Comput Stat (2011) 26:355–369
DOI 10.1007/s00180-011-0249-1

ORIGINAL PAPER

Classification of repeated measurements data using tree-based ensemble methods

Werner Adler · Sergej Potapov · Berthold Lausen

Received: 30 October 2009 / Accepted: 12 February 2011 / Published online: 26 March 2011

Abstract In many medical applications, longitudinal data sets are available. Longitudinal data, as well as observations from paired organs, show a dependency structure which should be respected in the evaluation. Adler et al. (Comput Stat Data Anal 53(3):718–729, 2009) proposed various bootstrapping strategies for ensemble methods based on classification trees for two measurements of paired organs. These strategies have shown to improve the classification performance compared to the traditional approach, where only one observation per subject is used. We extend the methodology to the situation, where an arbitrary number of observations per individual are available and investigate the performance of the proposed methods with bagged classification trees (bagging) and random forests in the situation of longitudinal data. Moreover, we adapt the estimation of classification performance criteria to repeated measurements data. The clinical data set consists of morphological examinations of both eyes of glaucoma patients and healthy controls over a time period of up to 7 years. The performance of our modified classifiers is evaluated by a subject-based leave-one-out bootstrap ROC analysis. Simulation results and results for the glaucoma data set demonstrate that our proposal is an improvement of adhoc strategies and of the use all measurements of each subject or block strategy.

Keywords Bagging · Bootstrap · Longitudinal data · Random forest · ROC analysis

W. Adler (✉) · S. Potapov
University of Erlangen-Nuremberg, Waldstrasse 6,
91054 Erlangen, Germany
e-mail: werner.adler@imbe.med.uni-erlangen.de

B. Lausen
University of Essex, Wivenhoe Park,
Colchester CO4 3SQ, UK

1 Introduction

Data in medical applications often show some dependency in their structure. Many organs of the human body are paired. In many progressive diseases it is common to repeat examinations over a longer time period. Thus, longitudinal data sets are available for the patients. Especially in the paired case, it is common practice to use one-either randomly or knowledge based-selected observation or to calculate the mean value. Although these evaluation procedures are statistically sound, information is lost. Similarly, learning sets for classification tasks often consist of only the newest observations if longitudinal data are available. The reason may be the intuitive argument that when a disease is progressive, newer examinations are more typical for the disease and therefore should lead to better classification results. Models like mixed-effects logistic regression are able to account for the dependency in the data structure. In the independent data case however, model-free ensemble methods like bootstrap aggregated classification trees (bagging) or random forests often show better classification performance than traditional statistical classifiers (Adler et al. 2008). In this work, we examine the question if the performance of ensemble methods can be increased by making use of a dependent data structure. Brenning and Lausen (2008) demonstrated that the variance of the error estimate can be reduced by paired cross validation in the case of ensemble methods applied to paired data. Recently, random forests was adapted for the analysis of cluster-correlated data (Karpievitch et al. 2009). Here, we adapt bootstrapping techniques proposed in Adler et al. (2011) to the case where more than two observations per subject are available. The performance of bagged classification trees and random forests with our bootstrap strategies is tested in a highly nonlinear but simple simulation model and with a medical data set consisting of repeated topography measurements of left and right eyes of glaucoma patients and healthy controls.

2 Ensemble classification of dependent observations

Our aim is to improve the classification of single future observations or single future measurements and not the classification of a new patient or experiment block. Bagging (Breiman 1996) is defined as the aggregation of several base learners which are trained on different bootstrap samples of the learning set. The base classifier in our case is the classification tree (Breiman et al. 1984). A modification of bagged classification trees is given by a random forest (Breiman 2001). Here, at each node of every tree a different subset of all variables is used for training. This leads to a reduced correlation between the trees and thus to a better classification performance of the ensemble.

Bagging and random forest do not consider any dependency structure in the data. This can have a negative effect on the classification performance, as a simple extreme example illustrates: pairs of observations can be highly correlated, say because they stem from a left and right organ of the same person. For illustration purposes we set the correlation to one, so drawing N observations of this learning set is equivalent to drawing $2 \cdot N$ observations of a learning set consisting of only one observation per subject. Thus, approximately $(1 - \exp(-2)) \cdot N = 0.865 \cdot N$ different observations are drawn. This is in contrast to the $0.632 \cdot N$ different observations that were drawn if the bootstrap sample size was of size N. Thus, a higher correlation between trees has to be expected, which is contrary to the idea of ensembles, as they are expected to perform better when the variation between the single trees is higher.

Recently, the method RF++, a variation of random forests has been proposed, which accounts for dependency structure in the data by subject based bootstrapping. Using our example from above, where the learning set consists of $2 \cdot N$ observations of N subjects, we draw N of N subjects with replacement and use the observations of these subjects for training. Because in this example the correlation between two observations of one subject is one, we draw $0.632 \cdot N$ different observations, where each one is doubled.

We further elaborate the idea of subject based bootstrapping and introduce more refined strategies. To keep our idea specific, we develop the strategies for our clinical example, which will be introduced in more detail in Sect. 5. In this example, the data consist of repeated measurements of left and right eyes from glaucoma patients and healthy controls. Formally, our learning set $\mathcal{L}$ consists of N persons with one or more (repeated) measurements from the left and/or right eye:

$$\mathcal{L} = \left\{(y_i^{L_{j(i)}}, x_i^{L_{j(i)}}, y_i^{R_{k(i)}}, x_i^{R_{k(i)}})\right\}, \\ i = 1, \ldots, N, \quad j(i) = 1, \ldots, J_i, \quad k(i) = 1, \ldots, K_i \tag{1}$$

where $x_i^{s_j}$ is a p-dimensional predictor variable, $x_i^{s_j}=(x_{i1}^{s_j}, \ldots, x_{ip}^{s_j}) \in \mathcal{R}^p$, and $s_j \in \{L_1, \ldots, L_{J_i}, R_1, \ldots, R_{K_i}\}$ denotes repeated measurements from the left (L) or right (R) eye (J_i and K_i are the numbers of repeated measurements per person i and eye). The class membership is given by $y_i^{s_j} \in \{0, 1\}$.

To create the learning samples for the individual base classifiers (trees), B bootstrap samples of the set of subjects $\mathcal{S} = \{1, \ldots, N\}$ are drawn with the drawn subjects denoted as i^*. To create the learning sample we introduce strategies τ, with $\tau \in \{1, \Omega\}$. For $\tau = 1$, one observation is drawn per subject. The strategy $\tau = \Omega$ means that all observations per subject are selected for the learning set. Thus, the learning samples $\mathcal{L}^*_{b,\tau}$ of the individual trees are defined as

$$\mathcal{L}^*_{b,\tau} = \{(v_l, u_l), l = 1, \ldots, N_\tau\}, \qquad b = 1, \ldots, B, \tag{2}$$

where (v_l, u_l) consists of the p-dimensional measurement u_l and class variable v_l of one eye and N_1 is N and $N_\Omega = \sum_{i^*}(J_{i^*} + K_{i^*})$ is the total number of all left (J) and right (K) observations per drawn subjects i^*.

The base classifier of the ensemble defines a mapping of new measurements $\tilde{u}$ to an estimated class membership $\hat{v}$:

$$C^{base}(\tilde{u}, \mathcal{L}^*_{b,\tau}) : \tilde{u} \to \hat{v} \tag{3}$$

with $\hat{v} \in \{0, 1\}$.

We define the estimated probability $\hat{p}(\hat{v} = 1|\tilde{u})$ given by an ensemble as:

$$\hat{p}_{\tau}^{ens}(\hat{v} = 1|\tilde{u}, \mathcal{L}) = \frac{1}{B}\sum_{b=1}^{B} C^{base}(\tilde{u}, \mathcal{L}_{b,\tau}^{*}) \tag{4}$$

The final classification found by our ensemble depends on a threshold $t \in [0, 1]$:

$$C_{\tau}^{ens}(\tilde{u}, \mathcal{L}, t) = I\left(\hat{p}_{\tau}^{ens}(\hat{v} = 1|\tilde{u}, \mathcal{L}) \geq t\right), \qquad t \in [0, 1]. \tag{5}$$

3 Classifier performance estimation

3.1 True and false positive rates (TPR and FPR)

Classification performance can be measured by the classification error or by ROC analysis. The classification error simply is given by the proportion of falsely classified observations. To determine the classification error, one specific cutpoint has to be chosen. This cutpoint is often set according to an a priori probability distribution of the response variable, when the marker is given by the class membership probability determined by the classification method. ROC analysis allows for a more general performance examination of the classifier, as no single cutpoint needs to be determined, but true and false positive rates (TPR and FPR, respectively) are calculated for several cutpoints. These rates are defined as:

$$\text{TPR} = \frac{\text{true positives classified as positive}}{\text{total true positives}},$$

and

$$\text{FPR} = \frac{\text{true negatives classified as positive}}{\text{total true negatives}}.$$

If sufficient data are available, the classifier performance can be reported using one part of the data for training of the classifier and another part of the data for testing. Here, the calculation of the classification error or the area under the ROC curve (AUC) is straightforward.

On the other hand, if only few data are available, as it is often the case in various applications, resampling methods as cross-validation or the bootstrap are used. There exists a tradeoff between bias and variance for both of these methods: while cross-validation tends to have the lower bias, it tends to suffer from a larger variance than the bootstrap (Hastie et al. 2001). Brenning and Lausen (2008) examine cross-validation for a paired data-structure.

Bootstrap estimation of classifier performance naively is done by comparing the true and false positive rates of all observations. We followed this approach in our simulation study, as the number of observations per subject was equal. In our glaucoma example on the other hand, the number of observations varies widely between different subjects and a naive performance estimation therefore would put higher weight on subjects with many observations. So in the glaucoma example, we estimate the classifier performance by a subject-based leave-one-out bootstrap with $B = 100$ bootstrap samples. For the definition of the bootstrap estimated TPR and FPR we adapt the proposal of Adler and Lausen (2009) who generalize the leave-one-out bootstrap proposal of Hastie et al. (2001):

$$\widehat{\text{TPR}}^{(1)}(t) = \frac{1}{|\mathcal{S}_1|}\sum_{i\in\mathcal{S}_1}\frac{1}{|\mathcal{B}^{-i}|}\sum_{b\in\mathcal{B}^{-i}} C_\tau^{ens}(x_i, \mathcal{L}^{*b}, t), \qquad t \in [0, 1], \tag{6}$$

and

$$\widehat{\text{FPR}}^{(1)}(t) = \frac{1}{|\mathcal{S}_0|}\sum_{i\in\mathcal{S}_0}\frac{1}{|\mathcal{B}^{-i}|}\sum_{b\in\mathcal{B}^{-i}} C_\tau^{ens}(x_i, \mathcal{L}^{*b}, t), \qquad t \in [0, 1], \tag{7}$$

where $\mathcal{B}^{-i}$ denotes the set of indices b of those bootstrap samples that do not contain observations of subject i, $|.|$ denotes cardinality, $\mathcal{L}^{*b}$ is the bth bootstrap sample obtained by drawing with replacement from $\mathcal{L}$ and $\mathcal{S}_c = \{i \in \mathcal{S} | y_i^{\cdot} = c\}$, $c \in \{0, 1\}$ is the set of subjects i in class c.

3.2 Bootstrap estimated TPR and FPR of dependent data

One possibility is to ignore the dependence structure and to compute estimates for the classification of single observations and measurements or of patients and blocks. Another possibility is to generalize the proposal of Brenning and Lausen (2008) for the classification of single observations and measurements. Here, we suggest a method to estimate TPR and FPR for the classification of patients and blocks.

To account for the dependent data structure, we first average the bootstrap estimated classification results of all repeated measurements of a single eye:

$$C_{i,\tau}^{s}(\mathcal{L}^{*b}, t) = \frac{1}{|\{j|\exists x_i^{s}\}|}\sum_{\{j|\exists x_i^{s}\}}\frac{1}{|\mathcal{B}^{-i}|}\sum_{b\in\mathcal{B}^{-i}} C_\tau^{ens}(x_i^{sj}, \mathcal{L}^{*b}, t), \tag{8}$$

i.e. $C_{i,\tau}^{s}$, $s \in \{L, R\}$ is the classification result of a single eye of subject i. The set $\{j|\exists x_i^{s}\}$ denotes the set of indices j which refer to all repeated measurements existing for eye side $s \in \{L, R\}$, where a different number of repeated measurements for different eye sides and subjects is allowed. Then, the classification results of both eyes of one subject are averaged, denoted as $C_{i,\tau}$:

$$C_{i,\tau}(\mathcal{L}^{*b}, t) = \frac{1}{|\{s|\exists x_i^{s\cdot}\}|}\sum_{\{s|\exists x_i^{s\cdot}\}} C_{i,\tau}^{s}(\mathcal{L}^{*b}, t). \tag{9}$$

Here, depending on subject i, $\{s|\exists x_i^{s\cdot}\}$ consists of the index for the left (L), or the right (R) eye. With the classification of a single subject, the leave-one-out bootstrap estimated TPR and FPR are defined as:

$$\widehat{\text{TPR}}^{(1)}_{dep}(t) = \frac{1}{|\mathcal{S}_1|} \sum_{i \in \mathcal{S}_1} C_{i,\tau}(\mathcal{L}^{*b}, t), \qquad t \in [0, 1], \tag{10}$$

and

$$\widehat{\text{FPR}}^{(1)}_{dep}(t) = \frac{1}{|\mathcal{S}_0|} \sum_{i \in \mathcal{S}_0} C_{i,\tau}(\mathcal{L}^{*b}, t), \qquad t \in [0, 1]. \tag{11}$$

Thus, with

$$\widehat{\text{ROC}}^{(1)}_{dep}(t) = \left\{ \widehat{\text{FPR}}^{(1)}_{dep}(t), \widehat{\text{TPR}}^{(1)}_{dep}(t), \quad t \in [0, 1] \right\}, \tag{12}$$

the area under the curve is defined as:

$$\widehat{\text{AUC}}^{(1)}_{dep} = \int_t \widehat{\text{ROC}}^{(1)}_{dep}(t) dt, \qquad t \in [0, 1]. \tag{13}$$

4 Simulation study

Being interested in nonparametric approaches we utilize the well known classification benchmark problem "Spirals" (Lang and Witbrock 1988) for our simulation study. In this problem, two intertwined spirals represent two classes and therefore a linear model must perform poorly while nearest-neighbours, random forest or bagging can be used instead. The data are two dimensional which allows us to have a very clear impression of how repeated measurements can be constituted and facilitates a transparent regulation of the data structure. The *R* package `mlbench` V1.1-6 (Leisch and Dimitriadou 2010) offers the possibility to simulate an arbitrary number of noisy observations of these spirals. We examined two simulation setups with standard deviation sd equal to 0.1 (simulation model 1) and 0.2 (simulation model 2), respectively. These simulated observations serve es "true" values of the "subjects" which are not observable in reality. To simulate repeated measurements, we added gaussian noise to these values with mean value 0 and standard deviation 0.01, 0.03, 0.05, and 0.1, denoted as σ_{obs}. A standard deviation equal to 0.01 represents very reliable measurements, while a standard deviation equal to 0.1 simulates the situation, in which repeated measurements can vary considerably. We examined samples of size 50, 100, 150, and 200. A varying number of measurements per subject was simulated, ranging from 1 to 10. A large reference data set consisting of 5,000 observations to illustrate the classification problem, the "true but not measurable" values of 200 subjects and three repeated measurements with σ_{obs} varying from 0.01 to 0.1 are illustrated in Fig. 1 for both simulation setups with standard deviation 0.1 and 0.2, respectively. For each of the

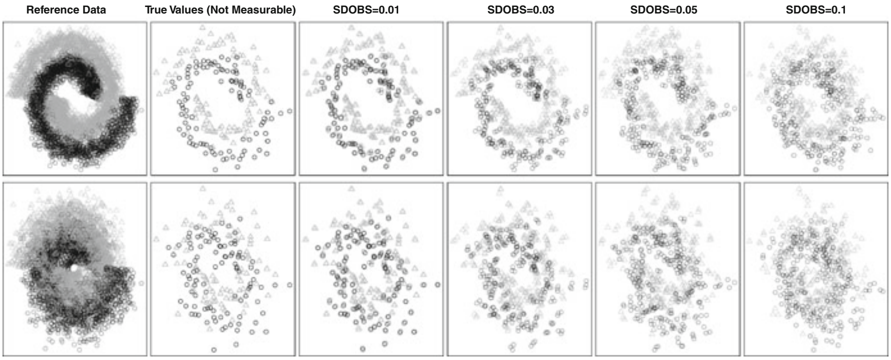

Fig. 1 Classification problem “Spirals”: two arms of a spiral represent two classes. We simulated the data set with a noise (*top row*: model 1, sd = 0.1; *bottom row*: model 2, sd = 0.2) added on the ideal values on the spirals. The figure shows from *left* to *right*: a large data set illustrating the classification problem; the “true” values per “subject” in one simulation run with 200 “subjecs”; three repeated measurements with increasing σ_{obs}

different parameter combinations we performed 100 iterations in which we simulated the according learning data set. A random forest consisting of 1,000 classification trees and bagging consisting of 100 classification trees were trained with the proposed bootstrap sampling strategies (1) $\tau = 1$ and (2) $\tau = \Omega$. Additionally, (3) a naive training was performed, where the data structure was not taken into account, (4) one of the measurements per subject was chosen randomly prior to training and training was then performed with a reduced data set, and finally (5) the mean values of all repeated measurements per subject were calculated and training was performed using a data set consisting of the mean values. In each of the 100 iterations, we also simulated a data set consisting of 100 observations with the appropriate noise (sd = 0.1 or 0.2) for testing. The classification performance of random forest and bagging was evaluated by calculating the AUC.

5 Medical example: Glaucoma classification

Glaucoma is one of the most common causes for blindness worldwide (Quigley 1996). The name glaucoma covers a variety of eye diseases which have in common the progressive and irreversible degeneration of the optic nerve head (ONH). Patients suffering from glaucoma typically notice the disease in a late stage, when visual field loss which is compensated by a high degree becomes obvious. Therefore, early detection by ophthalmologists is mandatory. The Heidelberg Retina Tomograph (HRT) is an important diagnostic instrument to detect morphological changes due to glaucoma. The HRT creates depth images of the eye background and calculates geometric parameters that can be used for classification. The Erlangen glaucoma registry is a data base containing longitudinal measurements of several diagnostic instruments of glaucoma patients and healthy controls. Our example data set is taken from the glaucoma registry and consists of 61 HRT variables from $N = 372$ subjects (182 healthy controls, 190 glaucoma patients). The classes are further split in subclasses, namely preperimetric

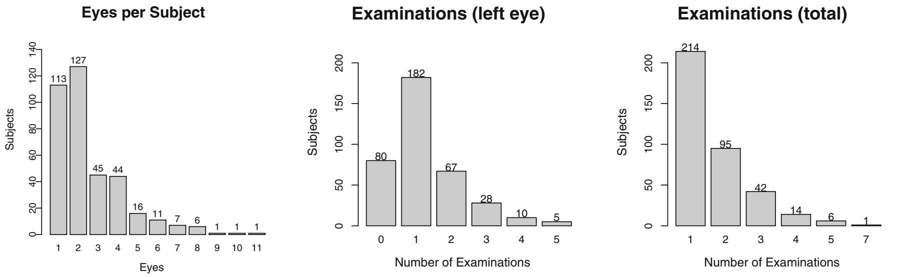

Fig. 2 Number of examinations

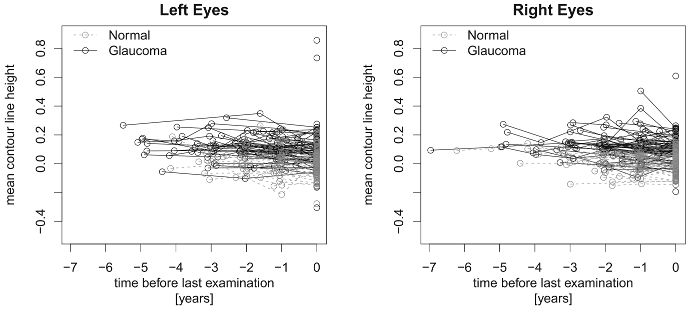

Fig. 3 Longitudinal measurements of a typical HRT parameter (“mean height of contour line”) of normal and glaucomatous left and right eyes: the parameter is plotted against the time to the most recent examination (in years)

and perimetric glaucoma (glaucoma class) and healthy and raised ocular hypertension (normal class). A total of 951 observations stem from 592 repeated measurements of the left or/and right eyes. Both eyes are available from 220 subjects, while from 152 subjects only one (left or right) eye side is available. Class membership is consistent within subjects, i.e. there do not exist observations with different class memberships from one subject. Figure 2 describes the distribution of the examinations per subject and eye and Fig. 3 shows the longitudinal pattern of the HRT parameter mean height of contour line. We will later compare the performance of our proposed strategies to the performance of a classifier trained with a reference data set consisting of only the newest observation of one selected eye per subject. In other words, the reference data set will not show the dependent data structure but consist only of independent observations. Thus, the size N_{ref} of the reference data set is given by the number N of subjects, $N_{ref} = N = 372$.

In a first step we examine the performance of bagged classification trees with B_{bag}=100 trees and a random forest with B_{rf}=1,000 trees using two classes (glaucoma and normal). The classifiers are trained either with the reference data set, with the naive method, i.e. observation based rather than subject based bootstrapping, and with the strategies $\tau \in \{1, \Omega\}$.

In a second step we make use of expert knowledge and use finer classes (pre- and perimetric glaucoma, raised intraocular pressure, and healthy) for training and testing. In the evaluation step, which is done by calculation of the dependent AUC, we again merge the subclasses and evaluate the results for the two main interesting outcomes, healthy or glaucomatous.

The performance estimation is done with the leave-one-out bootstrap with 100 bootstrap samples. The estimated $\widehat{\mathrm{AUC}}_{dep}^{(1)}$ is calculated. This procedure is iterated 10 times.

We did all our calculations using the programming language R v1.9.1 (R Development Core Team 2009). For classification we utilized the packages `ipred` V0.8-8 (Peters et al. 2002) and `randomForest` V4.5-34 (Liaw and Wiener 2002). The random forests were calculated using the default value for the parameter `mtry` which is the square root of the number of variables in the data set (7 in our case). For ROC analysis we used the package `Daim` V1.1.1 (Potapov et al. 2009).

6 Results

6.1 Simulated repeated measurements

The results for random forest trained with the different strategies and data sets for simulation model 1, where the standard deviation of the spirals observations sd equals 0.1 are illustrated in Fig. 4. In the case of $\sigma_{obs} = 0.01$, i.e. the situation with reliable measurements, there is hardly a difference between all strategies for all examined data set sizes. This was expected as the standard deviation of the model is smaller (0.1) and all repeated measurements do not differ by much. The classification performance increases with increasing data set size, as can be expected. Because the repeated measurements do not differ by much, the performance remains nearly constant for an arbitrary number of repeated measurements. As σ_{obs} increases, the superior performance of our proposed strategies becomes obvious. Interestingly, for sd = 0.1, strategy $\tau = \Omega$ performs nearly identical to $\tau = 1$ for $N = 50$. For larger data sets, $\tau = 1$ is slightly superior, especially for $\sigma_{obs} = 0.1$, i.e. the noisiest repeated measurements. There is an increase in classification performance when the number of repeated measurements increases from one to two, while for a larger number of repeated measurements no difference can be seen. The worst performing strategy, especially for $\sigma_{obs} = 0.1$, is to use a reduced data set with one randomly chosen observation for classification. Most prominently in the situation $\sigma_{obs} = 0.1$, and $N = 50$, the strategy where mean values are calculated, performs worse than the naive strategy which for $\sigma_{obs} = 0.1$ and $N = 50$ performs equal to both modified bootstrap strategies and for larger data sets still performs equal to $\tau = \Omega$. The performance of the different strategies with random forest in simulation model 2 (sd = 0.2) is shown in Fig. 5. The main difference compared to simulation model 1 is the overall worse performance and the poor performance of the naive strategy (not correcting for dependent data structure) in the situation $\sigma_{obs} = 0.01$. With increasing σ_{obs} the resemblance to simulation model 1 is closer in so far as the performance of the naive strategy becomes better and the performance of the strategy using the reduced data set becomes worse. Especially for $\sigma_{obs} = 0.1$, our strategy $\tau = 1$ is the best performing method, while $\tau = \Omega$, using mean values and the naive strategy are very close to each other (only for $N = 50$, the strategy using mean values performs slightly worse). Bagging with simulation model 1

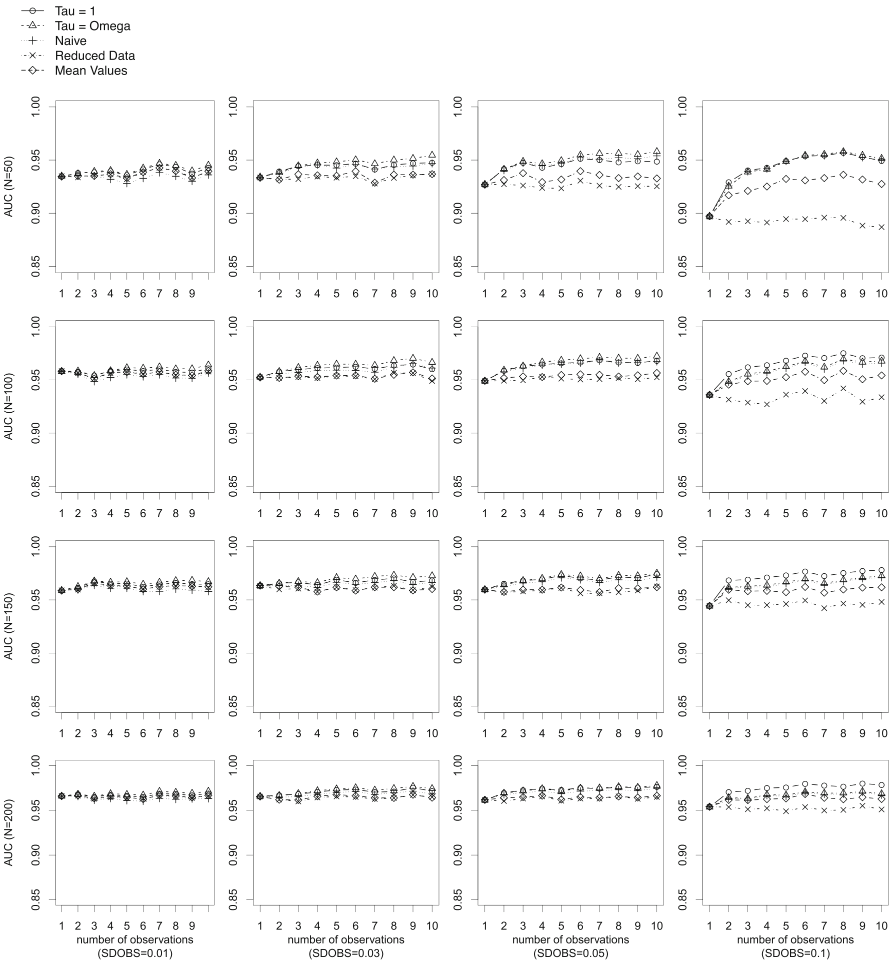

Fig. 4 Classification performance of random forest using different strategies for handling repeated measurements in the simulation study (spirals, sd = 0.1; 100 iterations) measured by AUC: 1 to 9 observations per subject; from *left* to *right*: $\sigma_{obs} = 0.01$, $\sigma_{obs} = 0.03$, $\sigma_{obs} = 0.05$, $\sigma_{obs} = 0.1$. From *top* to *bottom*: 50 subjects, 100 subjects, 150 subjects, and 200 subjects

is illustrated in Fig. 6. Except for the fact that that the naive strategy performs slightly worse than all other methods for $\sigma_{obs} = 0.01$, the performance of all strategies in all situations closely resembles the performance of random forests in simulation model 1. The same holds true in simulation model 2 (see Fig. 7).

6.2 Longitudinal measurements of glaucoma patients

The boxplots shown in Fig. 8 illustrate the classification performance of random forests and bagging with different bootstrap sampling strategies with two classes. The bootstrap estimated AUC (100 bootstrap samples) is simulated 10 times. The differ-

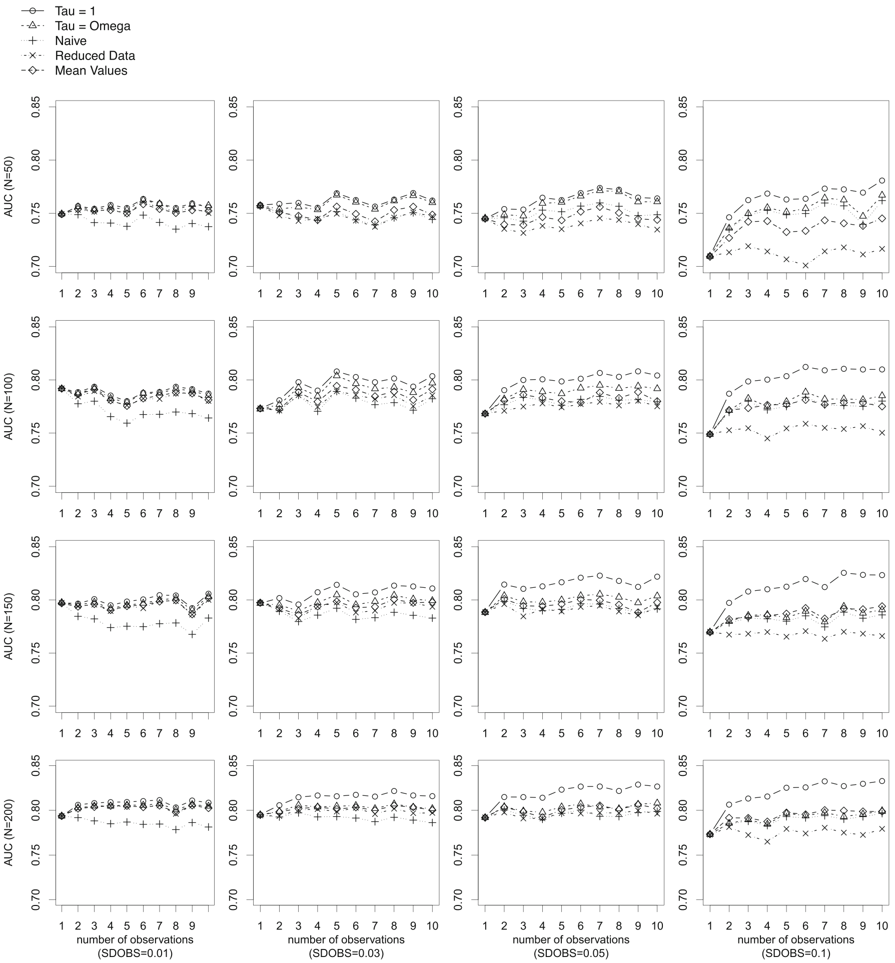

Fig. 5 Classification performance of random forest using different strategies for handling repeated measurements in the simulation study (spirals, sd = 0.2; 100 iterations) measured by AUC: 1 to 9 observations per subject; from *left* to *right*: $\sigma_{obs} = 0.01$, $\sigma_{obs} = 0.03$, $\sigma_{obs} = 0.05$, $\sigma_{obs} = 0.1$. From *top* to *bottom*: 50 subjects, 100 subjects, 150 subjects, and 200 subjects

ence between the strategies is tested by the Wilcoxon signed rank test (Adler et al. 2008). The performance of the naive approach, i.e. bootstrapping without respecting the dependent data structure, serves as the reference for the tests. For random forests, the strategy where one observation per subject is sampled randomly for each tree performs best. For bagging, the strategy where one random observation is used per tree performs comparable to the reference, although the variation of the estimation is reduced. For both classifiers, the case where a reduced data set with only one observation per subject is used for training leads to the worst performance. In the four class scenario which is illustrated in Fig. 9 an overall improvement of the classifier performance compared to the two class case can be observed, especially for bagging.

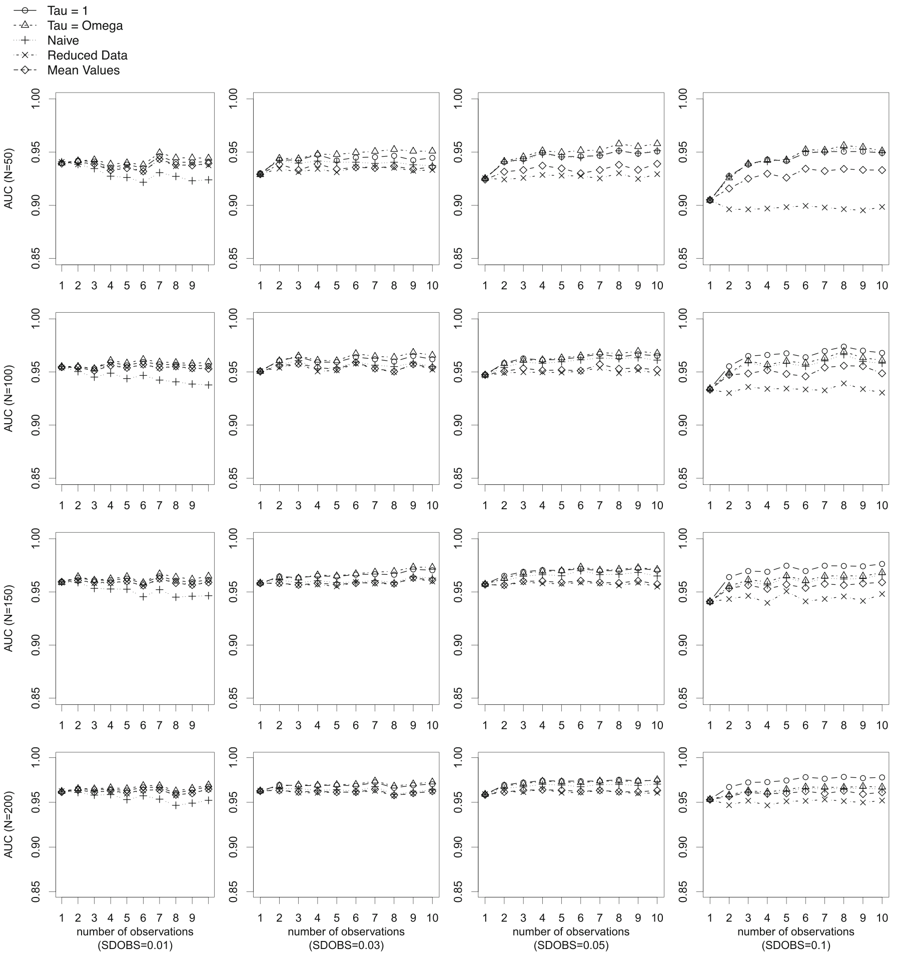

Fig. 6 Classification performance of bagging using different strategies for handling repeated measurements in the simulation study (spirals, sd $= 0.1$; 100 iterations) measured by AUC: 1 to 9 observations per subject; from *left* to *right*: $\sigma_{obs} = 0.01$, $\sigma_{obs} = 0.03$, $\sigma_{obs} = 0.05$, $\sigma_{obs} = 0.1$. From *top* to *bottom*: 50 subjects, 100 subjects, 150 subjects, and 200 subjects

The classifier trained with a reduced learning set performs worst for bagging, while for random forests using all observations in a subject-based bootstrap leads to the poorest performance. As in the two class case, the strategy where one randomly selected observation per subject is used shows the best performance.

The better performance by incorporation of additional classes was expected in our case, as these classes reflect the heterogenity of the data set so that additional classes lead to a more homogenous classification problem. Certainly it is advisable to use all available expert knowledge to simplify the problem but an introduction of new classes is no solution in all cases, if the available measurements do not reflect these classes.

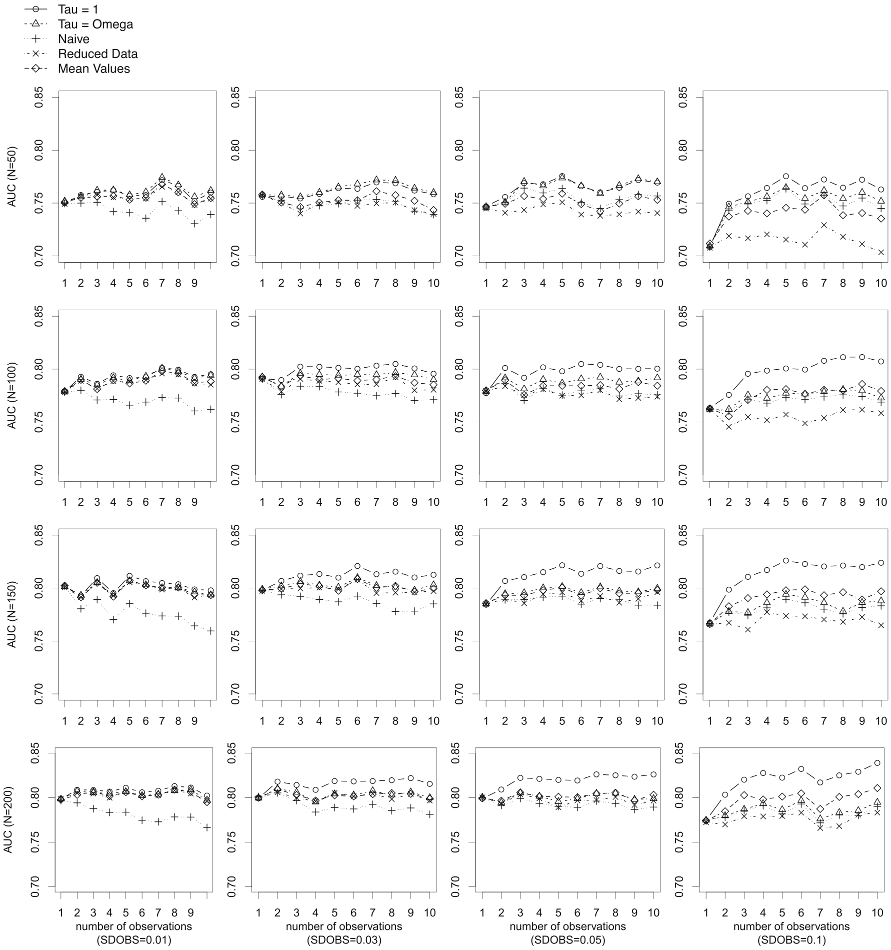

Fig. 7 Classification performance of bagging using different strategies for handling repeated measurements in the simulation study (spirals, sd = 0.2; 100 iterations) measured by AUC: 1 to 9 observations per subject; from *left* to *right*: $\sigma_{obs} = 0.01$, $\sigma_{obs} = 0.03$, $\sigma_{obs} = 0.05$, $\sigma_{obs} = 0.1$. From *top* to *bottom*: 50 subjects, 100 subjects, 150 subjects, and 200 subjects

7 Discussion

One way to account for longitudinal or paired data structure in classification problems is to use parametric mixed effects logistic regression models, e.g. (Morgan et al. 2009). These models have the disadvantage of posing assumptions on the distribution of the data which do not have to be fulfilled in reality. The same is true for GEE, e.g. proposed for regression analysis of eyes (Martus et al. 2004). Nonparametric tree based methods are more general as they do not need any underlying assumptions on the data. An improvement to single trees is given by tree ensembles; well known are bagged classification trees, random forests, and adaboost. For bagged trees and random forests the bootstrap samples that are used to train the single classification trees are independent

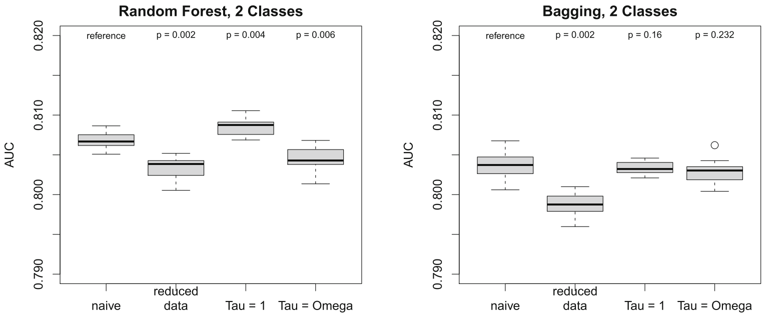

Fig. 8 Bootstrap estimated areas under the ROC curve (AUC) of random forest and bagging with two classes for the different strategies

of each other. This allows for an easy implementation of our strategies. A possible generalization of our proposal for boosting could be to implement the weights used for the learning sets on a subject basis. An increased/decreased weight could be realized by the inclusion of more/less follow-up examinations of the subject into the next iteration step.

In our study ignoring the dependency structure and using all observations for training led to a better performance than using only the newest observations of one eye. This could be due to the fact that the increased similarity between trees can be neglected compared to the improved performance by a larger training sample. Further examination in this direction could clarify this issue.

Recently, the method RF++ was proposed by Karpievitch et al. (2009). RF++ performs subject based rather than observation based bootstrapping and selects all observations per subject to create learning data for the single trees. Thus, the correlation between replicated measurements of the same subject is respected. The subject based bootstrap leads to trees which have observations of less subjects in common than in the naive case, where bootstrapping is performed on the observation level.

Our results demonstrate that sampling one observation is better compared to sampling all observations of each subject. The rationale for this approach is: when only one observation per subject is selected, likely different observations are used for the training of different trees, although the same subjects might be selected which further reduces similarity between trees. Thus, our approach incorporates advantages of both methods, subject based bootstrapping with all observations per subject and observation based bootstrapping with only one selected observation per subject: the total ensemble makes use of all available data in contrast to the latter case. This means that the information of all observations is represented in the ensemble. Faster training is possible as less observations are used in a single tree and as stated above, the similarity between the trees is further reduced. Moreover, our approach can be applied to unbalanced repeated measurement data.

A further aspect that is worth future research is the application of our strategy to subagging (Buehlmann and Yu 2002). The strategy to sample one per block is the same for sampling with or without replacement. In this study, it was not our aim to investigate, if bagging or subagging provides a better solution for the spirals problem or the glaucoma data set.

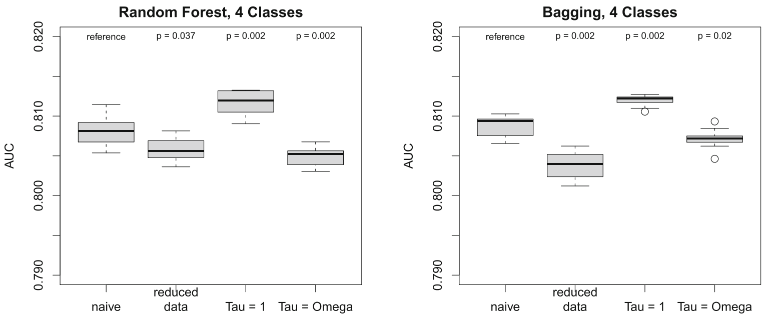

Fig. 9 Bootstrap estimated areas under the ROC curve (AUC) of random forest and bagging with four classes for the different strategies

Acknowledgements This research was partially funded by SFB, Grant 539A4/C1.

References

Adler W, Lausen B (2009) Bootstrap estimated sensitivities, specificities and roc curve. Comput Stat Data Anal 53(3):718–729

Adler W, Peters A, Lausen B (2008) Comparison of classifiers applied to confocal scanning laser ophthalmoscopy data. Methods Inf Med 47(1):38–46

Adler W, Brenning A, Potapov S, Schmid M, Lausen B (2011) Ensemble classification of paired data. Comput Stat Data Anal 55(5):1933–1941

Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, California

Brenning A, Lausen B (2008) Estimating error rates in the classification of paired organs. Stat Med 27(22):4515–4531

Buehlmann P, Yu B (2002) Analyzing bagging. Ann Stat 30:927–961

Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning. Springer, New York

Karpievitch YV, Hill EG, Leclerc AP, Dabney AR, Almeida JS (2009) An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. PLoS ONE 4(9):e7087

Lang KJ, Witbrock MJ (1988) Learning to tell two spirals apart. In: Touretzky D, Hinton G, Sejnowski T (eds) Proceedings of the connectionist models summer school, Morgan Kaufmann, Mountain View, CA, pp 52–59

Leisch F, Dimitriadou E (2010) mlbench: machine learning benchmark problems. R package version 2.0-0

Liaw A, Wiener M (2002) Classification and regression by random forest. R News 2(3):18–22. http://CRAN.R-project.org/doc/Rnews/

Martus P, Stroux A, Juenemann AM, Korth M, Jonas JB, Horn FK, Ziegler A (2004) GEE approaches to marginal regression models for medical diagnostic tests. Stat Med 23:1377–1398

Morgan WH, Hazelton ML, Balaratnasingamm C, Chan H, House PH, Barry CJ, Cringle SJ, Yu D (2009) The association between retinal vein ophthalmodynamometric force change and optic disc excavation. British J Ophthalmol 93:594–596

Peters A, Hothorn T, Lausen B (2002) ipred: Improved predictors. R News 2(2):33–36. http://CRAN.R-project.org/doc/Rnews/

Potapov S, Adler W, Lausen B (2009) R package daim—R package version 1.0.0

Quigley HA (1996) Number of people with glaucoma worldwide. British J Ophthalmol 80:389–393

R Development Core Team (2009) R: A language and environment for statistical computing. R Found Stat Comput. http://www.R-project.org