feature. This is done before the learning process begins. Filter methods are very fast and thus very useful to select features in high dimensional data sets. Several filter methods have been proposed in the literature and have shown their effectiveness on selecting the most relevant features and improving the predictive performance. Some of the most popular filter methods are described in the following.

t-test filter: The statistical *t*-test is commonly used for feature selection. The *t*-test is used in the form that defines the score of a feature as the ratio of the difference between its mean values for each of the two classes and the standard deviation. The weight of each feature is thus given by its computed absolute score.

Minimum-Redundancy-Maximum-Relevance: The minimum Redundancy Maximum Relevance (mRMR) method proposed by Peng et al. [6] is a mutual information based method. It selects features according to the maximal statistical dependency criterion. The mRMR method selects a feature subset that has the highest relevance with the target class, subject to the constraint that selected features are mutually as dissimilar to each other as possible.

Relief: This method was proposed by Kira and Rendell [7]. Relief is based on instance learning, it selects features to separate instances from different classes. Robnik and Kononenko [8] related the relevance evaluation criterion of Relief to the hypothesis of margin maximization, which explains why the algorithm provides superior performance in many applications.

Filters can be used as a preprocessing step to reduce space dimensionality and overcome overtting [4,5]. When the number of features becomes very large, the filter model is usually chosen as it is computationally efficient, fast and independent of the classification algorithm. However, taking into account the predictive performance of a learning algorithm while selecting features could be of a big interest, since enhancing this performance is one of the main objectives of feature selection. Filters ignore this aspect and this is their major shortcoming.

2.2. Wrappers

It is of high interest that the search for the optimal feature subset takes into account the specific biases and performance of the predictive algorithm. Based on this, wrapper models use a specific classifier to evaluate the quality of selected features [4]. The performance measure of a learning algorithm along with a statistical re-sampling technique such as cross validation [9] is used to select the best feature subset. Given a predefined classifier, a typical wrapper model iteratively produces a set of features based on a searching procedure, then evaluates features using the performance of a classifier until a feature set with the desired quality is reached. A wide range of search strategies can be used and are described in the following.

Sequential feature selection: It is one of the most widely used wrapper techniques [4,10]. It selects a subset of features by forward or backward search, which consist on respectively adding or removing features until certain stopping conditions are satisfied.

Randomized feature selection: Randomized wrapper algorithms search the next feature subset at random [11]. Single features or several features can be added at once, removed, or replaced from the previous feature set based on the effect on the predictive performance. With these updates, the current set moves to the subset with the highest accuracy. The search procedure terminates when no subset improves over the current set.

Support vector machines and recursive feature elimination: While the search and the evaluation procedures are separated in the previous wrapper methods, there exist methods that use an embedded model where the search for an optimal subset of features is built into the classifier construction using its internal parameters. Guyon et al. [12] introduced a feature selection method using the weight vectors of a support vector machine (SVM) [13] in combination with recursive feature elimination (RFE) to form SVM.RFE. The ranking criterion is computed for all features based on their corresponding weights. This process is iterated and the features with the smallest rankings, i.e. weights, are removed. The remaining features are selected. This iterative procedure is a backward feature elimination [4]. The algorithm can be accelerated by removing more than one feature.

Wrappers usually provide the best performing feature set for a particular type of model and have the ability to take into account feature dependencies as they consider groups of features jointly.