On the Use of Variable Complementarity for Feature Selection in Cancer Classification

Patrick E. Meyer and Gianluca Bontempi

Université Libre de Bruxelles, (CP 212), 1050 Bruxelles, Belgique
{pmeyer, gbonte}@ulb.ac.be
http://ulb.ac.be/di/mlg/

Abstract. The paper presents an original filter approach for effective feature selection in classification tasks with a very large number of input variables. The approach is based on the use of a new information theoretic selection criterion: the double input symmetrical relevance (DISR). The rationale of the criterion is that a set of variables can return an information on the output class that is higher than the sum of the informations of each variable taken individually. This property will be made explicit by defining the measure of *variable complementarity*. A feature selection filter based on the DISR criterion is compared in theoretical and experimental terms to recently proposed information theoretic criteria. Experimental results on a set of eleven microarray classification tasks show that the proposed technique is competitive with existing filter selection methods.

1 Introduction

Statisticians and data-miners are used to build predictive models and infer dependencies between variables on the basis of observed data. However, in a lot of emerging domains, like bioinformatics, they are facing datasets characterized by a very large number of features (up to several thousands), a large amount of noise, non-linear dependencies and, often, only several hundreds of samples. In this context, the detection of functional relationships as well as the design of effective classifiers appears to be a major challenge. Recent technological advances, like microarray technology, have made it possible to simultaneously interrogate thousands of genes in a biological specimen. It follows that two classification problems commonly encountered in bioinformatics are how to distinguish between tumor classes and how to predict the effects of medical treatments on the basis of microarray gene expression profiles. If we formalize this prediction task as a supervised classification problem, we realize that we are facing a problem where the number of input variables, represented by the number of genes, is huge (around several thousands) and the number of samples, represented by the clinical trials, is very limited (around several tens). Because of well-known numerical and statistical accuracy issues, it is typically necessary to reduce the number of variables before starting a learning procedure. Furthermore, selecting features (i.e. genes) can increase the intelligibility of a model while at the

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 91–102, 2006.

same time decreasing measurements and storage requirements [1]. A number of experimental studies [2, 3, 4] have shown that irrelevant and redundant features can dramatically reduce the predictive accuracy of models builded from data.

Feature selection is a topic of machine learning whose objective is selecting, among a set of input variables, the ones that will lead to the best predictive model. Two well-known approaches in feature selection combine a search strategy with a stochastic evaluation function: the *filter approach* and the *wrapper approach* (see [3, 2]). In the wrapper approach, the evaluation function is the validation outcome (e.g. by leave-one-out) of the learning algorithm itself. In the filter approach, examples of evaluation functions are probabilistic distance, interclass distance, information theoretic or probabilistic dependence measures. These measures are often considered as intrinsic properties of the data, because they are calculated directly on the raw data instead of requiring a learning model that smoothes distributions or reduces the noise.

This paper will focus on the use of filter techniques for feature selection in supervised classification tasks. In particular, we present an original filter approach based on a new information theoretic selection criterion, called the *double input symmetrical relevance* (DISR). This criterion combines two well known intuitions of feature selection: first, a combination of variables can return more information on the output class than the sum of the information returned by each of the variables taken individually. This property will be made explicit by defining the notion of *variable complementarity*. Secondly, in absence of any further knowledge on how subsets of d variables should combine, it is intuitive to assume a combination of the best performing subsets of $d-1$ variables as the most promising set. This intuition will be made formal by the computation of a lower-bound on the information of a subset of variables expressed as the average of information of all its sub-subsets.

The DISR criterion can be used to select among a finite number of alternative subsets the one expected to return the maximum amount of information on the output class. As we intend to benchmark its performance with respect to state-of-the-art information theoretic criteria we define an experimental session where several filter algorithms with different selection criteria but the same search strategy are compared. In our experiments we compare the filter based on DISR with four state-of the art approaches: a Ranking algorithm [5] and three filters based on the same search strategy: the forward selection. The three state-of-the-art criteria are the Relevance criterion [6], the Minimum Redudancy Maximum Relevance criterion [7] and the Conditional Mutual Information Maximization criterion [8]. The assessment of the different filters is obtained by measuring the classification accuracy of several learning algorithms which adopt as inputs the set of variables returned by each of the filter methods. For our benchmark purposes, we use eleven public-domain multi-class microarray gene expression datasets. The experimental results show that the proposed technique is competitive with existing filter selection methods.