Available online at www.sciencedirect.com

ScienceDirect

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

ELSEVIER

Original Research Article

# Classification methods for high-dimensional genetic data

*Jan Kalina* *

*Institute of Computer Science of the Academy of Sciences of the Czech Republic, Department of Medical Informatics and Biostatistics, Prague, Czech Republic*

ARTICLE INFO

*Article history:*
Received 5 March 2013
Accepted 19 September 2013
Available online 31 December 2013

*Keywords:*
Multivariate statistics
Classification analysis
Shrinkage estimation
Dimension reduction
Data mining

ABSTRACT

Standard methods of multivariate statistics fail in the analysis of high-dimensional data. This paper gives an overview of recent classification methods proposed for the analysis of high-dimensional data, especially in the context of molecular genetics. We discuss methods of both biostatistics and data mining based on various background, explain their principles, and compare their advantages and limitations. We also include dimension reduction methods tailor-made for classification analysis and also such classification methods which reduce the dimension of the computation intrinsically. A common feature of numerous classification methods is the shrinkage estimation principle, which has obtained a recent intensive attention in high-dimensional applications.



## 1. Introduction

Classification analysis is a task of constructing (learning) a decision rule based on the data (training data set), which is able to automatically assign new data to one of two or several groups. Unfortunately, standard classification methods are known to be unsuitable for high-dimensional data with a number of variables exceeding the number of observations. In the analysis of molecular genetic data, classification procedures have to face serious challenges. Common classification procedures of multivariate statistics as well as data mining suffer from the curse of dimensionality. Other important issues include combining data of different types (continuous or categorical variables, signals, or images), combining classifiers, robustness, missing data, or sensitivity to the normalization of the data. High-dimensional information extraction tools have been developed in image analysis or chemometrics [1] and in recent years have penetrated to applications in molecular genetics, e.g. on high-throughput gene expression data. This paper presents an overview of a variety of new classification methods, which have recently been introduced for the analysis of molecular genetic data.

Examples of classification tasks in molecular genetics include the diagnostic decision support in patients based on gene expression measurements, the problem to find a small number of genes which contribute the most to the diagnosis (so-called gene filtering), or the task to find clusters of samples (patients) or clusters of variables (genes) in a clinical study. This paper fills the gap identified in [2] as a need for an overview of data analysis tools for high-dimensional data.

In general, we distinguish between supervised and unsupervised classification methods. Let us assume that each of the multivariate observations belongs to one of $K$ given groups. Supervised classification methods exploit the information about the group membership of each observation of the

* Corresponding author at: Institute of Computer Science AS CR, Pod Vodárenskou věží 2, CZ-182 07 Praha 8, Czech Republic.
E-mail address: kalina@cs.cas.cz (J. Kalina).

0208-5216/$ – see front matter 
http://dx.doi.org/10.1016/j.bbe.2013.09.007

training data set and their aim is to describe the variability among individuals groups, which are fixed and known for the training data set.

Unsupervised methods ignore the information about the group membership of individual samples from the training set. All observations of the training set are considered jointly and the methods aim at dividing the set of observations to certain natural clusters. As an example let us assume that a researcher has the information that the data belong to three different groups. An unsupervised classification method does use this information, but the researcher can require the method explicitly to divide the data e.g. into three clusters. Although the cluster analysis is extremely popular in the analysis of gene expression data, its results may be blurred by the presence of a large number of irrelevant genes.

Regularization can be described as solving ill-posed or insoluble high-dimensional problems by means of additional information, assumptions, or penalization [3]. We understand shrinkage methods as an important class of regularized approaches for analyzing high-dimensional data, especially in the classification task. Shrinkage methods go back to C. Stein [4] as a linear combination (weighted mean) of a usual (unbiased, naïve, raw) estimator and a structured (biased) estimator, which can be obtained under additional special assumptions or is equal to a simplification of the real situation (e.g. a constant or diagonal matrix). The parameter for weighting both estimators is obtained by a cross-validation, while an analytical expression of the optimal amount of shrinkage is available in some cases. In regression analysis, the concept of shrinkage regression is used to describe estimation methods suitable under multicollinearity [5]. Shrinkage approaches are particularly appropriate for genetic data with a vast majority of genes with a very low signal-to-noise ratio [6].

Robustness properties of shrinkage approaches have been observed empirically (e.g. [7]), although the parallel between shrinkage estimation and the theory of robustness has not been theoretically examined. D.L. Donoho and I.M. Johnstone [8] applied a shrinkage principle to wavelet coefficients in a general context of curve estimation (including image analysis). Their shrinkage wavelets are obtained by pulling each wavelet coefficient towards zero by at least the noise level. Additionally, wavelet coefficients with an absolute value below a given threshold are cut away. This can be interpreted as a robust approach to reducing noise.

This paper has the following structure. Section 2 describes dimension reduction tools tailor-made for classification problems. Section 3 is devoted to modifications of the linear discriminant analysis for high-dimensional data. Section 4 discusses regression methods, which are commonly used to solve high-dimensional classification problems. Section 5 is devoted to classification methods based on hypothesis testing. Section 6 studies recent classification methods which have been proposed in the context of data mining.

## 2. Dimension reduction in classification problems

Dimension reduction is commonly used as a preliminary step of the information extraction from high-dimensional data simplifying consequent computations. Particularly in the context of classification analysis, dimension reduction methods serve also to describe differences among groups, reveal the dimensionality of the separation among groups, and express the contribution of individual variables to this separation.

Dimension reduction methods in molecular genetics can be distinguished to two groups:

- Variable selection (gene selection)
- Feature extraction

Common gene selection methods are based e.g. on information theory or hypothesis testing. The latter has the aim is to rank the genes in order of evidence against the null hypothesis rather than to assign $p$-values to genes. However, the major drawback of common variable selection methods is a lack of stability as analyzed in [9]. On the other hand, feature extraction tools search for combinations of genes. The most common feature extraction tool is the principal component analysis (PCA). It has been criticized as ineffective for data observed in two or several groups [10]. Because the PCA does not exploit the information about the class membership of each sample, there have been new dimension reduction methods proposed, which are tailor-made for classification purposes.

The maximal relevance method is a general dimension reduction approach for variable selection based on searching for such genes, which have the largest relevance to the response. If the task is a classification to two groups, the response is an indicator variables (group 1 vs. group 2) and the relevance is commonly measured by means of a mutual information or correlation coefficient. The low classification performance of the method has been criticized [11], while the problem may actually consist in classification methods, which have only a weak performance for genes with a high redundancy.

X. Liu et al. [11] obtained significantly better results with the minimum redundancy and maximum relevance (MRMR) method, which selects genes related to the discrimination among groups while at the same time reducing the redundancy among the genes. The method selects genes with the strongest ability for predicting the differences in gene expression patterns in different classes. Therefore it is suitable for a consequent classification analysis, although it is not incorporated directly to the dimension reduction process. The computation is based on comparisons of different gene sets in terms of their relevance and redundancy.

K. Vanden Branden and M. Hubert [12] proposed a procedure called robust soft independent modelling of class analogies (RSIMCA). It starts with the computation of the robust principal component analysis for high dimensional data (ROBPCA), which is performed on each group separately. This summarizes the data in each group in a different dimension. A new observation is classified by means of its deviations to the different robust PCA models, exploiting a robust Mahalanobis distance. The robustness is ensured by the minimum covariance determinant estimator, which can be alternatively replaced by its weighted counterpart [13].

V. Zuber and K. Strimmer [14] proposed a CAR (correlation-adjusted correlation) score, as a criterion for variable ranking