Bundling classifiers by bagging trees

Torsten Hothorn*, Berthold Lausen

Institut für Medizininformatik, Biometrie und Epidemiologie, Friedrich-Alexander-Universität
Erlangen-Nürnberg, Waldstraße 6, D-91054 Erlangen, Germany

Received 9 February 2004; received in revised form 28 June 2004; accepted 29 June 2004
Available online 22 July 2004

Abstract

The quest of selecting the best classifier for a discriminant analysis problem is often rather difficult. A combination of different types of classifiers promises to lead to improved predictive models compared to selecting one of the competitors. An additional learning sample, for example the out-of-bag sample, is used for the training of arbitrary classifiers. Classification trees are employed to bundle their predictions for the bootstrap sample. Consequently, a combined classifier is developed. Benchmark experiments show that the combined classifier is superior to any of the single classifiers in many applications.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Bagging; Ensemble-methods; Method selection; Error rate estimation

1. Introduction

The toolbox for discriminant analysis or supervised learning problems contains a huge set of possible candidates for the construction of classification rules, e.g. linear discriminant analysis, classification trees, k nearest neighbors, as well as aggregated classifiers such as bagging, boosting and random forests. The individual classifiers perform well in some situations and fail under other conditions. The identification of the best procedure is often rather difficult, especially when the performance of a number of candidates is evaluated for learning samples of limited size. Dudoit et al. (2002) demonstrated this by comparing the
error rate of nine procedures for tumor classification using the Lymphoma gene expression data (Alizadeh et al., 2000) with only 81 observations.

Instead of selecting one single procedure, combining the competitors may improve classification rules. There are several approaches to the combination of different classifiers. A linear combination of the estimated conditional class probabilities is suggested by LeBlanc and Tibshirani (1996) and Mojirsheibani (1997), this is related to linear combinations of regression models (Breiman, 1996c; LeBlanc and Tibshirani, 1996). Merz (1999) uses correspondence analysis to combine the prediction of different classifiers. Majority voting of the predictions of the different classifiers is introduced by Mojirsheibani (1999, 2002). A framework for a “mix” of regression models is given by Yang (2001, 2003).

We propose to add the outcome of arbitrary classifiers (linear discriminant variables, predicted conditional class probabilities or predicted classes) to the set of original predictors for bagging classification trees (Breiman, 1996a, 1998). A number of classifiers are constructed for each bootstrap sample using an additional set of observations, for example, the out-of-bag sample or an independent bootstrap sample. The predictions of these classifiers are computed for the observations in the bootstrap sample and are used as additional predictors for a classification tree. The trees implicitly select the most informative predictors, and, in a sense, “bundle” the additional classifiers. The procedure is repeated sufficiently often and a new observation is classified by averaging the predictions of the multiple trees. Complex functional relationships between the predictor variables and the response are hard to detect by searching for the best rectangular split in a single node of a tree, but may be appropriately modelled by one of the additional classifiers. If, for example, the classes are separable by a hyperplane identified by a linear discriminant analysis, the values of the linear discriminant functions serving as additional predictor variables cover this knowledge and enable the trees to search for splits in highly informative linear combinations of the original predictors (“double-bagging”, Hothorn and Lausen, 2003b). The idea is in the spirit of Breiman (2001b): Instead of reducing the dimensionality, the number of possible predictors available to the classification trees is enlarged and the procedure is stabilized by bootstrap aggregation.

The suggested combination of classifiers has an additional merit: The error rates of the individual classifiers do not need to be estimated because the classifiers are implicitly selected by classification trees. Therefore, an honest estimate of the combined procedures’ misclassification error can be computed for example using cross-validation or the 0.632+ bootstrap (Efron and Tibshirani, 1997).

The paper is organized as follows. The used framework of discriminant analysis is discussed in Section 2 and we define the combined classifier in Section 3. Benchmark experiments in Section 4 show that the combined classifier is at least comparable to the best of the single classifiers or leads to an improvement with respect to misclassification error.

2. Model and error rates

Consider a standard model in discriminant analysis: a \( p \)-dimensional vector of predictors \( \mathbf{X} \in \mathbb{R}^p \), \( \mathbf{X} = (X_1, \ldots, X_p) \) is observed and associated with a class label \( Y \in \{1, \ldots, J\} \). The joint distribution function of the predictors and class labels is denoted by \( \mathcal{F} \). We
observe a learning sample $L_n$ of $n$ independent and identically distributed random samples from $\mathcal{F}$:
$$L_n = \{(y_i, x_i); i = 1, \ldots, n\}.$$  
Furthermore, the marginal distribution function of the predictors $X$ is denoted by $\mathcal{F}_X$.

We try to predict the class for a new observation $x_{\text{new}}$ based on the learning sample by a rule $C$, called classifier. A classifier is a function $C : \mathbb{R}^p \rightarrow \{1, \ldots, J\}$ which maps the $p$-dimensional predictors into the class labels. The function $C$ is a composition of two functions $C = g \circ c$ where $c : \mathbb{R}^p \rightarrow \mathbb{R}^{(J-1)}$ and $g : \mathbb{R}^{(J-1)} \rightarrow \{1, \ldots, J\}$.

The value $c(x_{\text{new}})$ may be the vector of the conditional class probability estimators and $g$ the argmax function. For the linear discriminant analysis (LDA), $c(x_{\text{new}})$ are the discriminant variables and $g$ is a function defining the class. The classifier is trained using the learning sample $L_n$, and we denote the dependence of the classifier $C$ on the learning sample by writing
$$C(x_{\text{new}}; L_n) = g(c(x_{\text{new}}; L_n)).$$

By convention, $c_j(x_{\text{new}}; L_n)$ is the $j$th element of the vector $c(x_{\text{new}}; L_n) \in \mathbb{R}^{(J-1)}$.

The performance of a classifier is measured by the misclassification error, i.e. the loss function
$$L(C) = P(\mathcal{F}(C(X) \neq Y).$$

The Bayes classifier is the classifier with minimum loss and is of the form,
$$C^{\text{Bayes}}(x) = \arg\max_{j=1,\ldots,J} P_j(x),$$

where $P_j(x) = P(Y = j | X = x)$ are the conditional class probabilities. The misclassification error of the Bayes classifier is called Bayes error,
$$L_{\text{Bayes}} = L(C^{\text{Bayes}}).$$

Since this misclassification error is hardly measurable in realistic setups, we focus on the conditional error rate, where the condition is on the learning sample $L_n$:
$$L(C(\cdot; L_n)) = P(\mathcal{F}(C(X; L_n) \neq Y|L_n).$$

3. Combining classifiers

The aim of this paper is to combine a main classifier $C^{\text{main}} = g^{\text{main}} \circ c^{\text{main}}$ and $K$ additional classifiers $C^k = g^k \circ c^k; k = 1, \ldots, K$ of arbitrary type. Superscripts are used to distinguish between the different classifiers. The combined classifier is defined as the main classifier
trained on the original $p$ predictors and additional $K(J - 1)$ transformations of the original predictors in the combined learning sample:

$$C_{\text{comb}}(x_{\text{new}}; L_n) = C_{\text{main}}(x_{\text{new}}; L_{\text{comb},n}),$$

where the combined learning sample $L_{\text{comb},n}$ is given by

$$L_{\text{comb},n} = \{(y_i, x_i, c^1(x_i), \ldots, c^K(x_i)); i = 1, \ldots, n\}.$$

We will choose classification trees as the main classifier. The motivation for adding transformations to the set of original predictors is that classification trees allow for rectangular splits of the form $X_i \leq b$ but classification trees based on the combined learning sample $L_{\text{comb},n}$ allow for more general splits of the form $c^k(X) \leq b$. For example, if the discriminating variables of a linear discriminant analysis are used as additional predictors, linear splits of the form $\beta^T X \leq b, \beta \in \mathbb{R}^p$ are possible. Note that in this setup, $c^k(x); k = 1, \ldots, K$ may or may not depend on the learning sample $L_n$.

One version of random forests (Forest-RC, Breiman, 2001a) use random linear combinations of the predictors. In contrast to random combinations we will add possibly non-linear transformations $c^k(X)$ of the predictors that depend on the data. However, an additional learning sample for training $c^k$ is often not available and simple sample splitting is inefficient for small learning samples. Computing $c^k$ and $c_{\text{main}}$ using the same data is not appropriate. In this case, $c_{\text{main}}$ is likely to select transformations derived from the individual classifier $c^k$ which overfits the data most. Therefore we will use either the out-of-bag sample or a second independent bootstrap sample for the bootstrap aggregated combined classifiers as learning samples for the training of the additional classifiers as follows.

Let $L_n^*$ denote a bootstrap sample of size $n$ from the empirical distribution function $\hat{F}$:

$$L_n^* = \{(y_i^*, x_i^*); i = 1, \ldots, n\} \sim \hat{F}.$$

In standard bagging, the multiple trees trained on the bootstrap samples $L_n^*$ are aggregated by class majority voting, i.e. voting of the class predictions for a new observation. However, Breiman (1996a) mentioned that averaging the conditional class probability estimators and choosing the class with highest average conditional class probability leads roughly to the same results. We therefore define the bootstrap aggregated main classifier by the expectation of $c_{\text{main}}$ with respect to $\hat{F}$:

$$C_{\text{bagmain}}(x_{\text{new}}; L_n) = g_{\text{main}} \left( E_{\hat{F}} c_{\text{main}}(x_{\text{new}}; L_n^*) \right) = g_{\text{main}} \left( \int c_{\text{main}}(x_{\text{new}}; L_n^*) d\hat{F}(L_n^*) \right)$$

and approximate it by a finite number of $B$ bootstrap samples

$$C_{B_{\text{bagmain}}}(x_{\text{new}}; L_n) = g_{\text{main}} \left( B^{-1} \sum_{b=1}^B c_{\text{main}}(x_{\text{new}}; L_n^{*(b)}) \right). \quad (1)$$

In each bootstrap sample, roughly one third of the original observations are left out due to sampling $n$ out of $n$ with replacement. Breiman (1996b) calls the observations in $L_n \setminus L_n^*$,
i.e. the observations which are not element of the bootstrap sample, “out-of-bag” and uses this additional sample for improving estimators of the conditional class probabilities, because the correlation between the bootstrap sample and the out-of-bag sample vanishes as the sample size increases. Bylander (2002) proposes a correction of the out-of-bag error rate estimator and Rao and Tibshirani (1997) discuss a weighted prediction where the weights are determined from the out-of-bag samples. In contrast to those suggestions, we will use the out-of-bag observations $L_n \setminus L_n^*$ for training of additional classifiers. Another possibility is to draw a second independent bootstrap sample $L_n^{**}$ for training of the classifiers $c^1, \ldots, c^K$.

For a bootstrap sample $L_n^*$, the combined learning sample is defined by an additional learning sample $L_a$:

$$L_{comb,n}^* = \{(y_i^*, x_i^*, c_1(x_i^*; L_a), \ldots, c^K(x_i^*; L_a)); i = 1, \ldots, n\},$$

where either the the out-of-bag sample $L_a = L_n \setminus L_n^*$ or a second independent bootstrap sample $L_a = L_n^{**}$ is now used as learning sample for $c^k, k = 1, \ldots, K$. This learning sample consists of the original predictors as well as $K(J - 1)$ transformations of them, induced by additional classifiers $c^1, \ldots, c^K$. Note that the observations including the additional predictors in $L_{comb,n}^*$ are independently and identically distributed only when the classifiers $c^1, \ldots, c^K$ are trained using a second independent bootstrap sample.

The bootstrap aggregated combined classifier is now given by

$$C_{bagcomb}(x_{new}; L_n) = g_{main}(E \hat{F}_{comb} (x_{new}; L_n^*)), \quad \text{where}$$

$$c_{comb}(x_{new}; L_n^*) = c_{main}(x_{new}; L_{comb,n}^*)$$

(2)

and again we approximate $C_{bagcomb}$ by a finite number of $B$ bootstrap replications:

$$C_{bagcomb_B}(x_{new}; L_n) = g_{main}\left(B^{-1} \sum_{b=1}^{B} c_{comb}(x_{new}; L_n^{*(b)})\right).$$

Each of the $B$ classification trees works as a main classifier which selects and combines the predictions of the additional classifiers $c^1, \ldots, c^K$. Consequently we will call the procedure “bundling”. Instead of sampling $n$ out of $n$ with replacement, Bühlmann and Yu (2002) suggested subsampling (“subagging”), i.e. sampling 50% of the data without replacement. This ensures that the out-of-bag sample is an independent learning sample which always contains half of the observations. In analogy, we call this procedure “subundling”.

4. Benchmark experiments

In this section, we illustrate the performance of the combination of classifiers via bundling using three artificial, five small and three larger benchmark problems. The experiments were conducted using the ipred package (Peters et al., 2002, see Appendix B) in the R statistical environment (version 1.5.1, Ihaka and Gentleman, 1996, http://www.R-project.org).
Table 1
Sample size of the learning samples, the number of predictors $p$ and number of classes $J$ for the benchmark datasets under consideration

<table>
<thead>
<tr>
<th>Data set</th>
<th>$n$</th>
<th>Test size</th>
<th>$p$</th>
<th>$J$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twonorm</td>
<td>300</td>
<td>18.000</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Threenorm</td>
<td>300</td>
<td>18.000</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Ringnorm</td>
<td>300</td>
<td>18.000</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>699</td>
<td>—</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>—</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Diabetes</td>
<td>768</td>
<td>—</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>—</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>196</td>
<td>—</td>
<td>62</td>
<td>2</td>
</tr>
<tr>
<td>Satellite</td>
<td>4435</td>
<td>2000</td>
<td>36</td>
<td>6</td>
</tr>
<tr>
<td>Shuttle</td>
<td>5803</td>
<td>5802</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>DNA</td>
<td>2000</td>
<td>1186</td>
<td>180</td>
<td>3</td>
</tr>
</tbody>
</table>

The breast cancer, ionosphere, diabetes, glass, satellite, shuttle and DNA datasets from the UCI machine learning repository (Blake and Merz, 1998) are assembled in the R package mlbench (Leisch and Dimitriadou, 2001). The code for generating the twonorm, threenorm and ringnorm data is part of the mlbench package as well. A detailed description of these artificial problems is given in Breiman (1998). In addition, we use the glaucoma dataset which is part of the ipred package. In this ophthalmological application, 62 predictors derived from laser scanning images of the optic nerve head shall be used to classify 196 eyes as either normal or glaucomatous (Hothorn and Lausen, 2003a, b).

We study bundling of three individual classifiers: linear discriminant analysis (LDA), $k$ nearest neighbors ($k$-NN, with $k = 5$ and 10) as well as the logistic regression model (LR). The multinomial model is used for problems with more than two classes.

Since these classifiers are trained using the out-of-bag observations only, this subset of the learning sample can be too small. For example, the out-of-bag sample in the glaucoma problem contains 72 observations on average. The estimation of a covariance matrix for the LDA with 62 predictors is usually infeasible. There are two strategies to deal with this problem. One possibility is to reduce the number of possible predictors. We use a stabilized LDA (sLDA, Läuter, 1992) based on low-dimensional PC-$q$ scores (Läuter et al., 1996, 1998, see Appendix A for more details). Another possibility is to give more weight to the out-of-bag sample via “subbundling” using subsampling as an alternative resampling plan.

The values of the linear discriminant functions of the stabilized LDA as well as the predicted conditional class probabilities of nearest neighbors and the logistic regression model are combined. For bagging and bundling, 50 unpruned trees are used. We additionally report the error rates of random forests with 50 trees (Forest-RI, Breiman, 2001a) using R package randomForest (Liaw and Wiener, 2002, version 3.3-2), where the number of randomly selected predictors in each node is chosen as the ceiling of $\log_2(p + 1)$.

The misclassification error for the artificial problems is the average over 100 simulation runs, where the learning samples are of size 300 and the error rate is computed using one single test sample of size 18000. For the larger datasets, a test sample is selected randomly for the larger problems. Table 1 reports the sample sizes of learning and test samples for the
Table 2
Simulated misclassification errors for the artificial problems with standard deviations given in parentheses

<table>
<thead>
<tr>
<th></th>
<th>sLDA</th>
<th>5-NN</th>
<th>10-NN</th>
<th>LR</th>
<th>Bagg</th>
<th>Bund</th>
<th>SBund</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twonorm</td>
<td>2.5 (0.1)</td>
<td>4.0 (0.3)</td>
<td>3.4 (0.2)</td>
<td>5.2 (1.2)</td>
<td>7.5 (1.4)</td>
<td>2.8 (0.4)</td>
<td>2.6 (0.2)</td>
<td>5.2 (0.5)</td>
</tr>
<tr>
<td>Threenorm</td>
<td>17.3 (0.5)</td>
<td>18.2 (1.2)</td>
<td>16.9 (1.2)</td>
<td>18.0 (0.6)</td>
<td>20.2 (1.3)</td>
<td>16.9 (0.5)</td>
<td>16.8 (0.5)</td>
<td>18.6 (0.7)</td>
</tr>
<tr>
<td>Ringnorm</td>
<td>38.8 (2.0)</td>
<td>47.5 (0.4)</td>
<td>49.3 (0.2)</td>
<td>38.0 (1.3)</td>
<td>10.2 (1.8)</td>
<td>6.8 (0.7)</td>
<td>8.1 (0.8)</td>
<td>7.6 (1.2)</td>
</tr>
</tbody>
</table>

The Bayes errors are 2.3 (twonorm), 10.5 (threenorm) and 1.3 (ringnorm). The classifiers are stabilized LDA (sLDA), nearest neighbors (k-NN), the logistic regression model (LR), bagging (Bagg), bundling (Bund) and subundling (SBund) as well as random forests (RF).

real world applications. The misclassification error for the smaller problems is estimated by averaging the misclassification error of ten runs of ten-fold cross-validation. The results for the artificial problems are given in Table 2, and Fig. 1 shows a graphical representation. The estimated misclassification errors for the real world applications are displayed in Table 3.

Bundling is, except for the diabetes data, at least as good as any of the single classifiers or bagging. For the twonorm and threenorm problem, the misclassification error of bundling is very close to the misclassification error of sLDA or 10-NN, respectively. In the breast cancer data, the error of bundling is comparable to sLDA and for the glass and shuttle data, the misclassification error is in the order of the error estimated for bagging. For the ringnorm, ionosphere, glaucoma, satellite and DNA data, bundling improves the misclassification error. Subundling does not lead to an additional improvement, the error rates of bundling are comparable for both resampling plans under investigation. The error rates of bundling or subundling are comparable to the misclassification error estimated for random forests for the real world datasets (except for the glass data) and slightly smaller for the artificial ones.

5. Discussion

Classification trees (Breiman et al., 1984) search for the best rectangular split in any of the predictors in each node. There are several approaches to relax this restriction in order to allow for splits in more general structures. For example, Breiman et al. (1984) suggest to search for splits in linear combinations of the predictors in each node and Kim and Loh (2001) propose multiway splits in linear combinations. Buttrey and Karo (2002) incorporate nearest neighbors in the leaves of a tree.

Our proposal is to enlarge the possible predictors available to any of the single classification trees in bagging, i.e. using trees to “bundle” the predictions of arbitrary classifiers. The additional predictors are transformations of the original predictors derived from arbitrary classifiers which are trained using either the out-of-bag samples or additional bootstrap samples as learning samples.

The investigations on benchmark datasets show that the combination of linear discriminant analysis, nearest neighbors and the logistic regression model leads to improved predictive models. Meyer et al. (2003) performed a more extensive simulation study including a
combination of LDA and classification trees and observed similar results. Beside the good performance of bundling, the need for an explicit selection of one of the individual classifiers disappears. As a consequence, only the error rate of bundling needs to be estimated.
Table 3
Estimated misclassification errors for the benchmark datasets, see Table 2 for a description

<table>
<thead>
<tr>
<th></th>
<th>sLDA</th>
<th>5-NN</th>
<th>10-NN</th>
<th>LR</th>
<th>Bagg</th>
<th>Bund</th>
<th>SBund</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>3.3</td>
<td>6.4</td>
<td>8.3</td>
<td>7.4</td>
<td>4.0</td>
<td>3.0</td>
<td>3.1</td>
<td>3.0</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>13.8</td>
<td>15.6</td>
<td>16.6</td>
<td>11.8</td>
<td>8.3</td>
<td>6.5</td>
<td>6.6</td>
<td>6.6</td>
</tr>
<tr>
<td>Diabetes</td>
<td>27.3</td>
<td>28.5</td>
<td>26.0</td>
<td>22.4</td>
<td>24.3</td>
<td>24.1</td>
<td>24.7</td>
<td>23.8</td>
</tr>
<tr>
<td>Glass</td>
<td>42.4</td>
<td>32.7</td>
<td>37.3</td>
<td>35.7</td>
<td>23.2</td>
<td>24.9</td>
<td>26.3</td>
<td>22.1</td>
</tr>
<tr>
<td>Glaucoma</td>
<td>16.0</td>
<td>20.1</td>
<td>20.2</td>
<td>22.9</td>
<td>15.9</td>
<td>15.4</td>
<td>15.0</td>
<td>15.1</td>
</tr>
<tr>
<td>Satellite</td>
<td>19.3</td>
<td>8.7</td>
<td>9.6</td>
<td>19.2</td>
<td>8.4</td>
<td>7.4</td>
<td>7.7</td>
<td>7.4</td>
</tr>
<tr>
<td>Shuttle</td>
<td>8.2</td>
<td>0.4</td>
<td>0.6</td>
<td>2.9</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>DNA</td>
<td>8.1</td>
<td>19.5</td>
<td>16.2</td>
<td>10.4</td>
<td>4.5</td>
<td>3.1</td>
<td>3.1</td>
<td>5.6</td>
</tr>
</tbody>
</table>

For a comparison with other candidate classifiers like support vector machines and boosting we refer to the results reported by Meyer et al. (2003).

Especially for small learning samples with a large number of predictors, the out-of-bag sample may be too small for computing some of the additional classifiers. We propose two approaches to deal with the problem: either use a stabilized form of the individual classifier (for example stabilized LDA) or choose an alternative resampling plan, namely sample splitting. Another possibility is to select a small random subset of the predictors for the computation of the additional classifiers, an approach not studied here. Of course, the use of a small random selection of predictor variables available to split in each node of the classification trees as in random forests is possible for bundling as well.

Acknowledgements

We would like to thank Christine Vogel and Achim Zeileis for their very helpful comments and Janice Hegewald for proof-reading. Financial support from Deutsche Forschungsgemeinschaft, Grant SFB 539-A4/C1 is gratefully acknowledged.

Appendix A. Stabilized linear discriminant analysis

Let $\tilde{x}$ denote the $p$-dimensional mean vector over all $n$ observations and let

$$W = \sum_{i=1}^{n} (x_i - \tilde{x})(x_i - \tilde{x})^\top$$

denote the total sum of products matrix as defined in Läuter et al. (1998). The stabilized linear discriminant analysis uses the PC-$q$ scores (Läuter, 1992; Läuter et al., 1998), i.e. a $q$-dimensional linear combination $\tilde{x}_i = x_iD_q$ (with $q < p$), instead of the original $p$-dimensional predictors $x_i$. The matrix $D_q$ is the $p \times q$ matrix of eigenvectors with eigenvalues greater one in the eigenvalue problem $WD = \text{Diag}(W)D_A$. Theorem 2 of Läuter et al. (1998) states that $\tilde{x}_i$ are left-spherically distributed random variables. A stabilized
linear discriminant analysis can therefore be computed based on the $q$-dimensional linear combination $x_iD_q$ of the original predictors.

Appendix B. Software

Bundling as proposed here is implemented in the R package ipred (Peters et al., 2002), the software is released under GPL and is available from http://CRAN.R-project.org. Bundling of sLDA, k-NN ($k = 5$ and 10) and LR for the glaucoma data (Section 4) can be computed along the following lines:

```r
R> library(ipred)
R> data(GlaucomaM)
R> mybundle <- list(
    # stabilized LDA
    list(model = slda, predict = function(object, newdata)
      predict(object, newdata)$x),
    # 5-NN
    list(model = function(...) ipredknn(..., k = 5),
      predict = predict),
    # 10-NN
    list(model = function(...) ipredknn(..., k = 10),
      predict = predict),
    # LR or multinomial model, resp.
    list(model = function(...) multinom(..., trace = FALSE),
      predict = function(obj, newdata)
      predict(obj, newdata, type = "prob"))
)
R> bagging(Class ~ ., data = GlaucomaM, comb = mybundle,
          nbagg = 50)
```

References


