Anal Bioanal Chem (2008) 390:1261–1271
DOI 10.1007/s00216-007-1818-6

ORIGINAL PAPER

Assessing and improving the stability of chemometric models in small sample size situations

Claudia Beleites · Reiner Salzer

Received: 4 October 2007 / Revised: 7 December 2007 / Accepted: 14 December 2007 / Published online: 29 January 2008

Abstract Small sample sizes are very common in multivariate analysis. Sample sizes of 10–100 statistically independent objects (rejects from processes or loading dock analysis, or patients with a rare disease), each with hundreds of data points, cause unstable models with poor predictive quality. Model stability is assessed by comparing models that were built using slightly varying training data. Iterated k-fold cross-validation is used for this purpose. Aggregation stabilizes models. It is possible to assess the quality of the aggregated model without calculating further models. The validation and aggregation methods investigated in this study apply to regression as well as to classification. These techniques are useful for analyzing data with large numbers of variates, e.g., any spectral data like FT-IR, Raman, UV/VIS, fluorescence, AAS, and MS. FT-IR images of tumor tissue were used in this study. Some tissue types occur frequently, while some are very rare. They are classified using LDA. Initial models were severely unstable. Aggregation stabilizes the predictions. The hit rate increased from 67% to 82%.

Keywords Chemometrics/statistics · Small sample size · Model aggregation · Model stability · IR spectroscopy/Raman spectroscopy · Brain tumor

Introduction

Several multivariate models for brain tumor diagnosis have been developed [1, 2 and references therein, 3]. The certainty of the results of such data analyses is determined by the number of samples. One major problem encountered during the development of such models is therefore to deal with the uncertainty caused by very limited numbers of samples.

Another very prominent example of the uncertainty of statements based on only a few measurements is the recent discussion about the Atlantic meridional overturning circulation (MOC) [4, 5]; see also the comments [6, 7]. Between 1957 and 2004, five transatlantic cross-sections were measured. These measurements showed a decline in the MOC. The expected error, as indicated by a modeling study, was about 2/3–3/4 of the observed trend [4]. Unfortunately, the expected error did not play a significant—if any—role in the broad public discussion. Recently, measurements over a complete year were published. From the observed variations it was concluded that the past measurements are indeed too sparse to establish trends. It is now estimated that ten years of continuous measurements are needed in order to characterize and quantify the variations [5].

Sparse data sets comparable to the five transatlantic cross-sections are quite common in analytical chemistry. In an environmental study there might be only a few contaminated samples. In a calibration using samples taken from a process only few samples of extreme values of that parameter might be available. In classification tasks like loading dock analysis, or the diagnosis of a particular disease, etc., often the number of samples available may be sufficient, but this varies widely among different classes.

Despite the enormous amounts of data produced by modern analytical instruments, this restricted number of actual samples in the least populated class determines the quality of the results. In this situation, a particularly careful analysis of the uncertainty of the results becomes essential. In addition, the data analysis strategy needs to be adapted to the special requirements of such sparse data sets in order to avoid unnecessarily high uncertainties.

Spectrometers easily acquire data sets of hundreds or thousands of data points per spectrum, and thousands of spectra of a single sample if an array detector is used. Thus, a typical spectroscopic data set might have one or two orders of

C. Beleites (✉) · R. Salzer
Institute for Analytical Chemistry,
Dresden University of Technology,
Bergstrasse 66,
01062 Dresden, Germany
e-mail: Claudia.Beleites@chemie.tu-dresden.de

magnitude more data points per spectrum than the number of samples actually measured. In such situations, the number of samples in the smallest class or the sparsity of samples at the extremes of the calibration range determine the quality of the data analysis. Such models are likely to be unstable. Unstable models have poor predictive quality.

We discuss these problems with an example of a medical diagnosis: 133404 spectra (151 FT-IR micro-images) of 53 samples of glioma brain tumors of different grades and and of five samples of control (healthy) tissue are classified. Classification is achieved by linear discriminant analysis of the mean absorbances using an optimized set of eight spectral regions.

Bias and variance

Bias and variance of chemometric models

All measurements are subject to uncertainty. Chemometric models are characterized by their model parameter values, which determine the relationship between the variates. These parameters are estimated from the data set. Therefore, models are subject to uncertainty, just like any measured value.

Systematic errors (the bias) originate from incorrect assumptions about the model used to address the problem. Thus, systematic errors can, in principle, be avoided by using appropriate functions to model the relationship between the variates. On the other side, this would often require very complex models; in other words, models with many degrees of freedom, and lots of parameters.

Variance is the random error in the fit model. In the case of parametric models it is located in the parameters. Variance in the model causes variance in the predictions of the model. The variance in the model reflects the fact that the model is fitted to a data set with a certain information content (due to the measurement method and, particularly, the finite—and in this context small—sample size) with respect to the problem considered. Models suffering high variance are termed *unstable*. Thus, variance depends on the model's complexity: as the number of estimated parameters increases (the more degrees of freedom the model has), the variance increases too.

In addition, classification models may have an uncertainty that cannot be avoided, the so-called *irreducible error*. This reflects the true overlap of the classes. The total uncertainty is therefore composed of the irreducible error, the bias and the variance.

As mentioned above, both the bias and the variance of the model depend on the complexity of the model. More complex models adapt better to the given training data. Therefore, they are less biased. On the other hand, more complex models have more parameters that are estimated during the training of the model. However, the available data set only has a finite information content, particularly in the case of small sample sizes. Thus, estimating more parameters (using more complex models, i.e., models with more degrees of freedom) means that there is more variance in the model. This tradeoff between systematic and random error in the model parameters is the so-called *bias–variance tradeoff* (see, e.g., [8, pp. 193]). Therefore, it is not always desirable to have a model that can adapt perfectly to the problem. A model with a restricted number of degrees of freedom can have considerably less total uncertainty.

There is an optimal model complexity that depends on the available sample size. For a given data set, one needs to consider what model complexity can be afforded. To put it another way, the question is: how many samples are needed to train a model of a given complexity?

As an example, consider spectra with $p=200$ data points each. Modeling one output as a linear function of $p+1$ inputs (p data points per spectrum plus one more to model an offset) yields a model with $p+1=201$ parameters that are to be estimated.

This does *not* mean that 201 samples need to be measured: just two samples with $p=200$ data points would render the model-building process mathematically solvable. This resembles how linear calibration can be achieved using just two points. In fact, if no offset (the additional parameter) is allowed, a single spectrum would be sufficient to calculate a solution. This would force the model to go through the origin. But the fact that a solution can be found mathematically does not address the problem of how reliable such a model is. In order to reliably estimate the model parameters, more samples are needed.

The full quadratic model already has $^1/_2(p+1)(p+2) = 20301$ parameters. The number of parameters can be reduced by applying restrictions though. If, for example, the interactions between the inputs are not modeled, only 401 parameters are needed. Considering $p=200$ a moderately small number of variates for many kinds of measurements (e.g., spectroscopic or chromatographic data), it becomes clear that the required sample size for nonlinear models is far beyond the scope of the small sample size situations discussed here. Indeed, even linear models might still be unstable.

If the model is not complex enough, it cannot adapt to the problem appropriately. The result will be that the total uncertainty is dominated by the bias. In this case, the underlying relations between the variables have to be approximated more appropriately, and/or more complex models will need to be built, respectively.

If the model is too complex, it is unstable. In this case the variance is the main contributor to the total uncertainty. Better models are obtained by applying stronger restrictions to the model. Such restrictions could be linear instead of higher-order models, variate reduction of the data set, or aggregation, as discussed below.

Uncertainty in quality measures of classification models

After the model is built, a validation is needed to measure its quality. Again, the concept of systematic and random errors applies.

The quality of classification models is often expressed as a hit or error rate, a sensitivity, a specificity, a positive predictive value (ppv) and a negative predictive value (npv), among others. These values are relative frequencies, e.g., the number of correctly classified measurements divided by the number of test measurements gives the hit rate. These values need to be obtained by testing the model with new samples. The results of these tests are summarized in the so-called *confusion matrix* (Fig. 1). This is a table that compares the prediction of the model (columns) with the true class of the test samples (rows). The diagonal gives the correct predictions, while all other cells indicate wrong predictions. Figure 1a shows the setup of the confusion matrix for the astrocytoma data set used throughout this study. Figure 1b–d demonstrate for the first class (here: G, glioblastoma) which parts of the confusion matrix are summed up (gray-coded) and set into relation to each other in order to calculate hit rate, sensitivity (how well the model recognizes samples of a particular class), and specificity (how well the model recognizes that the sample does *not* belong to the particular class).

Considering these calculations as Bernoulli trials, the expected distribution of such a value is binomial:

$$\Pr(s) = \binom{n}{s} p^s (1-p)^{n-s} \tag{1}$$

where s is the number of "successful" observations (i.e., the enumerator in Fig. 1b–d), p is the true probability (e.g., the model's true hit rate), and n is the number of tests (i.e., the denominator in Fig. 1b–d).

Expectation and variance are

$$\mathrm{E}(s) = np \tag{2}$$

and

$$\mathrm{Var}(s) = np(1-p) \tag{3}$$

Having observed s successes, the probability is estimated to be

$$\widehat{p} = \frac{s}{n} \tag{4}$$

Assessing the quality of the model, we are interested in the confidence interval for the true value p if s successes during n tests are observed. The binomial distribution can be approximated by a normal distribution using the variance as given in Eq. 3. Unfortunately, this approximation should only be used if $n\,p>5$ for $p<0.5$, or $n\,(1-p)>5$ for $p<0.5$. This means that at least $n=10$ samples need to be in the denominator of the proportion we are interested in (with a true probability $p=0.5$). With more realistic (or desirable) values for p, e.g., $p=0.95$, at least 100 test samples are needed. As such sample sizes are not available in small sample size applications, the normal distribution gives only a very rough approximation here; no more than a "guesstimate."

Better confidence intervals can be constructed from the binomial distribution itself. Equation 1 gives the probability of observing s successes in n tests given a true probability p. The confidence interval we are interested in can be expressed as:

$$\int_{\text{lower bound}}^{\text{upper bound}} \Pr(s)dp = \text{confidence level} \tag{5}$$

This integration can be done numerically using the normalization

$$\int_{p=0}^{1} \Pr(s)dp = 1 \tag{6}$$

which says that the probability is 1 that the observed s successes during the n trials stem from any true probability p.

Start at $\widehat{p} = \frac{s}{n}$ and expand the bounds of the integral stepwise into the direction with higher probability until the

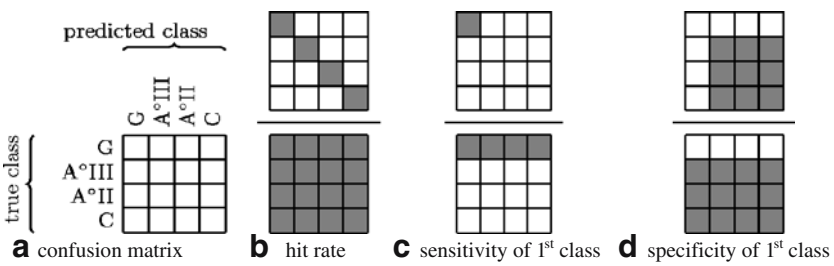

Fig. 1 Confusion matrix and model quality measures. The classes for the data set used in this article are Glioblastoma (G), Astrocytoma °III (A°III), Astrocytoma °II (A°II), and Control (C)

desired confidence level is covered. This gives the bounds of the smallest confidence interval. One-sided confidence intervals are obtained starting with the lower bound 0 or the upper bound 1.

Still, this calculation needs s and n to be known. This may sound trivial. However, consider a data set consisting of huge numbers of spectra from different locations for each sample, but only a few samples. In this case, the question of how to obtain a statistically relevant sample size is not easily answered. The spectra of one sample are not statistically independent. However, they are measured because they add information. In this case, the effective sample size (i.e., the number of samples that need to be measured to get the information content using only one spectrum per sample) is somewhere between the number of samples and the number of spectra. More detailed discussions of cluster sampling, design effects, and effective sample size are provided in the literature, e.g. [9, p. 425]. Using the number of real samples for n results in a conservative bound for the confidence interval. "Conservative" means that there is no more variance in the validation results due to the finite test sample size than that expressed by these bounds.

Resampling: iterated k-fold cross-validation and out-of-bootstrap validation

When we are already hampered by too few samples to build good models, we cannot afford to reserve samples for testing alone. Therefore, so-called *resampling* techniques are applied: the data set is split into a training subset and a testing subset in such a way that each sample appears in only one of these two sets. Individual (i.e., new) models are built for each training subset. We call these *submodels*. The submodels are tested with their respective testing subsets.

k-fold cross-validation is a resampling scheme widely used for model validation. The training samples are constructed by excluding $1/k$ of the samples for each of the k training sets so that each sample is excluded once. This is "drawing without replacement" (see Fig. 2). The excluded samples are then used as statistically independent test samples for this model. During iterated k-fold cross validation, the whole procedure is iterated using different random splits of the data. Then each sample serves as independent test sample once per iteration. Iterated k-fold cross validation estimates of model quality parameters such as hit or error rates are subject to less variance uncertainty compared to the values obtained with the standard procedure (only one iteration).

Out-of-bootstrap validation differs from k-fold cross-validation in the construction principle of the training data sets: samples are drawn with replacement (Fig. 3). Usually the constructed bootstrap set has the same sample size as the data set the samples are drawn from. Drawing with replacement means that a particular sample can be drawn more than once into the bootstrap set. Then other samples are left out and can serve as independent test samples for a model trained on the bootstrap set. On average, 63% of the samples show up in this training set, and the remaining 37% are used for testing. A large number of bootstrap models are built and tested with the left-out samples to obtain the out-of-bootstrap validation results. Bootstrap methods are discussed in detail by Efron and Tibshirani [10]. The book also includes a discussion of the jackknife.

The total uncertainty in the hit and error rates was found to be similar for iterated k-fold cross-validation and for out-of-bootstrap validation for simulated spectral data and several multivariate data sets (from the UCI pattern-recognition database [11]). While out-of-bootstrap estimates expressed higher bias, they in turn had lower variance than k-fold cross-validation [12, 13].

Model validation using resampling methods assumes that all of the models built and evaluated during the validation procedure are sufficiently similar to the model built on the whole data set. The validation results for these submodels can be used directly instead of an extra (independent) test set. This assumption breaks down in the small sample size situations discussed here: the submodels are on average worse than the model built using all samples for training. This is indicated by the well-known pessimistic bias of the resampling validation results.

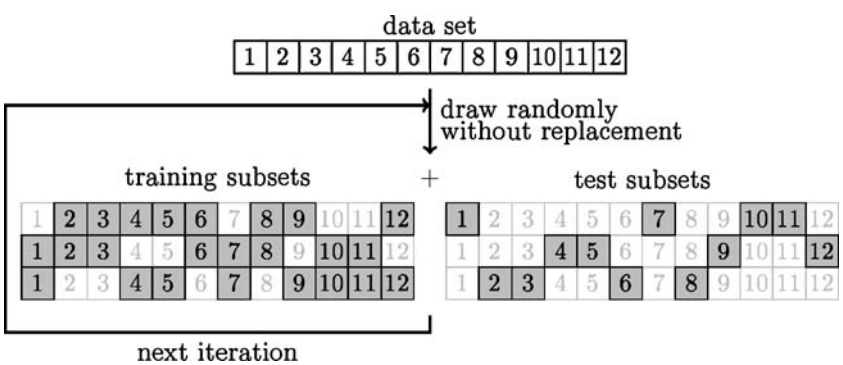

Fig. 2 Iterated k-fold cross-validation. k=3 models (rows in this figure) are built in one iteration. Each model is built with a training subset excluding $\frac{n=12}{k=3} = 4$ samples. They form the test subset for this model. The data set is split randomly; each sample appears in a test subset exactly once in each iteration. Additional iterations require additional random splits

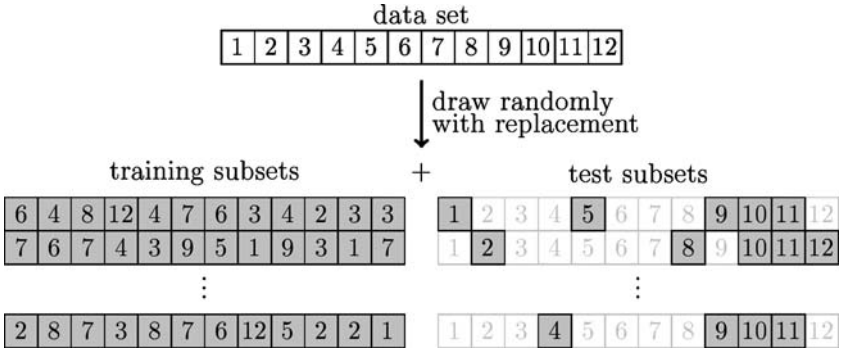

Fig. 3 Out-of-bootstrap. Each training subset (rows in this figure) is constructed by drawing $n=12$ samples with replacement. The samples not drawn are the test subset. The test subsets differ in length. By accident, a particular sample may not show up in any of the test subsets, as it is used in all training subsets (samples 3, 4, 6, and 7), or vice versa (samples 10 and 11). A large number (e.g., 100) of subsets are usually drawn in this way

A second, weaker assumption for the resampling validation is that the submodels are sufficiently similar to each other that it is permissible to average the test results of several submodels. In other words, the submodels are assumed to be stable with respect to the variation in the subsamples used to train the submodels.

Model stability

Unstable models react very sensitively to changes in the training data: small variations in the training data lead to large variations in the model built. Models can be built on slightly varying training sets. Their stability can then be assessed by examining the model parameters. Alternatively, these models can be tested with the same test sample. Model stability is then measured by comparing the variation in the results for this sample.

Such slightly varying data sets are produced during resampling-based validation schemes, as discussed above. The stability of the submodels can be assessed by comparing the model parameters of the submodels. An alternative is to compare the test results obtained for the same test sample with different submodels.

If the models turn out to be unstable, model aggregation may help.

Model aggregation

Model aggregation means combining the predictions of a number of different models. Figure 4 illustrates two common methods of aggregation: for regression models, the mean (or median) of the submodels' predictions can be used as the prediction of the aggregated model. This aggregation scheme is not restricted to regression models. In particular, classification models that predict probabilities rather than class labels can also be aggregated by averaging. Alternatively, classification models can be aggregated by *majority vote*: the class that was chosen by the submodels most often is predicted. Models calculated during the mentioned resampling procedures can be used for aggregation purposes. Aggregating models based on bootstrap resampling is termed *bagging* [14].

Model aggregation sharpens the prediction. This is the same strategy as measuring a value several times and using the mean to get a measure with less variance of the true value. The sharpening effect of aggregation is clearly seen with majority vote aggregation: a few wrong predictions within mostly correct predictions are outvoted by the correct ones. If the wrong predictions are in the majority, the model will either refuse to give a prediction (if no class has the majority, or a threshold for the majority, e.g., > 2/3 of the votes, is not met), or one of the wrong classes may win. Thus, if the submodels are not only unstable (imprecise) but also bad (inaccurate), the aggregated model may even be worse, as it could emphasize the wrong tendency. Still, this will only happen for really bad models. As an example, consider the use of majority vote in classification models: worsening caused by wrong predictions will only happen if the models that are aggregated are on average worse than guessing.

In addition, the distribution of the predictions of the submodels can be evaluated. Just as repeated measurements allow us to calculate the standard deviation, the standard deviations of the predictions of the submodels can be calculated. This allows us to check for the uncertainty of the aggregated prediction: e.g., a prediction can be refused if the standard deviation of the predictions of the submodels is too large.

Aggregation is a restriction on the models. Just as the mean of a number of measured values has a lower variance than the measured values themselves, the mean of many models has lower variance than the models themselves. As long as aggregating improves stability more than it introduces bias, aggregation will improve the model. Aggregation thus helps to reduce the total uncertainty if the problem with the submodels is instability.

In general, all of the submodels must be evaluated in order to obtain the aggregated prediction. A linear model is expressed as a vector of the linear combination coefficients for the variates. Such models can be aggregated directly to speed up the prediction. The aggregation then is done by

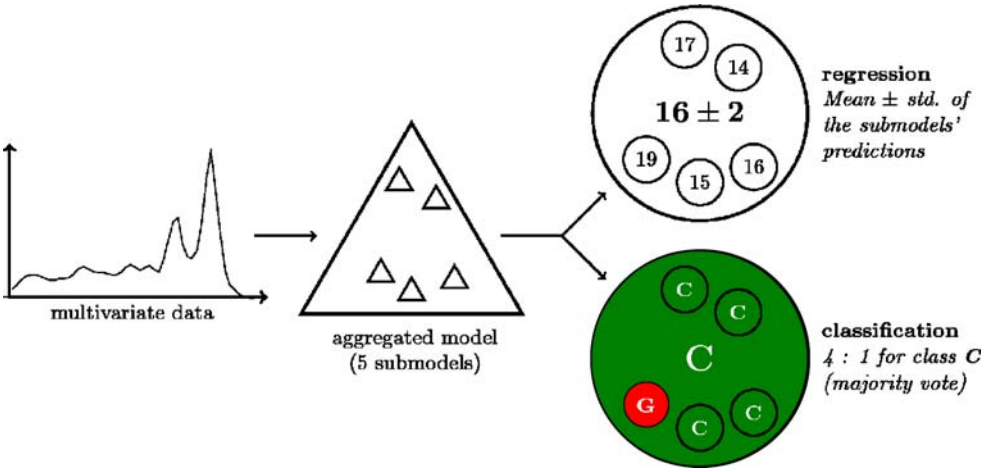

Fig. 4 Model aggregation. Multivariate data (*left*) is presented to several submodels (*small triangles*) that each give a prediction (*small circles on the right*). These predictions can be aggregated by using the mean of the predictions of the submodels (*large circle to the upper right*). Even the standard deviation can be calculated. The predictions of classification models can also be combined by majority vote (*large circle to the lower right*)

aggregating the model parameters themselves instead of aggregating the predictions of the submodels: e.g., the mean vector of the linear coefficient vectors of all submodels is used as the parameter vector for the aggregated model. Then only one vector multiplication is needed to calculate the aggregated model's prediction instead of one vector multiplication per submodel plus an averaging of the predictions of the submodels.

The aggregated model needs to be tested like any other chemometric model. The hit rate, etc., of the aggregated model's quality can be calculated without building further models. Again, test samples are needed that are independent of the training samples the aggregated model was built with. The resampling strategies described above construct the training subsamples in such a way that a particular sample is excluded from the training subset more than once. For iterated *k*-fold cross-validation, each sample is excluded once per iteration.

Thus, there are several different submodels (one per iteration of *k*-fold cross-validation) that were built without using that particular sample. Aggregating these submodels (instead of all available submodels) yields an aggregated model that is statistically independent of that particular sample. This model can be tested with that sample. This can be done for all samples giving results from statistically independent test samples. For bootstrap-aggregated (bagged) models, this is called the *out-of-bag* estimate [15]. This method is illustrated in Fig. 5 for *k*-fold cross-validation.

Data set and data analysis setup

We tested the discussed data analysis methods by classifying a data set of FT-IR spectra of different grades of brain tumors. The data set is characterized by widely varying sample numbers (between 3 and 43) in the classes and a large number of variates (> 200 data points per spectrum).

Gliomas are the largest group of primary brain tumors. The tumor classes analyzed in this study form a histological continuum from benign astrocytoma to malignant glioblastoma. Astrocytoma °II are early-stage tumors and are usually undiagnosed or are not surgically removed, and so this sample class is most rare. "Control" means healthy brain tissue, which is usually not surgically removed either.

The data set used here consists of 150 FT-IR micro-images from 58 samples of 58 patients with astrocytoma of different histological grades and of control patients (see Table 1). The sample sizes of the classes vary widely (ranging from three samples of astrocytoma °II to 43 glioblastoma samples) and reflect the different relative frequencies and sizes of the different tumor grades.

Samples were obtained during surgical treatment, snap-frozen in liquid nitrogen, and 10-μm cryo sections were prepared on CaF_2 slides for spectroscopic measurements. FT-IR micro-images were captured by the imaging spectrometer "Hyperion" (Bruker Optik GmbH, Ettlingen, Germany). This system comprises an IFS 66 continuous scan spectrometer, an infrared microscope (IRscope II) with a 15×Cassegrain objective, and a 64×64 MCT Focal Plane Array (FPA) detector. The sample area of one FT-IR image captured in transmission mode is 270×270 μm. A spectral resolution of 8 cm^{-1} was used and the spectra were measured in transmission mode. FT-IR micro-images were collected from all visually distinct parts of the sample in order to cover all tissue types.

A filter was applied to exclude spectra with minimum absorbances of below −0.1 or above 0.3, and spectra with maximum absorbances of less than 0.1 or more than 1.0. This filter excludes bad pixels of the detector as well as areas without sample or tissue that is too thick (due to folding of the thin section). The spectra were cut to 1000–1800 cm^{-1} and minimum–maximum-normalized: the minimum is at ca. 1800 cm^{-1} and the maximum is the amide I band. Thereafter, a histogram filter calculates the histogram of the absorbance values at each point and accepts spectra that are within the range populated by at least 16

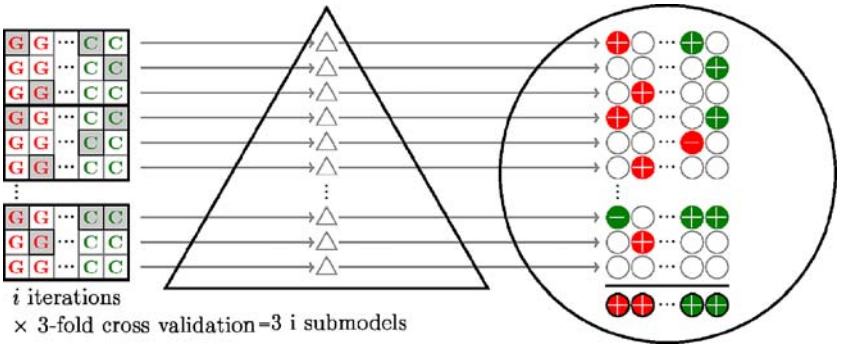

Fig. 5 Testing the aggregated model. Each submodel is built by excluding certain samples (indicated by the *gray background* in the rectangle on the left; the letters G and C indicate the classes). Each sample is excluded from the training data once per iteration (*thick rectangles*). For example, the first sample was not used to build submodels 1 of the first and last iterations, and 3 of the second iteration. During each iteration, k=3 submodels are built. Thus, a total of $3i$ submodels are built during the i iterations (*triangles*). The predictions of the submodels (*small circles*) are aggregated (*circles below the thick line*). During testing, only the predictions of those models that were built by excluding the particular sample are aggregated. Here, aggregation is done by majority vote. For example, the predictions shown for sample 1 are 2 G (*red*) : 1 C (*green*); thus it is assigned to class G. The predictions are compared with the true classes: G is correct, as indicated by +, and C is wrong (as indicated by −) for sample 1. The mean error rate of the shown samples is $\frac{2}{12} = 17\%$ (the results above the line), while the aggregated model has $\frac{0}{4} = 0\%$ errors (below the line)

absorbance values per histogram bin [16]. Thus, any heavy tails in the distribution of absorbance values are trimmed; linear discriminant analysis (LDA) is known to be sensitive to heavy tails [17]. In order to improve the signal-to-noise ratio and minimize the memory size requirements, as well as to speed up calculations, a pixel binning of 2 was performed for each FT-IR micro image. All preprocessing was done separately for each FT-IR micro image using Matlab scripts (Mathworks Inc., Natick, MA, USA) written in-house.

The data set consists of a total of 133404 spectra of the classes "Glioblastoma," "Astrocytoma °III," "Astrocytoma °II," and "Control." Class means and standard deviations of the preprocessed spectra are shown in Fig. 6.

Classification was done using a two-step procedure that optimized the selection of input variates and classification. First, a given number of spectral regions were selected. The mean absorbances of these regions were then used as input variates for linear discriminant analysis (LDA). This classification model was optimized by genetically optimizing the selected spectral regions with respect to the quality of the LDA model. This was done using the program *ga_ors* developed at the National Research Council of Canada, Institute for Biodiagnostics [18]. We used a population size of 200 models optimized over 25 generations with eight spectral regions. Model training was done on the mean spectra of the FT-IR micro-images, and the spectra were weighted so that all samples within a class had the same weight and all classes had the same weight or prior probability.

The prediction of the program is the posterior probability for the classes; i.e., the probability that the spectrum belongs to tissue of that particular class.

Classification is done by assigning the class with the highest posterior probability to the spectrum. In addition, the posterior probability can be used to distinguish how sure the classification is: if the posterior probability is below a certain threshold, the spectrum is rejected as it cannot be classified with a high level of certainty. If the posterior probability is above this threshold, the spectrum is accepted for classification. As the choice of this parameter is not the main topic of this paper, it was fixed as follows.

The relation between the required posterior probability (threshold) and the total hit rate (accuracy) of the predictions of the submodels is depicted in Fig. 7a (thin line). The thick line corresponds to the aggregated model. The aggregated model is clearly better than the submodels in terms of the given threshold for the posterior probability.

Figure 7b gives the proportion of the spectra that the submodels accepted for prediction (black: all spectra, colored: the four classes) as a function of the chosen threshold for the posterior probability. The higher that the posterior probability is required to be in order to accept a prediction, the better the overall hit rate of the model. On the other hand, less and less spectra are accepted. The thin line in Fig. 7c gives the relation between the hit rate and the percentage of spectra accepted for prediction by the submodels. The thick line characterizes the aggregated model. If the submodels are compared to the aggregated model in terms of hit rate with respect to the number of

Table 1 Tumor data set

Class	Number of samples	Number of images	Number of spectra
Glioblastoma	43	102	89573
Astrocytoma °III	7	23	20799
Astrocytoma °II	3	5	4802
Control	5	20	18230
Total	58	150	133404

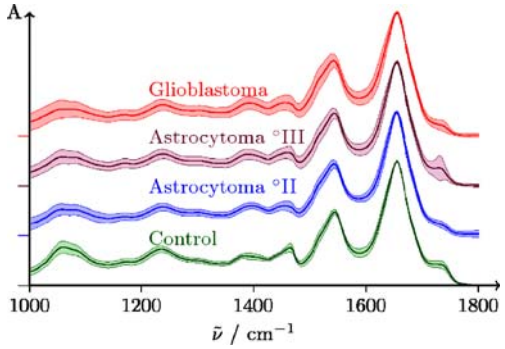

Fig. 6 Mean spectra ± one standard deviation for the classes. Offsets were introduced for clarity

spectra accepted for prediction, the aggregated model again clearly performs better than the submodels.

The threshold for the submodels was chosen to be a posterior probability of 0.85. Thus the submodels rejected predictions with less posterior probability as being "uncertain." This corresponds to accepting about half of the spectra for prediction (Fig. 7b).

As cross-validation ensures that all samples have the same weight in the data analysis, we performed 40 iterations of a fivefold cross-validation. A total of 200 optimized models was calculated. Cross-validation was carried out outside of both region selection and the training of the LDA model in order to ensure that the test samples are statistically independent of the tested model. Cross-validation sets were drawn stratified; i.e., the proportions of the classes were required to be approximately constant for all subsets. As the spectra for one patient (sample) are not statistically independent, the splits for the cross-validation were done sample-wise.

Aggregated model Model aggregation was then done using the mean class membership probability (posterior probability) over the 40 models that were built excluding the particular sample in order to get estimates of hit rate, sensitivities, and specificities of the model. Upon comparing Fig. 7b and d, it becomes clear that the aggregated model accepted less spectra than the submodels accepted for a given threshold of posterior probability. This can be explained by the process of averaging.

One might argue that the advantage of the aggregated models is the exclusion of more spectra from the prediction. But, as discussed above, Fig. 7c shows that the aggregated model is also superior to the submodels with respect to the number of spectra that are accepted for prediction.

A threshold was applied to filter out spectra that had average class posterior probabilities of below 0.65. The threshold was chosen so that more than half of the spectra were accepted for classification (see Fig. 7d), which equals the proportion of spectra accepted by the submodels at a threshold of 0.85 (Fig. 7b). The aggregated model has higher hit rates at the same threshold value for the posterior probability and for the same percentage of spectra accepted for prediction.

In order to check the stability of the aggregated model, five aggregated models consisting of eight submodels each were tested in addition to the model aggregating 40 submodels.

Results and discussion

The submodels

Figure 8 shows the overall hit rate and sensitivity and specificity for the four classes. The box–whisker plots give the observed ranges for these values for the 40 iterations. The crosses and circles are the observations for aggregated models and are discussed below in the "Aggregated model" section.

Confidence intervals on these model quality measures can be obtained in two ways: parametric or nonparametric. The parametric calculation uses the known distribution of the values, while nonparametric methods do not rely on any particular distribution. The parametric calculation can be done using the binomial distribution with $\widehat{p} = 67.1\%$ as

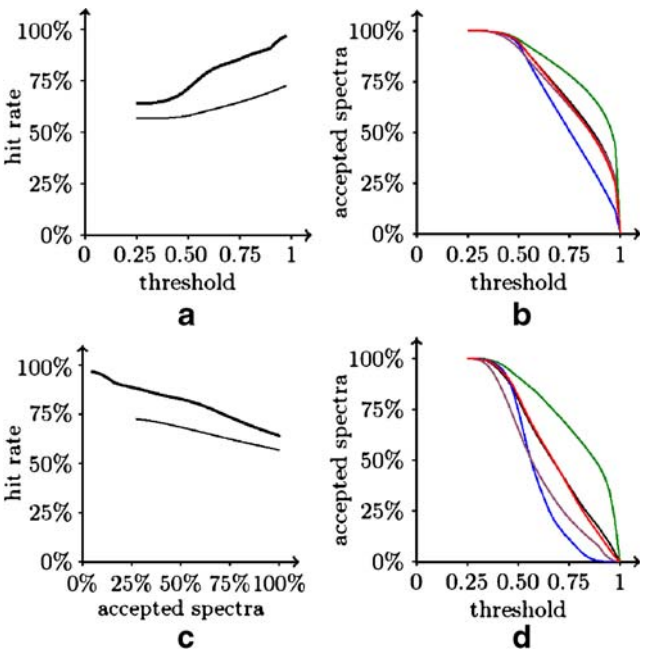

Fig. 7 Choosing the threshold for posterior probability in order to reject predictions as "not sure". Colors: glioblastoma (red), astrocytoma °III (violet), astrocytoma °II (blue), and control (green). **a** Hit rate of the submodels (thin) and the aggregated model (thick) depending on the threshold of posterior probability. **b** Percentage of spectra accepted for prediction using the submodels depending on the threshold of posterior probability. **c** Hit rate of the submodels (thin) and the aggregated model (thick) over the percentage of spectra accepted for prediction. **d** Percentage of spectra accepted for aggregated prediction depending on the threshold of posterior probability

described in the "Uncertainty in quality measures of classification models" section. As explained above, the statistically relevant effective sample size for this parametric calculation is not known. Still, it cannot be less than the number of test patients. On the other hand, the applied resampling scheme permits the nonparametric estimation of the confidence intervals.

By comparing the results from the iterated cross-validation to those from the binomial distribution, the effective sample size of the test set can be roughly estimated.

For a binomial distribution with $\widehat{p} = 67.1\%$ and $n=58$ test samples, the 90% confidence interval ranges from 55 to 74%. Compared to this, 90% of the observations are in the range 60–74%, indicating that the thousands of test spectra actually contain more information than that present in a statistically relevant sample size of 58. The next structures in the data set are the FT-IR micro-images measured in order to cover visually distinct areas of the sample. Therefore, one might hope that the effective sample size equals the number of FT-IR micro-images, although this cannot be proven to be the sample size for a classification model that is supposed to work across patients. The expected range for the hit rates with a test sample size of $n=150$ is 61–73%. Thus, the observed hit rates are close to the expected distribution of hit rates for experiments with $n=150$ and a true hit rate of $p=67.1\%$. On the other hand, extending this idea to testing with an effective sample size of $n=1/2 \times 33404$ spectra (the factor 1/2 acknowledges that only half of the predictions are accepted), the observed hit rates should vary by much less than 1% (90% of the observations should be between 66.8 and 67.4%; here the normal distribution can be safely used for approximation). These results permit the conclusion that measuring FT-IR micro-images from visually different areas of the samples actually improved the information content of the data set, while the spectra of each image behave more like repeated measurements rather than like independent samples.

The discussion above did not provide a conclusion about whether the observed wide distribution of hit rates stems completely from the variance due to the finite test sample size or whether model instability is also a main contributor to the observed variance between the test results. In order to tackle this question, we assessed the model stability by comparing predictions for the same spectrum given by different models.

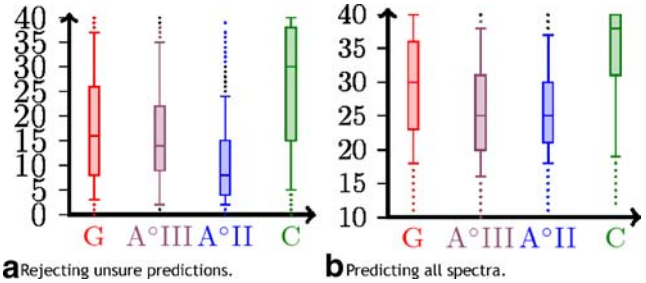

Fig. 9 Stability of the models: number of models that selected the most commonly chosen class for each spectrum (out of a total of 40 models). **a** Rejecting prediction at a threshold of 85% posterior probability. Values of less than 10 indicate that most models rejected the prediction for this spectrum. **b** If all spectra are predicted, the lowest possible value is 10. *Box*: median and quartile values for all spectra of one class; *whiskers*: 5th–95th percentile. Classes: glioblastoma (G), astrocytoma °III (A°III), astrocytoma °II (A°II), and control (C)

The box-whisker plots in Fig. 9 show how often the most commonly chosen class was actually chosen for a particular spectrum. A perfectly stable model would show the same class 40 out of 40 times (regardless of whether it is the correct class or not). Three classes show severe instability, while the control samples are classified as being somewhat more stable despite the fact that they are the second smallest class.

The bars in Fig. 10 show, for each point in the spectrum, the relative frequency that this point was chosen in one of the spectral regions that were used for the LDA. In other words, this is the proportion of models that used this particular point.

On the one hand, this allows for inference with regard to model stability: while there are clear trends, the models are not stable as the positions and spectral widths of the selected spectral regions vary. On the other hand, this allows us to check for the biochemical meanings of the selected spectral regions. The most frequent selections

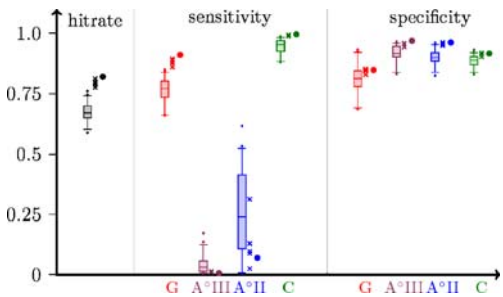

Fig. 8 Hit rates, sensitivities, and specificities of the models. *Box*: median and quartile values for the 40 submodels; *whiskers*: 5th–95th percentile. *Crosses*: Observed values for the five models aggregating eight submodels each; *circles*: observed values for the model aggregating 40 submodels. Classes: glioblastoma (G), astrocytoma °III (A °III), astrocytoma °II (A °II), and control (C)

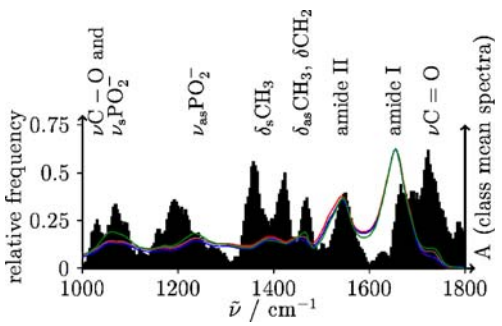

Fig. 10 Model parameter stability: the relative frequency that a point in the spectrum is chosen by a model. The class mean spectra are also drawn for easier interpretation. Colors: relative frequency (black bars), glioblastoma (red), astrocytoma °III (violet), astrocytoma °II (blue), and control (green)

include spectral ranges commonly ascribed to the ν(C=O) ester stretching vibration at 1730 cm^{-1}, amide groups (the shoulder of amide I at 1680 cm^{-1} and amide II 1560 cm^{-1}), and lipids (acyl >CH_2 bending at 1480 cm^{-1}). The spectral window 1020–1120 cm^{-1} is usually associated with >C–O stretching modes of carbohydrates (e.g., in gangliosides).

The spectral regions of phosphate stretching at 1040 cm^{-1} (ν_s phosphate and ν(C–O)) and 1240 cm^{-1} (ν_{as} phosphate) might be of particular interest with respect to the data analysis: both are chosen in a considerable number of models. These two regions carry biochemically equivalent information. If they are exchanged for each other in two models, there is instability with respect to the model parameters in chemometric terms, but it should not lead to instability in terms of of prediction or molecular information. Therefore, one must carefully check the chemistry behind the model in order to judge the importance of features and the importance of variation between models.

Aggregated model

Figure 11 visually demonstrates the power of model aggregation. Figure 11a depicts the results for 40 models that were built excluding sample no. 50. These 40 models were subsequently applied to classify the spectra of sample no. 50. In terms of the median, ca. 55% of the models agree in their prediction of these spectra (without threshold for posterior probability). The median of all astrocytoma °II spectra is 62.5%. Moreover, the astrocytoma °II class was the smallest class in the data set, containing only three patients. Thus, this is one of the most problematic samples of the whole dataset. Note that each of the 40 little images on the left gives the result for the *same* FT-IR image; i.e., exactly the same spectra are classified. The difference between the images is solely due to the differences between the classification models.

Figure 11b (right) shows that the predictions of the aggregated model for sample no. 50 have greatly enhanced accuracy. Despite the fact that the 40 submodels yielded very different predictions, only very few spectra are wrongly classified (12; i.e., about 10 % of the predictions). Still, no prediction is given (indicated by gray pixels) for most of the spectra. The required posterior probability of 65% is not met for those spectra. This is due to the fact that the predictions of the submodels were very different for those spectra. In this case, it is preferable to refuse a prediction rather than to run the risk of providing a prediction that is likely to be wrong.

The crosses in Fig. 8 indicate the observed hit rates, sensitivities, and specificities for the models aggregating eight tests. Note the improved mean hit rate and the far closer distribution (i.e., improved stability) compared to the distributions of the predictions of the different submodels. The filled circles show the results of aggregating all available test predictions: the hit rate is even higher.

Normal and glioblastoma show different developments regarding the sensitivities and specificities compared to the intermediate astrocytoma classes. While the classification of normal and glioblastoma classes improved substantially, the sensitivity for the astrocytoma decreased. Taking into account the small propotion of astrocytoma spectra of grades II and III that are accepted for prediction, the aggregated models—particularly those with higher thresholds for the posterior probability—develop towards a two-class model that distinguishes normal from highly malignant glioblastoma.

Astrocytoma °II and III have median sensitivities of less than 25% in the submodels, meaning that the models are worse than guessing with respect to the sensitivity.

Considering the small sample numbers available for astrocytoma °II and III and the large overlap in the classes (see spectra in Fig. 6), the grading of these intermediate tumor stages is very difficult. There is a continuum in the tumor classes, as they are three different grades of the same kind of brain tumors (glioma). Astrocytoma °II can develop into astrocytoma °III and eventually become glioblastoma. The classes were set up according to the histological grading. This scheme discriminates between tumor grades mainly based on morphology. However, this does not mean that biochemical changes in the cell occur at the same time as morphological changes. Also, a histologist may report that a particular tumor is between two of these grades (e.g., it is not a grade II tumor any longer, but it has not yet completely become grade III). This is a first reason for a class overlap. Furthermore, a high-grade tumor may still have areas of lower-grade tissue. The diagnosis is then the highest tumor grade that the histologist observed. Thus more class overlap may come from "mislabeling" due to the rules for grading tumors.

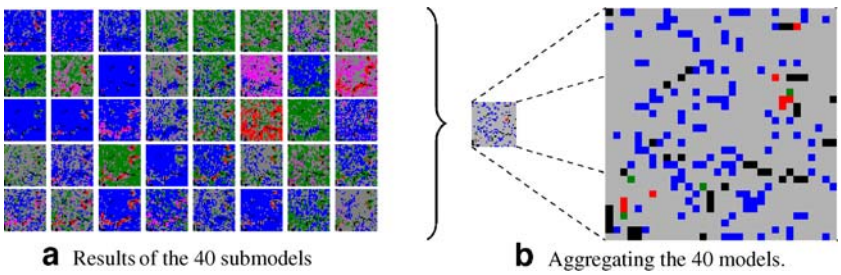

Fig. 11 Checking the predictions visually: sample 50 (Astrocytoma °II, smallest class) classified by 40 different models (**a**) and aggregation of these predictions (**b**). Colors: glioblastoma (red), astrocytoma °III (violet), astrocytoma °II (blue), and control (green), rejected (gray), and spectrum filtered out (black)

Glioblastoma differ from astocytoma °III mainly in that glioblastoma have necroses. However, they might also have areas that are morphologically like tumor tissue of grade II. Thus, a sample of a glioblastoma might contain lower-grade tissue. In addition, gliomas grow by infiltrating the surrounding tissue. Thus it is extremely hard for a surgeon to find the correct tumor border. This leads to the possibility that there can even be normal tissue in the tumor samples.

Recent information on similar samples indicates that lower-grade tissue may be as much as half of the tissue. We even found that up to 10% of the high-grade astocytoma (astrocytoma °III and glioblastoma) consisted mainly of non-tumor tissue.

Under these circumstances, any validation showing perfect distinction of the intermediate tumor classes would lead to suspicions of being overtrained and tested with statistically dependent data.

The distinction between control and tumor tissue is good despite the fact that two intermediate astrocytoma classes are between these tissue types. This is already indicated by a visual inspection of the complete spectra (including the ν(C–H) region), which shows a strong decrease in the ratio between ν(C–H) and amide I band (indicating that the lipid: protein ratio changes) from normal to tumor tissue. Spectroscopically, the information for the ν(C–H) region is also available in the fingerprint region (bending vibrations). Thus the high-wavenumber spectral region does not need to be included in the data analysis (see, e.g., the band at 1480 cm^{-1} in Figs. 6 and 10).

This permits very good separation between the spectra of normal and tumor tissue. Several spectral regions carrying the same chemical information lead to more stable predictions even when the model parameters themselves may not be that stable. In addition, the variablility in the spectra of normal tissue is less than that in the spectra of tumor classes (see Fig. 6).

Summary

Chemometric models are built using measured data. Therefore they are subject to systematic and random error (bias and variance), just like any measured value. Model stability (variance) can be measured by comparing the model parameters of several models built on slightly varying training data. As an alternative, the predictions from several such models for the same sample can be compared. Resampling schemes such as those used during iterated *k*-fold cross-validation or bootstrap samples can be used for this purpose.

In addition, these resampling schemes provide model quality estimates with low variance uncertainty. They also allow us to check the variance of these estimates.

Unstable models can be stabilized by model aggregation. This lowers the variance of the model, and thus improves its predictive quality. For classification models, this will also sharpen the prediction so that models with good average performance will improve. Bad models (e.g., classification models that are on average wrong, or regression models that use inappropriate approximations) will get even worse.

Model aggregation can be accomplished using the models built during a resampling-based validation procedure. In addition, even estimates of the quality of aggregated models can be calculated without training any further models by aggregating. In order to get statistically independent test results, only those predictions that were made by leaving out a particular sample are aggregated. Thus the prediction of this sample is statistically independent of this aggregated model and can serve as test sample.

For the presented data set of four classes of glioma (brain tumors) and control tissue, the overall error rate was reduced from 33 to 18% by model aggregation, and the stability of the predictions greatly improved.

References

1. Krafft C, Sobottka SB, Geiger KD, Schackert G, Salzer R (2007) Anal Bioanal Chem 387:1669–1677
2. Krafft C, Thümmler K, Sobottka SB, Schackert G, Salzer R (2006) Biopolymers 82:301–305
3. Beleites C, Steiner G, Sowa MG, Baumgartner R, Sobottka S, Schackert G, Salzer R (2005) Vib Spectrosc 38:143–149
4. Bryden HL, Longworth HR, Cunningham SA (2005) Nature 438:655–657
5. Cunningham SA, Kanzow T, Rayner D, Baringer MO, Johns WE, Marotzke J, Longworth HR, Grant EM, Hirschi JJM, Beal LM, Meinen CS, Bryden HL (2007) Science 317:935–938
6. Schiermeier Q (2007) Nature 448:844–845
7. Church JA (2007) Science 317:908–909
8. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning; data mining, inference and prediction. Springer, New York
9. Forthofer RN, Lee ES, Hernandez M (2007) Biostatistics, 2nd edn. Elsevier, Amsterdam
10. Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York
11. Asuncion A, Newman D (2005) UCI machine learning repository. http://archive.ics.uci.edu/ml/. Accessed 24 December 2007
12. Beleites C, Baumgartner R, Bowman C, Somorjai R, Steiner G, Salzer R, Sowa MG (2005) Chem Intell Lab Syst 79:91–100
13. Kohavi R (1995) In: Mellish CS (ed) Proc 14th Int Joint Conf Artificial Intelligence, Montréal, Québec, Canada, 20–25 August 1995. Morgan Kaufmann, San Francisco, CA, pp 1137–1145
14. Breiman L (1996) Machine Learning 24:123–140
15. Breiman L (1996) Out-of-bag estimation. Technical report, Statistics Department, University of California, Berkeley, CA
16. Beleites C (2003) Chemometrische Auswertung von IR-Images und -Maps. Master's thesis, Dresden University of Technology, Dresden
17. Huberty CJ (1994) Applied discriminant analysis. Wiley, New York
18. Nikulin A, Dolenko B, Bezabeh T, Somorjai R (1998) NMR Biomed 11:209–216