Feature Ranking Using
an EDA-based Wrapper Approach

Yvan Saeys, Sven Degroeve, and Yves Van de Peer

Department of Plant Systems Biology, Ghent University, Flanders Interuniversity Institute for
Biotechnology (VIB), Technologiepark 927, B-9052 Ghent, Belgium
{yvan.saeys,sven.degroeve,yves.vandepeer}@psb.ugent.be

Summary. Feature subset selection is an important pre-processing step for classification. A
more general framework of feature selection is feature ranking. A feature ranking provides an
ordered list of the features, sorted according to their relevance. Using such a ranking provides
a better overview of the feature elimination process, and allows the human expert to gain
more insight into the processes underlying the data. In this chapter, we describe a technique to
derive a feature ranking directly from the estimated distribution of an EDA. As an example, we
apply the method to the biological problem of acceptor splice site prediction, demonstrating
the advantages for knowledge discovery in biological datasets with many features.

1 Introduction

Reduction of data dimensionality has become an apparent need in machine learn-
ing during the past decades. Examples of large datasets with instances described by
many features include problems in image processing, text mining and bioinformat-
ics. To efficiently deal with such data, dimension reduction techniques emerged as a
useful pre-processing step in the flow of data analysis. A subset of these techniques
is referred to as feature (subset) selection techniques. These techniques differ from
other reduction techniques (like projection and compression techniques) in that they
do not transform the original input features, but merely select a subset of them.

The reduction of data dimensionality has a number of advantages: attaining
good or even better classification performance with a restricted subset of features,
 faster and more cost-effective predictors, and the ability to get a better insight in the
processes described by the data. An overview of feature selection techniques can be
found in [10] and [5].

Techniques for feature selection are traditionally divided into two classes: fil-
ter approaches and wrapper approaches [12]. Filter approaches usually compute a
feature relevance score such as the feature-class entropy, and remove low-scoring
features. As such, these methods only look at the intrinsic properties of the dataset,
providing a mechanism that is independent of the classification algorithm to be used
afterwards. In the wrapper approach, various subsets of features are generated, and
evaluated using a specific classification model. A heuristic search through the space of all subsets is then conducted, using the classification performance of the model as a guidance to find promising subsets. In addition to filter and wrapper approaches, a third class of feature selection methods can be distinguished: embedded feature selection techniques [1]. In embedded methods, the feature selection mechanism is built into the classification model, making direct use of the parameters of the induction model to include or reject features. Examples of these methods are the pruning of decision trees, and recursive feature elimination (RFE) using the weight vector of a linear Support Vector Machine [6].

In this chapter, we will focus on the wrapper approach for feature selection. Wrapper based methods combine a specific classification model with a strategy to search the space of all feature subsets. Commonly used search strategies are sequential forward or backward selection [11], and stochastic iterative sampling methods like genetic algorithms (GA, [13]) or estimation of distribution algorithms (EDA, [14]). The EDA approach to feature selection is shown in Fig. 1. In this case, each individual in the population represents a feature subset, coded as a binary string. Each bit represents a feature, a 1 indicating the presence, a 0 the absence of a particular feature. Individuals are evaluated (step 2, Fig. 1) by training a classification model with the features present in the individual (i.e. the ones having a 1), and afterwards validating it, either by cross-validation on the training set, or by using a separate training and holdout set. The feature subset returned by the algorithm is then the best subset found during the search.

Instead of using the traditional crossover and mutation operators, inherent to GA, an EDA explicitly constructs a model of the selected feature subsets (step 4). Depending on the complexity of the model, univariate, bivariate or multivariate interactions between the encoded features are modelled. In a subsequent step (step 5), the new population is created by sampling feature subsets from this model. The new population can either be completely sampled from the distribution, or can partly consist of sampled subsets and subsets retained from the previous population (elitists). The use of EDAs for feature subset selection was pioneered in [8] and the use of EDAs for FSS in large scale domains, was reported to yield good results [9, 22].

2 EDA-based Feature Ranking

2.1 Feature Ranking

As mentioned in the introduction, the standard approach to using EDA for feature subset selection (FSS), is to select the best feature subset encountered in the iterative process as the final solution. However, selecting the single best subset of features provides a rather static view of the whole elimination process. When using FSS to gain more insight in the underlying processes, the human expert has no idea of the context of the specific subset. Questions about how much and which features can still be eliminated before the classification performance drastically drops down provide