BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2563–2570
doi:10.1093/bioinformatics/btr434

Gene expression

Advance Access publication July 22, 2011

Estimating classification probabilities in high-dimensional diagnostic studies

Inka J. Appel*, Wolfram Gronwald and Rainer Spang

Institute of Functional Genomics, University of Regensburg, 93053 Regensburg, Germany

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Classification algorithms for high-dimensional biological data like gene expression profiles or metabolomic fingerprints are typically evaluated by the number of misclassifications across a test dataset. However, to judge the classification of a single case in the context of clinical diagnosis, we need to assess the uncertainties associated with that individual case rather than the average accuracy across many cases. Reliability of individual classifications can be expressed in terms of class probabilities. While classification algorithms are a well-developed area of research, the estimation of class probabilities is considerably less progressed in biology, with only a few classification algorithms that provide estimated class probabilities.

Results: We compared several probability estimators in the context of classification of metabolomics profiles. Evaluation criteria included sparseness biases, calibration of the estimator, the variance of the estimator and its performance in identifying highly reliable classifications. We observed that several of them display artifacts that compromise their use in practice. Classification probabilities based on a combination of local cross-validation error rates and monotone regression prove superior in metabolomic profiling.

Availability: The source code written in R is freely available at http://compdiag.uni-regensburg.de/software/probEstimation.shtml.

Contact: inka.appel@klinik.uni-regensburg.de

Received on May 27, 2011; revised on July 13, 2011; accepted on July 19, 2011

1 INTRODUCTION

Diagnosis, prognosis and prediction of treatment response based on transcriptomic, proteomic or metabolomic profiles is a well developed field (Michiels *et al*., 2007; Sotiriou and Piccart, 2007). A plethora of classification algorithms have been proposed and critically compared (Fan *et al*., 2010; MAQC Consortium, 2010; Zervakis *et al*., 2009). It very much depends on the classification problem at hand, whether an almost error-free classifier can be developed or whether classification errors are unavoidable regardless of what algorithm is chosen.

In the latter case, it is natural that a clinician asks for the reliability of an individual diagnosis before moving on to treatment decisions. Classification algorithms are typically evaluated by the frequency of misclassifications in cross-validation or on an independent test set. These performances are averages over many predicted cases. They say little about the reliability of an individuals diagnosis. The case might be easier or more difficult to diagnose than the average in the test set. For each case, every class is assigned a value $p_j \in [0,1]$, which is an estimated probability that the case belongs to that class, given the profiling data.

In microarray-based classification, the performance of classification algorithms has been analyzed and compared in great detail (Dudoit *et al*., 2002; Wessels *et al*., 2005). However, little attention has been given to the usefulness of probability estimates and this is even more true for metabolomic analyses. In fact, only relatively few classification algorithms estimate class probabilities and in the majority of clinical papers on the performance of classifiers, case-specific probabilities are not shown. A class probability estimator is most useful, if it flags incorrect classifications as low confidence classifications. In other words: if a classifier produces confident class probabilities close to one, these should be correct classifications.

In this article, we compare class probability estimators in the context of high-dimensional data-based diagnosis. We briefly review a selection of class probability estimators including those that are most frequently used in the context of gene expression analysis like Naive Bayes estimators or binary regression. In addition, we discuss alternative approaches from different fields of application like text categorization and digit recognition and adapt them to metabolomics analysis. We complement the pool of methods by a novel approach based on smooth local error rates. The approaches are compared on a recently published metabolomics dataset of patients with various types of kidney disease.

We found that artifacts can compromise the utility of some frequently used methods. A widely observed problem is the dependence of classification probabilities on the number of features used in a diagnostic signature. The more features are used by a classifier, the more confident the classification probabilities, even in cases where the classification is incorrect.

Moreover, class probabilities need to be estimated from test data or cross-validated classification scores since training scores display a better but unrealistic separation of classes. This overfitting phenomenon can greatly affect class probabilities. In our comparative metabolomics study, class probabilities derived from local error rates proved to be the method of choice.

2 CLASS PROBABILITY ESTIMATORS

We first review a selection of class probability estimators. In order to separate the class probability estimation problem from the classifier learning problem, we compare methods based on the same type of classifier. We chose a sparse linear classifier, which has been shown to be among the best-performing

*To whom correspondence should be addressed.

algorithms in microarray two-class classification problems (Wessels *et al.*, 2005).

Notations: let x_{ij} be a data matrix with $i=1,2,\dots,m$ denoting features (metabolites) and $j=1,2,\dots,n$ denoting cases. Further let C_k be a vector storing the true class membership (disease types) of cases.

The end product of all linear classification algorithms is a classification that assigns a case to Class 1 if $s(x)<0$ and otherwise to Class 2, where $x=(x_1,x_2,\dots,x_p)$ is the vector of intensity values of p signature features, $w=(w_1,w_2,\dots,w_p)$ a vector of corresponding feature weights and $s(x)=\langle w,x\rangle - b$ with distance to the origin b. Note that the vector w spans a line orthogonal to the separating hyperplane, $\langle w,x\rangle$ is the orthogonal projection of profile x onto this line and $s(x)$ the distance of x to the hyperplane. Intuitively, cases that fall closer to the separating hyperplane are less reliably classified than those that are further away.

Naive Bayes Estimates (NB): Tibshirani *et al.* (2002) proposed a method for class prediction in DNA microarray studies based on nearest shrunken centroids. Each class k is represented by a shrunken centroid $\bar{x}_k$ and a case x is assigned to the class with the nearest centroid. Formally, the discriminant score

$$\delta_k(x)=\sum_{i=1}^{p}\frac{(x_i-\bar{x}_{ik})^2}{\sigma_i^2}-2\cdot log\pi_k$$

is calculated independently for each class, where the sum runs over all genes with non-zero weights after shrinkage Δ, σ_i is the pooled within-class SD of gene i and π_k the estimated proportion of cases from class k in the entire population. A case is assigned to the group k with minimal $\delta_k(x)$. Classification probabilities are derived from the $\delta_k(x)$ by

$$p_k(x)=\frac{e^{-\frac{1}{2}\delta_k(x)}}{e^{-\frac{1}{2}\cdot\delta_1(x)}+e^{-\frac{1}{2}\cdot\delta_2(x)}}. \tag{1}$$

Note that the nearest shrunken centroid classification rule defines a separating hyperplane with normal vector $w_i=\frac{2\cdot(\bar{x}_{i1}-\bar{x}_{i2})}{\sigma_i^2}$, and Equation (1) can be translated to

$$p_k(x)=\frac{1}{1+e^{-\frac{1}{2}\cdot s(x)}}.$$

In this approach, every gene in the classifier is assumed to contribute independent evidence as to whether the case x belongs to class k or not. This assumption is mostly not justified biologically and produces artifacts that will be discussed in Section 3.

Compound Bayes Estimates (CB): Wright *et al.* (2003) introduced Compound Bayes Estimates to microarray analysis. In line with the PAM approach, the CB estimator models both classes individually using normal distributions. However, unlike the PAM approach the CB estimator models the 1D distributions of the projected data $s(x_j)$ instead of the multidimensional distributions of the original data x_j. Given a separating hyperplane with normal vector w and associated classification scores $s(x_j)=\langle w,x_j\rangle - b$, one assumes that the $s(x_j)$ are distributed normally in both classes with possibly different means μ_k and SDs σ_k. Bayes rule yields

$$p_k(x)=\frac{\phi_1(s(x);\hat{\mu}_1,\hat{\sigma}_1^2)}{\phi_1(s(x);\hat{\mu}_1,\hat{\sigma}_1^2)+\phi_2(s(x);\hat{\mu}_2,\hat{\sigma}_2^2)}$$

where $\phi(x;\mu,\sigma^2)$ represents the normal density function with mean μ and variance σ^2. The four parameters $\hat{\mu}_1,\hat{\mu}_2,\hat{\sigma}_1,\hat{\sigma}_2$ are estimated from the projected data $s(x_j)$.

Binary Regression (BReg): another well-established approach is binary regression, which has many ramifications some of which have been applied to genomic data (de Hoon *et al.*, 2004; West *et al.*, 2001). In our evaluation here, we represent the class of binary regression models by the approach described in Platt (2000), which fits the logistic model

$$p_k(s(x))=\frac{1}{1+e^{A\cdot s(x)+B}}$$

by minimizing a cross-entropy error function to adjust the parameters A and B. Although Platt (2000) use this estimator in combination with linear and non-linear support vector machines, it can also be used together with our linear classifiers without changes, since the regression simply operates on a set of precalculated classification scores $s(x_j)$ without exploiting any properties implied by the method that generated these scores.

Local Error Frequencys [LEF(Bin)]: if the shape of the regression function that relates classification scores to class probabilities is unknown, it can be estimated from local misclassification frequencies. Zadrozny and Elkan (2002) sort cases by the scores $s(x_j)$ and split them into equally sized disjoint bins. The local class k frequency $F_k(x)$ of case x is then calculated as the relative frequency of class k cases that fall into the same bin as x. The estimates $F_k(x_j)$ do not need to be strictly monotonous in $s(x_j)$. Hence, a few cases that are closer to the separating hyperplane might be judged more reliably classified than some of those that are further away from it, in contrast to intuition.

To assure monotonicity, Zadrozny and Elkan (2002) use monotone regression as implemented in the pair-adjacent violators algorithm (PAVA) (Ayer *et al.*, 1955). PAVA replaces $F_k(x_j)$ and $F_k(x_{j+1})$ with their average when the monotonicity constraint is violated. This averaging process is continued until an ordered set of probabilities is obtained. Note that local misclassification frequencies do not need to be combined with the PAVA algorithm but can be combined with any regression method. As with previous methods, it needs to be noted that the original work by Zadrozny and Elkan (2002) used the estimator in combination with a different classification algorithm (decision trees). However, since this method is also based only on post-processing classification scores, it can be used in the linear classifier context as well.

Smooth Local Error Frequencies [LEF(Smooth)]: a drawback of the binning approach is its coarseness. All cases in a bin receive the same local error frequency estimate $F_k(x)$ regardless of where they fall in the bin, and scores in the same bin can vary substantially. While the monotone regression step is partly compensating for this artefact, we argue that it cannot fully adjust for the binning effect and more smooth estimates of $F_k(x_j)$ are needed. Next we will describe two modifications of the LEF concept that combine monotone regression with smooth estimates of local error frequencies. We propose to use Gaussian smoothing kernels

$$K(s(x_j))=\frac{1}{\sqrt{2\pi\lambda^2}}\exp\left(-\frac{(s(x_j)-s(x))^2}{2\lambda^2}\right)$$

to estimate the local error frequencies $F_k(x)$ by

$$F_k(x)=\frac{\sum_{j\in C_k}K(s(x_j))}{\sum_j K(s(x_j))}.$$

The bandwidth λ is the same for all cases x_j (one fits all approach). It is a tuning parameter whose calibration will be discussed in the next section. Note that the kernels are centered at $s(x)$ and that the classification accuracy of the algorithm for cases with scores similar to $s(x)$ still determines the local error frequencies. Different to the binning approach, the actual distances of neighboring cases are now taken into account. Once the local error rates are estimated, we proceed like in the binning method of Zadrozny and Elkan (2002). We use monotone regression on the $F_k(x_j)$ employing the PAVA algorithm to achieve class probabilities.

Adaptive Local Error Frequencies [LEF(Adapt)]: this constant λ assumption is problematic, if the density of scores $s(x_j)$ is far from uniform. For these situations, we propose an adaptive estimator of $F_k(x_j)$. We propose to use the neighborhood adaptive Gaussian smoothing kernels

$$K_{x_j,l}(s(x_j))=\frac{1}{\sqrt{2\pi\lambda(x,l)^2}}\exp\left(-\frac{(s(x_j)-s(x))^2}{2\lambda(x,l)^2}\right).$$

Note that the kernels are centered at $s(x)$ and that their bandwidths depend on a tuning parameter l and vary across cases. The bandwidth λ is adapted to the local density of scores around $s(x)$. It is narrower in regions where we have many cases with similar scores $s(x_j)$. We achieve this by setting $\lambda(x,l)$

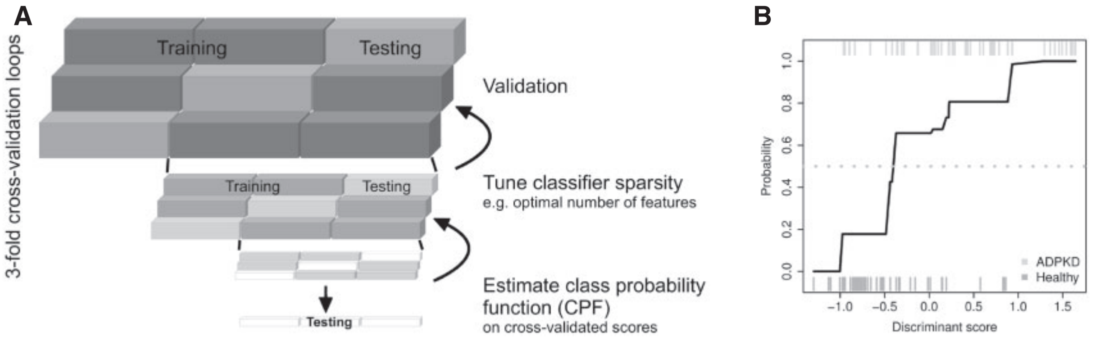

Fig. 1. (**A**) Schematically shows the 3-fold nested cross-validation. In the innermost loop, the class probability function CPF is optimized, in the next loop classifier sparsity is tuned and in the outermost loop validation is performed. The different loops are marked by the size of the corresponding bars. Note that in the Figure, each loop is further separated into different rows to indicate that data present in that loop are split several times in training and test data to ensure that all data of a specific loop are used once for testing. The outermost loop contains all data, of this the training data of the outermost loop are passed to the middle loop where the passed data are split again in training and test data and the training data of the middle loop are transferred to the innermost loop where again a splitting in training and test data is performed. (**B**) Shows estimated classification probabilities for cross-validated scores for method LEF(Adapt) exemplarily. The stripes on the x-axis show the cross-validated classification scores $s(x_j)$ from ADPKD patients (top) and healthy donors (bottom).

equal to the empirical variance of the l nearest neighbors of $s(x)$. We will discuss tuning of the parameter l in the next section. The adaptation of the kernel bandwidths ensures that the local error frequencies are supported by roughly the same number of neighboring cases.

Training and validating class probability functions: the end product of all methods is always an estimated class probability function (CPF) $p_k(s(x))$ that maps the score $s(x)$ to a number between 0 and 1 that we interpret as the probability that case x is in class k. The CPF needs to be trained on sets of classification scores. Training includes the estimation of distribution parameters like the class means and variances in the Compound Bayes Estimator, the regression parameters A and B of the binary regression approach or the local error frequencies $F_k(x_j)$ and the non-parametric regression curves resulting from monotone regression. In addition, training might depend on tuning parameters like the number of bins, the bandwidth of a Gaussian kernel, or the number of nearest neighbors in the adaptive LEF approach. Once a CPF is estimated, it can be used to predict classification probabilities of test cases x_j' by first calculating the scores $s(x_j')$ using a linear classifier and then plugging them into a CPF to gain the class probability $p_k(s(x_j'))$. This requires a CPF and a classifier and both need to be previously learned from data. It is important that the test cases x_j' were not included in any of these learning processes.

With respect to CPF estimation, it is important to distinguish between training scores $s(x_j)$ and test scores $s(x_j')$, since it is known that their distributions can be greatly different (Ambroise and McLachlan, 2002). Compared to test scores, training scores display a better but unrealistic separation of classes. This overfitting phenomenon can greatly affect the estimated CPFs as we will show in the next section. Here, we use a 3-fold nested cross-validation that covers the processes of classifier estimation, CPF estimation, parameter tuning and evaluation. Parameters that need to be calibrated include the tuning parameters of the local error frequency approaches, which we call Θ and the shrinkage parameter Δ of the nearest shrunken centroid classification algorithm that controls the sparsity of the classifier.

(1) CPF estimation
In the most inner loop Δ is fixed. N_1 cases are left out and the remaining cases are used to learn a classifier, which is applied to the leftout cases yielding scores $s_\Delta(x_j)$. By leaving out all cases in turn, we achieve cross-validated scores for all cases that entered the most inner cross-validation loop. A CPF is estimated from these scores for a variety of values of the parameter Θ. For each value, the $p_k^{\Delta,\Theta}(x_j)$ are computed using one of the methods described above. They are evaluated with respect to their classification performance by calculating the negative log-likelihood of true classes

$$-\log(L(\Theta)) = -\sum_k \sum_{j\in C_k} \log p_k^{\Delta,\Theta}(x_j) \tag{2}$$

The Θ with minimal $-\log(L)$ is chosen and the corresponding CPF $p_k^{\Delta}(\cdot)$ is returned to the middle cross-validation loop.

(2) Tuning classifier sparsity
In the middle loop, N_2 cases are left out. The remaining cases are forwarded to the inner loop varying Δ. For every Δ, the inner loop returns a CPF $p_k^{\Delta}(\cdot)$ which is applied to the leftout cases of the middle loop. These are evaluated by their misclassification rate and the optimal value of Δ and with it the optimal number of features is determined. The optimized CPF $p_k(\cdot)$ is returned to the outer loop.

(3) Validation
In the outer loop N_3 cases are left out. The remaining cases are forwarded to the middle loop, which returns a CPF $p_k(\cdot)$ to be applied to the leftout cases. Finally, this leaves a set of cross-validated probabilities $p_k(x_j)$ which we will next evaluate with respect to different criteria.

Our design allows different CPF estimation procedures to use different numbers of features for classification. This is enabled by the middle cross-validation loop. This is important since some CPF estimators for instance the PAM estimator are sensitive to the number of features. Our entire cross-validation design is summarized in Figure 1. For data used here, we set $N_1 = 1$, and $N_2 = N_3 = 2$.

3 RESULTS

We compared the class probability estimators in the context of a recently published metabolomic profiling study on kidney diseases (Gronwald *et al*., 2011). The dataset comprised 168 urine samples measured using 1D nuclear magnetic resonance (NMR) spectroscopy. 54 samples were obtained from patients with autosomal polycystic kidney disease. The challenge is to separate them from samples taken from healthy volunteers (46 samples), and samples from patients with compromised kidney function but no ADPKD (52 samples from diabetes mellitus patients and 16 samples

Table 1. Patient groups defined within the ADPKD dataset

Description	Index	Size
ADPKD	1	54
ADPKD with medication	1A	35
ADPKD without medication	1B	19
Healthy	2	46
Other CKD		
Renal transplant without rejection	3	16
Diabetes with microalbuminuria	4	30
Diabetes without microalbuminuria	5	22

Groups 1A and 1B correspond to ADPKD patients with and without medication for arterial hypertension, Group 2 consists of healthy volunteers. Patients 3 months after renal transplantation without rejection are assigned to Group 3 and diabetes mellitus type 2 patients with and without microalbuminuria are in Groups 4 and 5, respectively.

from patients 3 months after renal transplantation). More details on the composition of cases in the study can be found in Table 1. NMR 1D spectra were split into 701 equally sized buckets and globally normalized to the signal of the CH_2 group of creatinine to ensure sample to sample comparability. Furthermore, compatibility across metabolites was ensured by applying the glog transformation (Parsons *et al*., 2007). For full details of sample preparation and data preprocessing see Gronwald *et al*. (2011).

We learned a shrunken centroid classifier aiming at the separation of ADPKD patients and healthy controls. Across all patients, we observe a classification performance of 76% correct classification in cross-validation. Figure 1B resolves this classification performance further. The ticks on the x-axis show the cross-validated classification scores $s(x_j)$ from ADPKD patients (top) and healthy donors (bottom). In line with the global performance of only 76%, one can observe that there is no perfect separation of the two groups. Scores between -0.96 and $+0.86$ can be observed in both classes. The separating hyperplane lies in the middle of these points. It assigns 41 of them to the ADPKD class and 40 to the healthy donor class. Confined to this range of scores, the classifier has a performance of 71% correct classifications. These diagnoses should be flagged ‘unreliable’. However, scores >0.86 are only found among ADPKD patients and hence reliably indicate that a patient suffers from ADPKD. Identifying this group of patients boils down to estimating ADPKD probabilities for all patients. The y-axis exemplarily shows such estimates obtained using a local error frequency approach [LEF(Adapt)]. All patients with a score >0.93 receive probabilities close to one indicating their reliable classification as ADPKD positives. Patients with scores between -0.97 and $+0.22$ obtain probabilities between 0.2 and 0.8 flagging them as problematic classifications.

We next compare the six CPF estimators, Naive Bayes (NB), Compound Bayes (CB), binary regression (BReg) and the local error frequency methods (LEF(Bin), LEF(Smooth), LEF(Adapt)), with respect to several criteria including modification of classification performance, sparseness bias, calibration and the performance in identifying reliable classifications.

3.1 Modification of classification performance

Qualitative classifications can be obtained directly from the linear classifier. No CPF estimation is necessary. Nevertheless, once a CPF is estimated it is natural to assign cases with a class probability $p_k(x_j)$ >0.5 to class k and those with probabilities <0.5 to the other class. This might lead to a reassignment of some cases, as in the example shown in Figure 1 where 14 cases were reassigned.

Table 2. Classification performances of the outer cross-validation loop for the six probability estimation methods, Naive Bayes (NB), Compound Bayes (CB), binary regression (BReg) and the local error frequency methods using binning (Bin) and smoothing (Smooth/Adapt)

Group comparison		Classification performance					
		NB	CB	BReg	Bin	Smooth	Adapt
1	2	76.0	75.0	76.0	71.0	74.0	76.0
1	3	97.1	95.7	92.9	97.1	97.1	98.6
1	4	82.1	82.1	85.7	88.1	85.7	85.7
1	5	82.9	88.2	86.8	85.5	84.2	85.5
1	2,3,4,5	75.0	78.6	76.8	76.2	76.8	76.2
1A	2	86.4	86.4	86.4	85.2	80.2	84.0
1B	2	63.1	60.0	66.2	64.7	64.6	66.2
Average		80.4	80.9	81.5	81.1	80.4	81.7

For each patient group, comparison (rows) performances are listed for each estimation method. The assignment of patient groups to the indices can be found in Table 1.

Table 2 shows the global classification performance in several pairwise classifications using the plain shrunken centroid classifier [equal to the Naive Bayes (NB) estimator] and its modifications resulting from the different CPF estimators. For group sizes, see Table 1.

The local error frequency method LEF(Adapt) reaches the highest average performance of 81.7%, followed by binary regression (BReg) and LEF(Bin) with 81.5 and 81.1%, respectively. Overall, the classification accuracies of all CPF estimators differ 3.6–5.7% depending on the pair of groups. Table 2 also shows that the performance of a given method depends on the investigated pair of groups. Therein, CB, BReg and LEF(Adapt) won most frequently.

3.2 Sparseness bias

Estimated class probabilities should not depend on the number of features used by the classifier except for reflecting shifts in the overlaps of scores. Figure 2 shows discriminant scores and a PAM-based CPF (NB) of a 5 feature and a 200 feature classifier. In order to ensure comparability, discriminant scores are standardized to mean zero and SD one. Both refer to a comparison of ADPKD patients with healthy donors. One can observe that the overlap of scores only marginally changes for the two classifiers. However, the estimated CPFs change dramatically. For the 5 feature curve, all patients receive probabilities between 0.35 and 0.70 flagging them all as unreliable. This result does not reflect the score distributions well. In contrast, for the 200 features curve all probabilities are either close to 0 or close to 1. Hence, the curve considers all classifications reliable, which is misleading given the mix of classes in cases with scores between -1.1 and $+1$. The PAM estimator (NB) has an obvious sparseness bias. The more features are included in a classifier, the more confident the estimator becomes regardless of how the classes overlap. We next investigate to which extent the individual methods suffer from such a sparseness bias.

Figure 3 shows scatter plots of estimated probabilities for classifiers including different numbers of metabolites. For each number of features, the class probabilities of all cases are subtracted by those obtained for the same patient by the 2 feature classifier. The differences (y-axis) are plotted against the number of features (x-axis, logarithmic). The density of points in the scatter plots is coded on a gray scale with dark regions indicating high density. The first plot clearly shows the sparseness bias of the PAM approach (NB). The two gene classifier gives probabilities ∼0.5. This is kept up for classifiers up to 10 features. Classifiers with many features produce probabilities near 0 or 1 leading to differences of ±0.5. This effect becomes manifest for classifiers with 75 features or more. None of the other methods showed this behavior.

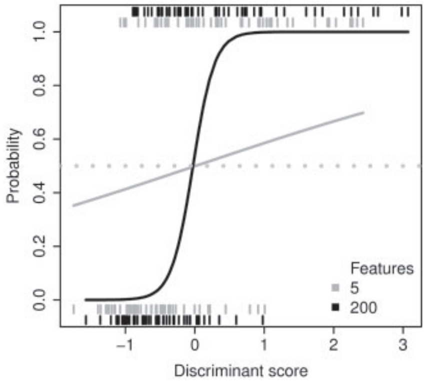

Fig. 2. Scores of a 5 feature classifier (gray) and a 200 feature estimator (black) together with a CPF estimated by the PAM approach are shown. Scores on the x-axis were standardized to mean zero and SD 1 for comparing CPFs directly. Cross-validated scores are indicated as gray and black stripes for the comparison of ADPKD patients (top) and healthy controls (bottom).

Although differences of class probabilities can reach high values, there is no systematic sparseness bias observable. For the Compound Bayes estimator and the binary regression estimator, the majority of differences stay close to zero. For the local error rate-based methods, the differences are greater but also here we do not observe a systematic trend towards more self-confident probabilities when more features are included.

3.3 Calibration

A straightforward criterion to evaluate probability estimators is calibration. An estimator is well calibrated, if in the long run the relative frequency of true classifications of cases with estimated class probabilities falling in a small interval $[p_0-\epsilon, p_0+\epsilon]$ is close to the estimated probability p_0 (Dawid, 1982). The ADPKD dataset is not large enough to test long run performance. That is why we simulated data with structures similar to the ADPKD data for the comparison of the ADPKD patients and healthy controls. Therein, pairs of samples (x_1, x_2) were drawn randomly from the original dataset and sample x_1 was shifted on the line spanned by samples x_1 and x_2 such that $x_1' = x_1 + \beta \cdot (x_2 - x_1)$ where $\beta \in [-1, 1]$. This procedure was repeated, for each class separately, until the simulated dataset was of the same size as the original one. In this way, 30 datasets were simulated and compared in terms of calibration.

Figure 4A compares estimated class probabilities to long run classification accuracies. Both estimated error probabilities and observed error frequencies were collected in bins of width 0.1. If the classifier is well calibrated, all points fall close to the $x = y$ line. Root mean square errors (RMSE) from this line quantify the calibration of the estimator. We found that the RMSE was large for the Compound Bayes estimator and for binary regression while

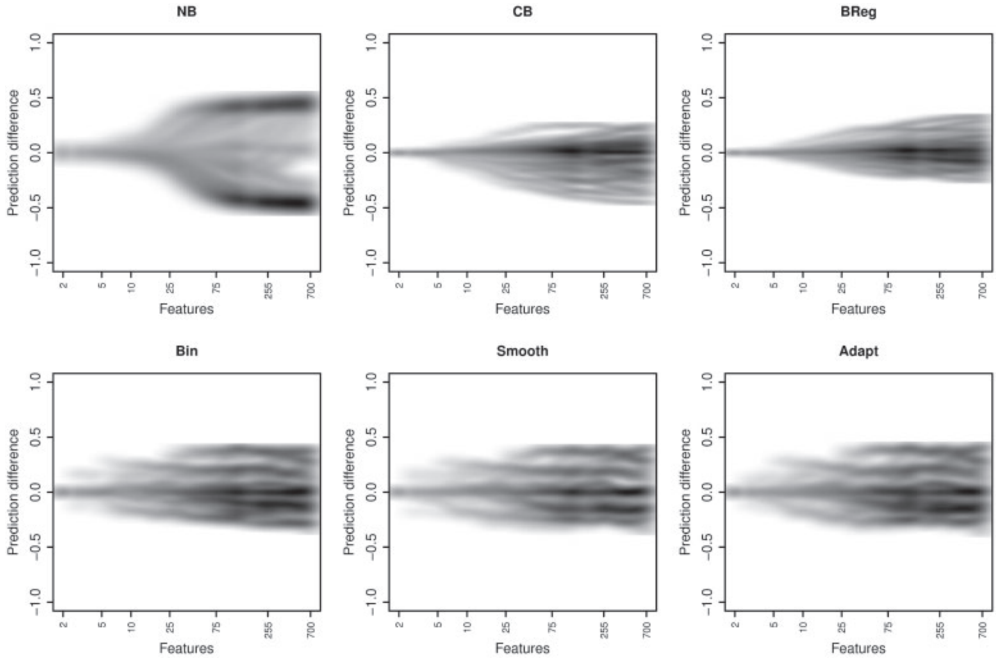

Fig. 3. Smoothed scatter plots of estimated probabilities for classifiers including different numbers of metabolites. For each patient, the class probability of the 2 feature classifier was subtracted from probabilities for all numbers of features for the given patient. The differences (y-axis) are plotted against the number of features (x-axis, logarithmic). The density of points in the scatter plots is coded on a gray scale with dark regions indicating high density.

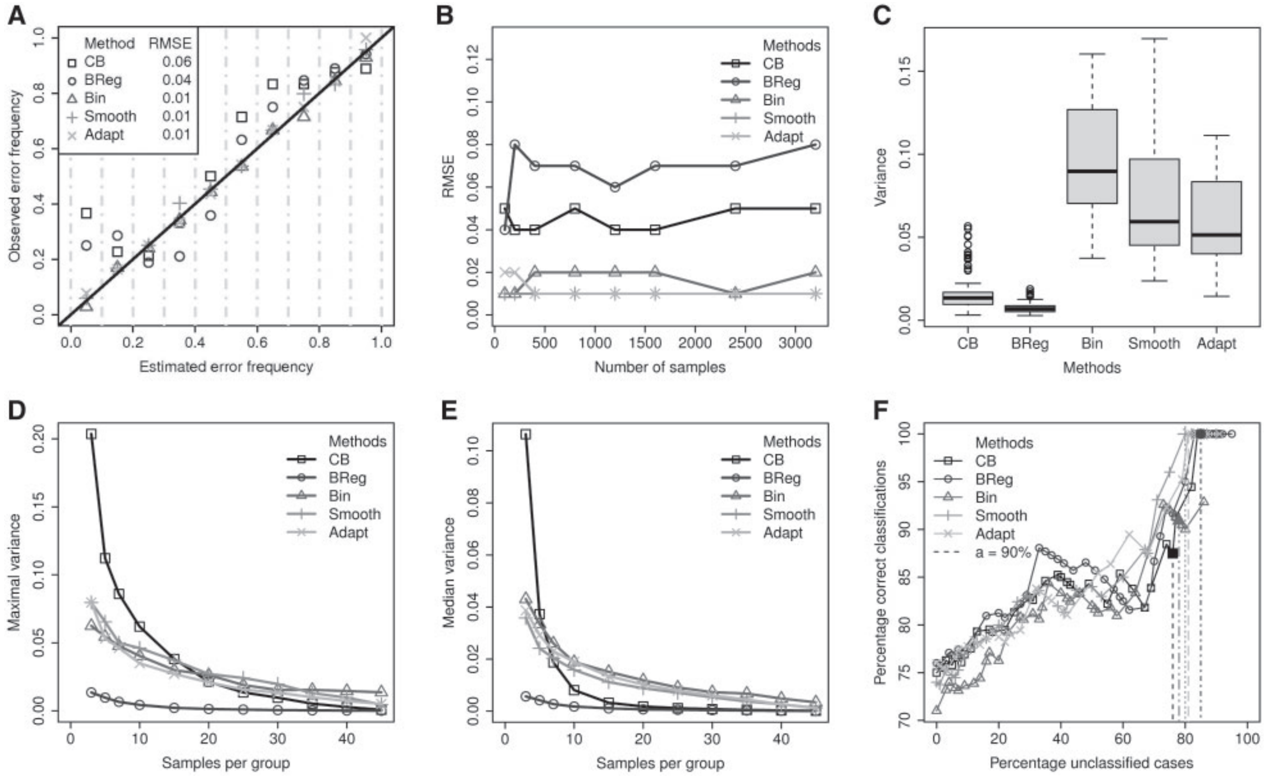

Fig. 4. (**A**) Estimated class probabilities to long run classification accuracies are shown in a reliability diagram. Both estimated error probabilities (x-axis) and observed error frequencies (y-axis) were collected in bins of width 0.1. If the classifier is well calibrated, all points fall close to the $x=y$ line. Root mean square errors (RMSE) from this line quantify the calibration of the estimator. (**B**) Calibration of CPFs with increasing sample size. To test whether calibration depends on the number of samples used for estimator training, the evaluation of calibration was repeated for sample sizes between 100 and 3200 and RMSEs were plotted for all CPFs with increasing sample size. (**C**) Variance of probability estimators across multiple simulation runs. Variances were calculated across cross-validation folds and simulation runs while the box plots displays the distribution of these variances across cases. (**D**) and (**E**) Dependency of the variance of estimated probabilities on the size of the study. Shown are the maximal (D) and median (E) variance across cases. (**F**) Reliability of classifications with increasing confidence level α. α is varied from 0.5 to 1 estimated probability. Samples below the threshold are left unclassified (x-axis), whereas the percentage of correct classifications is computed among samples above α (y-axis). The vertical dashed lines indicate the 90% confidence level for the different CPF estimators.

it was decreased by a factor of up to 6 for the local estimation methods, indicating that the local estimators are better calibrated. To test whether calibration depends on the number of samples used for estimator training, we rerun our evaluation for sample sizes between 100 and 3200. Figure 4B shows that in our study the RMSE values remain relatively constant, and that local estimators outperform the competing methods independent of the sample size, contrary to Niculescu-Mizil and Caruana (2005) who report a superior performance of the binary regression estimator for datasets with less then 1000 samples in the domain of text classification.

3.4 Variance of estimators

Another criterion for estimator evaluation is the variance of the estimator with respect to sampling. Typically, flexibility of an estimator comes at the price of increased variance. The binary regression approach is the most rigid in our collection with only two adjustable parameters. Compound Bayes has four parameters, while the non-parametric local estimators seem to be the most flexible ones. Here, we assess the variance of probability estimators across multiple simulation runs. For this purpose, the original data of ADPKD patients and healthy controls were enclosed as test set in the outer loop of the cross-validation scheme within all simulation runs (from Section 3.3). Hence, each case was predicted number of outer folds multiplied by the number of simulated datasets times for each CPF method. For each fold, a variance was calculated for each case across simulations. Finally, the median variance of each case across folds and methods was evaluated. Thus, we assess the variance for each sample individually. Figure 4C shows box plots of the variance of estimated class probabilities across all samples. We observed that the variance was smallest for the binary regression estimator followed by the Compound Bayes method and larger for the local error estimators, which is not surprising given the increased flexibility of these estimators. Moreover, we observed that the local estimators displayed a wider range of variances across samples showing high variance for samples in the gray zones between classes.

To test for the dependency of estimator variances on the size of the study, we drew random samples of cross-validated scores of ADPKD and healthy patients and re-estimated class probabilities.

The subsampled group sizes varied from 3 to 45 samples (about group size of healthy patients) and we drew 1000 random subsets per study size. The number of features was fixed to 50 in analogy to Gronwald *et al*. (2011). For each study size, every sample receives multiple probability estimates which are summarized into a variance for each sample. Figures 4D and E show the maximum and the median of these variances, respectively. We observed that the maximal and median variance was smallest for the binary regression method uniformly across study sizes. The Compound Bayes performed very poorly for small group sizes of three and five samples, but improved rapidly with studies getting bigger. The variances of the local error frequency-based methods were acceptable also for small studies but decreased only very slowly with studies getting larger.

3.5 Identifying reliable classifications

The ultimate goal of class probability estimation is the identification of those samples that can be reliably classified. Misclassifications should be rare among samples with high class probabilities. This property of an estimator is related to calibration in that we relate long run misclassification rates with estimated probabilities. However, the focus here is on extreme probabilities only. If an estimator is poorly calibrated for probabilities around 0.5, this is less of a problem since clinicians would not base treatment decisions on classifications that are labeled unreliable. If however, an estimated probability is close to 1, it must be reliable since a clinician might want to adjust treatment decisions based on this diagnostic result. Moreover, there is a trade-off between the reliability of a diagnosis and the number of samples that receive class probabilities close to 1. An estimator might assign extreme probabilities only to a small number of cases thus obtaining very low misclassification rates among these cases. However, this estimator might also miss many cases that could actually be reliably classified. Figure 4F shows the trade-off between the percentage of correct classifications and unclassified cases. Confidence threshold α is varied from 0.5 to 1. Samples below the threshold are left unclassified (x-axis), whereas the percentage of correct classifications is computed among samples above α (y-axis).

We observed that the percentage of correct classifications of the local methods does not fluctuate as much as that of the Compound Bayes and binary regression. More importantly, the local methods reached 100% correct classifications faster than the others. For $\alpha = 0.90$ (indicated by the vertical dashed lines in Fig. 4D), Compound Bayes left 76% of samples unclassified with 87.5% correct classifications, 2.5% below the confidence level and binary regression left 85% unclassified being 100% sensitive. The local methods left 78–81% cases unclassified and classified the remaining samples correctly.

4 DISCUSSION

Before a clinician decides on the treatment of a patient, the reliability of the diagnosis must be assessed. In machine learning, reliability can be expressed in terms of classification probabilities. Since little attention has been given to the usefulness of probability estimates so far, we have compared a selection of class probability estimators including Naive Bayes methods, binary regression and local error-based methods in the context of metabolomics-based diagnosis of disease. We found that local error-based estimators show superior performance for instance to more widely used methods, the PAM program, binary regression and Compound Bayes classifiers. Strikingly, the PAM approach displayed a strong sparseness bias. We do not recommend its use in clinical diagnosis. Binary regression-based estimators display the least variance, but are inferior with respect to all other criteria evaluated here. A collection of three local error-based estimators performed best overall with only marginal differences between the individual implementations. We conclude that this type of approach is the method of choice.

The computation time of the entire study is rather large due to three nested cross-validation loops combined with multiple simulation runs. For a practical use of the method, the determining factor is the estimation of class probabilities, which is much faster and corresponds to the innermost loop of the nested cross-validation scheme. The CPU times of an Intel Xeon E5320 1.86 GHz processor at a group size of 45 samples were 0.03, 1.54, 2.32, 2.92 and 9.24 seconds for Compound Bayes, LEF(Smooth), LEF(Adapt), binary regression and LEF(Bin), respectively.

Although the present study is confined to metabolomics data, we believe that similar results can be obtained for different forms of clinical diagnosis based on high-dimensional genome size readouts, e.g. proteomic or transcriptomic profiling data. Note, that from the perspective of CPF estimation the effective dimensionality is that of the signature and not that of the original dataset. The dimensionality of gene expression-based signatures described in the literature is well comparable to that of our metabolomics study. We believe that translational science will profit from further research on class probability estimation in the various contexts of omics data.

ACKNOWLEDGEMENT

The authors would like to thank Ann-Kathrin Immervoll and Matthias S. Klein for sample preparation and measurement. For provision of urinary samples, we thank Bernhard Banas, Carsten A. Böger, Kai-Uwe Eckardt, Stephan Reinhold, Bernd-Detlef Schulze and Raoul Zeltner. We thank the group members Claudio Lottaz, Tully Ernst and Peter Butzhammer of the Statistical Bioinformatics Department, University of Regensburg, for fruitful discussions on the manuscript.

Funding: BMBF grant (PKL-01GS0882-15); Bavarian Genome Network BayGene.

Conflict of Interest: none declared.

REFERENCES

Ambroise,C. *et al.* (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl Acad. Sci. USA*, **99**, 6562–6566.

Ayer,M. *et al.* (1955) An empirical distribution function for sampling with incomplete information. *Ann. Math. Stat.*, **26**, 641–647.

Dawid,A.P. (1982) The well-calibrated Bayesian. *J. Am. Stat. Assoc.*, **77**, 605–610.

Dudoit,S. *et al.* (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, **97**, 77–87.

Fan,X. *et al.* (2010) DNA microarrays are predictive of cancer prognosis: a re-evaluation. *Clin. Cancer Res.*, **16**, 629–636.

Gronwald,W. *et al.* (2011) Detection of autosomal dominant polycystic kidney disease by NMR spectroscopic fingerprinting of urine. *Kidney Int.*, **79**, 1244–1253.

de Hoon,M.J.L. *et al.* (2004) Predicting gene regulation by sigma factors in Bacillus subtilis from genome-wide data. *Bioinformatics*, **20** (Suppl. 1), i101–i108.

MAQC Consortium (2010) The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.*, **28**, 827–838.

Michiels,S. *et al.* (2007) Interpretation of microarray data in cancer. *Br. J. Cancer*, **96**, 1155–1158.

Niculescu-Mizil,A. and Caruana,R. (2005) Predicting good probabilities with supervised learning. In *ICML'05: Proceedings of the 22nd International Conference on Machine Learning*. pp. 625–632.

Parsons,H.M. *et al.* (2007) Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, **8**, 234.

Platt,J.C. (2000) Advances in large margin classifiers. In: Smola,A. *et al*. (eds) *Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods*. MIT Press, Cambridge, MA, pp. 61–74.

Sotiriou,C. and Piccart,M.J. (2007) Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care? *Nat. Rev. Cancer*, **7**, 545–553.

Tibshirani,R. *et al.* (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Wessels,L. *et al.* (2005) A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, **21**, 3755–3762.

West,M. *et al.* (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA*, **98**, 11462–11467.

Wright,G. *et al.* (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl Acad. Sci. USA*, **100**, 9991–9996.

Zadrozny,B. and Elkan,C. (2002) Transforming classifier scores into accurate multiclass probability estimates. In *SIGKDD'02*, Edmonton, Alberta, Canada, pp. 694–699.

Zervakis,M. *et al.* (2009) Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics*, **10**, 53.