Cancerclass: An R package for development and validation of diagnostic tests from high-dimensional molecular data

Jan Budczies, Daniel Kosztyla

May 6, 2012

Contents

1 Introduction

Progress in molecular high-throughput techniques has led to the opportunity of simultaneous monitoring of hundreds or thousands of biomolecules in medical samples, e.g. using microarrays. In the era of personalized medicine, these data form the basis for the development of prognostic and predictive tests. Because of the high dimensionality of the data and connected to the multiple testing problem, the development of molecular tests is sensitive to model overfitting and performance overestimation. Bioinformatic methods have been developed to cope with these problems, e.g. the multiple random validation protocol that was presented in [1].

Cancerclass integrates methods for development and validation of diagnostic tests from high-dimensional molecular data. In the past, simple classifiers were shown to have a good performance on high-dimensional data compared to more sophisticated methods [2]. Therefore, the protcol of **cancerclass** uses simple classification methods, while much attention is payed to validation and visualization of classification results. In short, the protocol starts with feature selection by a filtering step. Then, a predictor is constructed using the nearest-centroid method. The accuracy of the predictor can be evaluated using training and test set validation, leave-one-out cross-validation or in a multiple random validation protocol. Methods for calculation and visualization of continuous prediction score allow to balance sensitivity and specificity and define a cutoff value according to clinical requirements.

In the following, the functionality of **cancerclass** is illustrated using two sets of cancer gene expression data. A gene expression data set of two types of leukemia (AML and AML) [3] is delivered with **cancerclass**. Gene expression data of breast cancer with good and poor prognosis [4, 5] are obtained from the ExperimentData package **cancerdata**.

2 Multiple random validation protocol

First, the package **cancerclass** and an example data set are loaded. GOLUB1 is a gene filtered version of gene expression data from 72 leukemia patients [3, 1].

```
> library(cancerclass)
> data(GOLUB1)
> GOLUB1

ExpressionSet (storageMode: lockedEnvironment)
assayData: 3571 features, 72 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: X1, X2, ..., X72  (72 total)
  varLabels and varMetadata description:
    class: NA
    sample: NA
    ...: ...
    gender: NA
    (5 total)
featureData
  featureNames: AFFX-BioDn-3_at, AFFX-BioB-5_st, ..., M71243_f_at  (3571 total)
  fvarLabels and fvarMetadata description:
    symbol: NA
    description: NA
experimentData: use 'experimentData(object)'
Annotation: hu6800
```

Using a protocol similar to [1] we investigate the dependence of classification accuracy on the number of features (Fig. 1):

```
> nval <- nvalidate(GOLUB1[1:200, ], ngenes = c(5, 10, 20, 50,
+     100, 200))
```

```
> plot(nval)
```

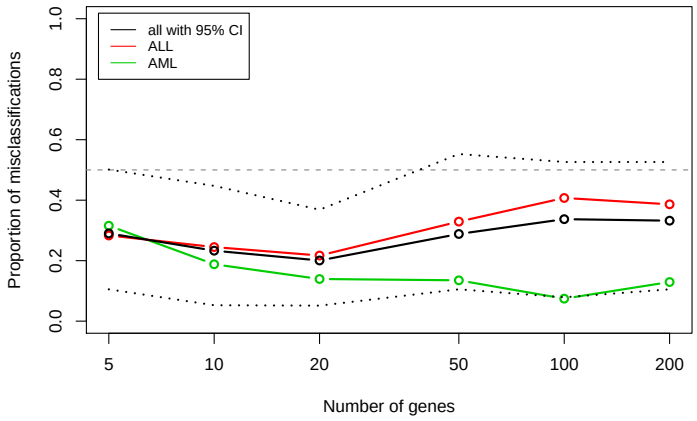

Figure 1: Missclassification rates in dependence of the number of genes.

The classification task is to distinguish between two types of leukemia, ALL and AML. Fig. 1 shows the overall classification accuracy, the sensitivity for prediction of ALL and the sensitivity for prediction of AML. The confidence interval of the overall classification rate is estimated from 200 random splits in training and test sets.

In order to reduce the computing time for the generation of the vignette, the gene expression data set has been reduced to the first 200 genes out of a total number of 3571 features. Classification rates will improve, when the calculation is done for the complete data set.

Next, we evaluate the performance of 10-gene predictors on the size of the training set (Fig. 2):

```
> val <- validate(GOLUB1[1:200, ], ngenes = 10, ntrain = "balanced")
> plot(val)
```

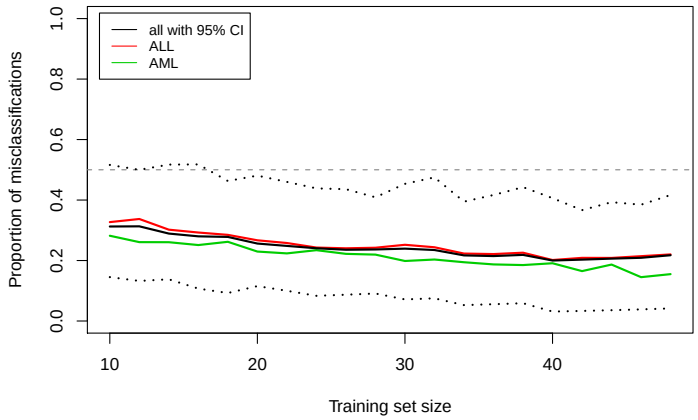

Figure 2: Missclassification rates for 10-gene predictors in dependence of the training set size. For each training set size, 200 splits in training and test sets were randomly drawn. Each training set contains an equal number of ALL and AML patients.

3 Predictor construction and validation

Two gene expression data sets of breast cancer are loaded. Both data sets were generated using the same type of microarrays. VEER is the original data set of 78 breast cancer samples [4]. VIJVER is a larger data set of 295 breast cancer samples including some of the profiles of the original data set [5]. An independent validation set VIJVER2 is obtained by removing the samples of VEER from VIJVER. A predictor of distance metastasis is fitted using the VEER data and validated in VIJVER2. Four methods `dist = "euclidean", "center", "angle", "cor"` are available for calculation of the distance between test samples and the centroids (see documentation of predict-method).

```
> library(cancerdata)
> data(VEER)
> data(VIJVER)
> VIJVER2 <- VIJVER[, setdiff(sampleNames(VIJVER), sampleNames(VEER))]
> predictor <- fit(VEER, method = "welch.test")
> prediction <- predict(predictor, VIJVER2, positive = "DM", dist = "cor")
```

The result of the prediction is a continuous score for each of the breast cancer patients. Three methods `score = "z", "zeta", "ratio"` are avaiable for calculation of the prediction score (see documentation prediction-class). The prediction score turns out to be significantly increased for patients that developed a distance metastasis within 5 years after surgery (Fig. 3). In fact, only three patients with prediction score zeta > 0.5 developed a distance metastasis.

```
> plot(prediction, type = "histogram", positive = "DM", score = "zeta")
```

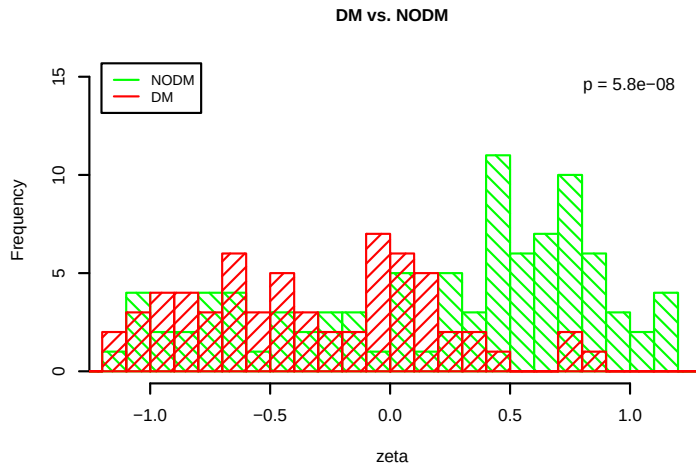

Figure 3: Histogram of the prediction score zeta patients that developed a distance metastasis within the first 5 years (DM) and patients that remained distance metastasis-free.

ROC analysis allows to trade off between sensitivity and specificity for the prediction of distant metastases. In fact, there is a cut off point for the prediction score yielding a sensitivity above 90% at a specificity of about 50% (Fig. 4). Confidence intervals of sensitivity and specificity are calculated by the Wilson procedure. The ROC curve runs significanlty above the diagonal with an area under the curve (AUC) of 0.74.

```
> plot(prediction, type = "roc", positive = "DM", score = "zeta")
```

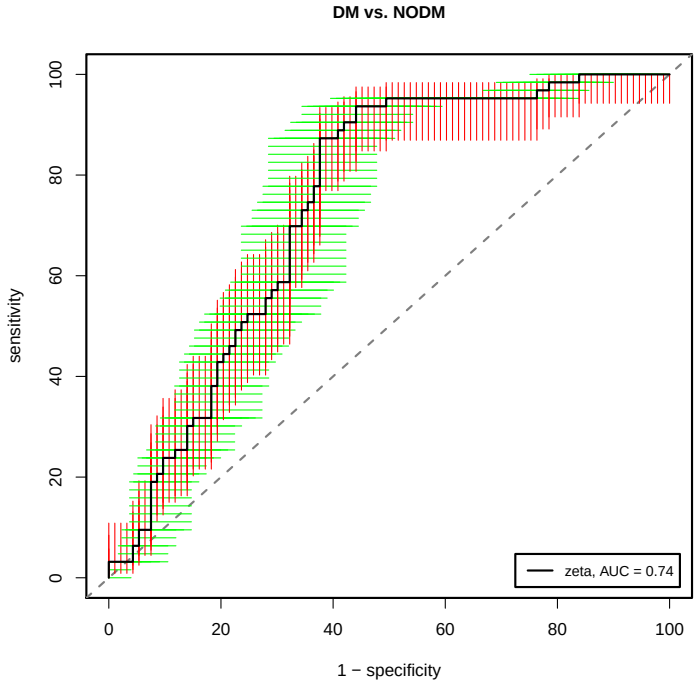

Figure 4: ROC curve for the prediction of distance metastases. 95% confindence intervals for sensitivity (red lines) and specificity (green lines). AUC = area under the curve.

Finally, a logistic regression model is fitted to the prediction score. Using Fig. 5, the probability of developing a distant metastasis within 5 years can be estimated from the gene expression based prediction score.

```
> plot(prediction, type = "logistic", positive = "DM", score = "zeta")

Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6610  -0.8354  -0.6052   1.0490   1.8924

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4716     0.1808  -2.608  0.00911 **
x            -1.4188     0.2997  -4.734  2.2e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 210.46  on 155  degrees of freedom
Residual deviance: 184.03  on 154  degrees of freedom
AIC: 188.03

Number of Fisher Scoring iterations: 4
```

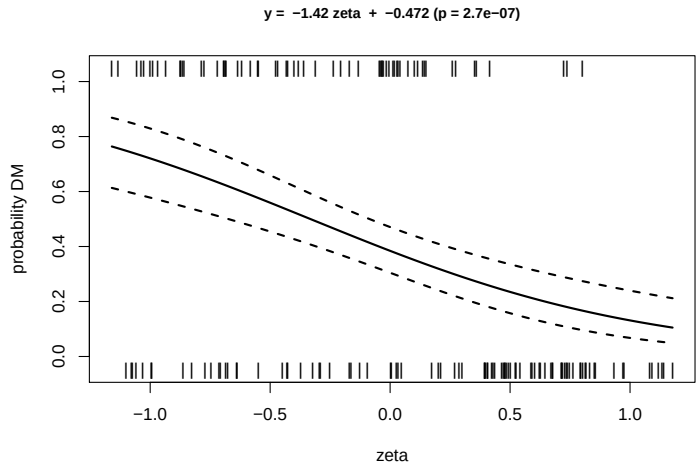

Figure 5: Probability of distance metastasis estimated in a logistic regression model including 95% confidence interval. Distribution of the patients with an unfavorable outcome (top track) and distribution of the patients with an favorable outcome (bottom track).

References

[1] Michiels S, Koscielny S, Hill C: *Prediction of cancer outcome with microarrays: a multiple random validation strategy.* Lancet 2005, 365:488-492.

[2] Wessels LF, Reinders MJ, Hart AA, *et al.*: *A protocol for building and evaluating predictors of disease state based on microarray data.* Bioinformatics 2005, 21(19):3755-62.

[3] Golub TR, Slonim DK, Tamayo P, *et al.*: *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science 1999, 286:531-537.

[4] van 't Veer LJ, Dai H, van de Vijver MJ, *et al.*: *Gene expression profiling predicts clinical outcome of breast cancer.* Nature 2002, 415:530-536.

[5] van de Vijver MJ, He YD, van't Veer LJ, *et al.*: *A gene-expression signature as a predictor of survival in breast cancer.* N Engl J Med 2002, 347:1999-2009.