Chapter 5
Fast-Ensembles of Minimum Redundancy Feature Selection

Benjamin Schowe and Katharina Morik

Abstract. Finding relevant subspaces in very high-dimensional data is a challenging task not only for microarray data. The selection of features is to enhance the classification performance, but on the other hand the feature selection must be stable, i.e., the set of features selected should not change when using different subsets of a population. Ensemble methods have succeeded in the increase of stability and classification accuracy. However, their runtime prevents them from scaling up to real-world applications. We propose two methods which enhance correlation-based feature selection such that the stability of feature selection comes with little or even no extra runtime. We show the efficiency of the algorithms analytically and empirically on a wide range of datasets.

5.1 Introduction

The growing dimensionality of recorded data, especially in bioinformatics, demands dimension reduction methods that identify small sets of features leading to a better learning performance. Along with the high dimensionality come small sample size and high variance, which makes it hard to find adequate feature subsets without being kept in local optima. The large number of features challenges the runtime of a selection algorithm. Hence, the main quality criteria are that the algorithm is

- **multivariate** – It takes into account inter-feature-dependencies;
- **stable** – It does not vary much for unseen data of the population;
- **amending learning** – The learning performance is enhanced;
- **fast** – It scales well for very large numbers of features.

Ensemble methods increase stability and, hence, are frequently used in feature selection. However, they usually slow down the procedure. We will present two
methods which speed up ensembles in a simple and effective way. A careful evaluation on several real world and simulated data sets investigates the quality of our new methods.

5.2 Related Work

Fast univariate filter approaches, like the t-test [5] or SAM-statistics [17], compute a scoring function on the features, disregarding feature interplay. Wrapper approaches [13] better solve this problem, at the cost of much longer runtime. Each feature set evaluation demands a cross-validated training of the used learning algorithm. Some learning algorithms provide the user with an implicit feature ranking which can easily be exploited for feature selection. Such embedded approaches are using the weight vector of a linear SVM [18] or the frequency of feature use of a Random Forest (RF) [3]. They are aware of feature interplay and faster than wrappers but biased towards the learning algorithm used.

A group of new algorithms has come up to bridge the gap between fast but univariate filters on the one hand, and slow but multivariate wrappers on the other hand. Their goal is to find a subset of features which is highly predictive with no or a minimum of redundant information. The correlation based feature selection (CFS) [8] performs a sequential forward search with a correlation measure in the evaluation step. CFS iteratively adds the feature which has the best ratio between predictive relevance of the feature and its correlation with the already selected features. Both, predictiveness and correlation, are measured by the entropy-based symmetrical uncertainty

$$SU(f_i, f_j) = \frac{2IG(f_i|f_j)}{H(f_i) + H(f_j)}, \quad (5.1)$$

where the information gain $IG$ of feature $f_i$ w.r.t. feature $f_j$ is divided by the sum of the entropies of $f_i$ and $f_j$. Since CFS uses symmetrical uncertainty, it is only suitable for discrete values.

Ding and Peng [4] reinvented CFS with the capability for handling numerical variables calling it Minimum Redundancy Maximum Relevance FS (mRMR). For numerical features the F-test is used. It reflects the ratio of the variance between classes and the average variance inside these classes. For a continuous feature $X$ and a nominal class variable $Y$ with $C$ classes, both from a data set with $n$ examples it is defined as

$$F(X, Y) = \frac{(n - C) \sum_c n_c (\bar{X}_c - \bar{X})^2}{(C - 1) \sum_c (n_c - 1) \sigma^2_c} \quad (5.2)$$

with per-class-variance $\sigma^2_c$ and $n_c$ the number of examples in class $c \in \{1, \ldots, C\}$. The redundancy of a numerical feature set is measured by the absolute value of Pearson’s correlation coefficient