2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

A Classifier-based Association Test for Imbalanced Data Derived from Prediction Theory

Johannes Mohr, Sambu Seo and Klaus Obermayer
Department of Electrical Engineering and Computer Science
Technische Universität Berlin
Bernstein Center for Computational Neuroscience Berlin
Germany
Email: {jm, seo, oby}@ni.tu-berlin.de

***Abstract*—How can we test for group differences in multidimensional input patterns, such as functional magnetic resonance imaging measurements or gene expression values? One solution is to split the available data into training and test set, and to estimate the generalization accuracy of a classifier that predicts the group variable from the input pattern. If this lies significantly above chance level, we can reject the null hypothesis of no association. This test is straightforward for balanced data, where all groups are equally frequent in the data set. However, data sets collected in observational studies are often imbalanced. Then accuracy is no longer a suitable measure of performance, and balanced accuracy should be used instead. In this paper, we give an overview on existing analytical tests and use the framework of prediction theory to derive a new test for the balanced accuracy of a classifier. We then use numerical simulations to evaluate the type I error rate and the power of two tests for imbalanced data.**

I. Introduction

In many scientific fields, researchers are confronted with the question whether there exists an association between a set of variables X and a factorial variable Y. They want to answer this question using a dataset $(X_i, Y_i), i = 1 \dots n$, that was either obtained observationally or by means of an experimental study. In tumor classification from microarray data [1]–[3] the goal is to check for associations between the expression levels X of genes and the tumor type Y. In the multivariate analysis of functional magnetic resonance imaging (fMRI) data [4], [5] the question is whether the measured blood oxygenation level dependent signal within a region of interest (such as certain brain area, or a local search light) is associated to a particular factor variable, such as experimental condition or subject group.

The null hypothesis is that there is no association between X and Y, and the task is to test whether this null hypothesis can be rejected at a given significance level. Since the existence of an association between X and Y implies that X is predictive of Y, machine learning methods can be applied to try to establish an association. If a classifier can be learned that predicts Y from X better than chance in the population, the null hypothesis of no association can be rejected. This is called classifier-based association testing. The parameters of a classification model are learned on a training set, and the model is used to predict the class labels of the examples in a test set. The task is to assess whether the predictions of the classifier are accurate enough that it seems unlikely that this could have been purely achieved by chance. In other words, we want to test whether the null hypothesis that the true accuracy μ of the classifier is not above chance level can be rejected at a chosen significance level δ, i.e $H_0 : \mu \leq 0.5$ versus $H_1 : \mu > 0.5$. Since the true accuracy of a classifier cannot be observed, one needs to define a hypothesis test based on the predictions the classifier made on a test set. For balanced class labels and binary classification, testing whether the classification results on the test set allow to reject the null hypothesis of no association between the multi-dimensional inputs and the binary group variable (class label) is rather straightforward. Two approaches will be briefly reviewed in section II. The first approach is based on Bayesian statistics, the second on the test set bound from prediction theory [6].

In experimental studies, the experimental design usually ensures that the data sets are balanced with respect to the group variables. However, for observational data, one group is often more prevalent than the other, making the data set imbalanced. This affects classifier-based association testing, since for such imbalanced data, the accuracy is not a suitable performance measure any longer. The reason for this is that high accuracy values can already be obtained by assigning all test data into the larger class, without taking the inputs even into account. Therefore, a better performance measure should be used, in which such a strategy would gain you nothing. One such measure is the balanced accuracy, which is defined as the arithmetic mean of sensitivity (percentage of the positive class that are correctly classified as positive) and specificity (percentage of the negative class that are correctly classified as negative). Classifying everything into the larger class results in a balanced accuracy of 50%. If the true balanced accuracy of a binary classifier is above 50%, we can claim to have learned a dependency between the input variables and the class variable.

In this work, we focus on classifier-based association tests for imbalanced data. We will first describe a test that is based on the posterior probability of the true balanced accuracy [7]. Then we will derive an alternative test within the framework of prediction theory. For this, we extend the test set bound on the generalization performance of a classifier [6] to the case of imbalanced class labels. This confidence bound is then used to derive a test for assessing whether the empirical balanced

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

accuracy is significantly higher than chance level. We evaluate the type I error rates and the power at different signal to noise levels for both the prediction theoretic test and the Bayesian posterior-based test on data sets of different class label ratios.

II. Balanced Data

A. Test for balanced data based on posterior of true accuracy

The first approach is based on calculating the posterior distribution of the true accuracy given the observed classification and the ground truth [4], [7]. It considers the prediction on the test examples as a series of Bernoulli experiments, in which balls are drawn with replacement from a bucket with an unknown mixture of "correct" and "incorrect" balls. The (unknown) proportion of correct balls in the bucket corresponds to the true accuracy μ. The task is to estimate the value of μ from the observations. The number of correct balls k in a series of n trials can be modeled by a binomial distribution,

$$\pi(k|\mu, n) = \binom{n}{k}\mu^k(1-\mu)^{n-k}. \tag{1}$$

The posterior distribution for μ is determined by the choice of prior. The conjugate prior for the binomial distribution is the beta distribution [8],

$$\pi(\mu, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}, \tag{2}$$

where $\Gamma(\cdot)$ is the Gamma function, and a and b are hyperparameters. The posterior distribution is again a beta distribution of the following form [8]

$$\begin{aligned}\pi(\mu|k, n, a, b) = &\frac{\Gamma(k+a+(n-k)+b)}{\Gamma(k+a)\Gamma(n-k+b)} \\ &\cdot\mu^{k+a-1}(1-\mu)^{n-k+b-1}.\end{aligned} \tag{3}$$

One can see that a and b have the effect of virtual observations in each of the two classes. In [7], the use of a flat prior for μ was suggested, which corresponds to $a = 1, b = 1$. This choice of prior therefore implies that we pretend to have drawn one correct and one incorrect ball before we observe our actual data. In this case the posterior, eq. (3), takes the form

$$\pi(\mu|k, n) = \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)}\mu^k(1-\mu)^{n-k}. \tag{4}$$

The p-value for the association test is then given by the probability that the true accuracy is smaller or equal to 0.5 (the chance level). This corresponds to the area under the left tail of the posterior distribution cut off at $\mu = 0.5$,

$$p = \int_0^{0.5} \pi(\mu|k, n)d\mu. \tag{5}$$

B. Test for balanced data based on prediction theory

The second approach is based on the test set bound that was derived in the framework of prediction theory [6]. This bound makes a statement about the future error rate of a trained classifier. Assume that the classifier has the error rate $1-\mu$, where μ is the true accuracy. Using the binomial distribution (eq. 1), the probability of making k or fewer errors on n examples is given by the binomial tail distribution,

$$Bin(n, k, 1-\mu) \equiv \sum_{j=0}^{k} \pi(j|1-\mu, n). \tag{6}$$

Then one can define the Binomial tail inversion,

$$\overline{Bin}(n, k, \delta) \equiv \max\{p : Bin(n, k, p) \geq \delta\}, \tag{7}$$

as the largest true classifier error rate such that the probability of observing k or more errors is at least δ.

The test set bound [6] states that if μ_S is the empirical accuracy on a test data set of size n drawn from a distribution D, then the probability of having a true error rate $1-\mu$ that is less than or equal to the Binomial tail inversion is greater than or equal to $1-\delta$.

Theorem II.1. *(Test Set Bound) For all classifiers f, all distributions D and all $\delta \in (0, 1]$*

$$\underset{S\sim D^n}{P}(1-\mu \leq \overline{Bin}(n, 1-\mu_S, \delta)) \geq 1-\delta. \tag{8}$$

The test set bound can be interpreted as a game where a "learner" tries to convince a reasonable "verifier" of the amount of learning which has occurred, however it is essential that the test examples are unknown to the learner [6]. The test set bound is a tight bound, i.e. if the true error is sufficiently large, it is violated exactly in a δ-portion of trials [6].

In an empirical investigation of several bounds on the true error rate of learned classifiers [9] it was found that there is trade-off between obtaining a high classification accuracy and having high confidence in the accuracy, and the test set bound was recommended for situations where the confidence is very important.

The test set bound can be used to assess whether a significant association could be established by a classifier: If, for a given δ, the binomial tail inversion $\overline{Bin}(n, 1-\mu_S, \delta)$ lies above the 50% error rate corresponding to chance level, then the null hypothesis that the classifier does not classify better than chance cannot be rejected at level δ.

III. Imbalanced data

A. Test for imbalanced data based on posterior of true accuracy

The method in section II-A was recently extended to the case of imbalanced class labels [7]. If q_1 is the true accuracy on the positive class and q_2 is the true accuracy on the negative class, the balanced accuracy is given by $\eta = 0.5(q_1 + q_2)$. Using a beta distribution with $a = 1, b = 1$ as prior for both positive and negative class, in [7] the posterior density of the true balanced accuracy was derived as

$$\begin{aligned}\pi(\eta|k_1, n_1, k_2, n_2) = \int_0^1 &\pi(2(\eta - z)|k_1 + 1, n_1 + 1) \\ &\cdot\pi(2z|k_2 + 1, n_2 + 1)dz,\end{aligned} \tag{9}$$

where k_1 are the true positives, k_2 the true negatives, n_1 is the number of positive examples in the test set, and n_2 is

the number of negative examples in the test set. Thus for imbalanced class labels the p-value that the predicted outcome was generated by a classifier with a true balanced accuracy of at most chance level is obtained as

$$p = \int_0^{0.5} \pi(\eta|k_1, n_1, k_2, n_2)d\eta. \tag{10}$$

B. Prediction theoretic test for imbalanced data

In the following, we first derive the test set bound for the balanced accuracy, then we use this to obtain a significance test for the empirical balanced accuracy of a classifier.

1) Test set bound for imbalanced data: Consider a classification function $f : X \to Y = \{-1, 1\}, x \mapsto y$ that maps an input space X to a binary output space Y. We assume an unknown underlying distribution D over $X \times Y$, from which a test set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ of n examples has been drawn independently. We further assume that from these examples, n_1 have the class label 1 and n_2 have the class label -1.

Definition III.1. *(True Balanced Accuracy) The true balanced accuracy η of the classifier is defined as*

$$\eta \equiv \frac{1}{2}\left(\underset{(x,1)\sim D}{P}(f(x) = 1) + \underset{(x,-1)\sim D}{P}(f(x) = -1)\right) \tag{11}$$

Thus η corresponds to the average of specificity and sensitivity. While the true balanced accuracy is not observable, one can use the prediction of the classifier on the test set S to obtain an estimate $\hat{\eta}_S$.

Definition III.2. *(Empirical Balanced Accuracy) The empirical balanced accuracy $\hat{\eta}$ of the classifier is defined as*

$$\begin{aligned}\hat{\eta}_S \equiv \frac{1}{2}&\left(\underset{(x,1)\sim S}{P}(f(x) = 1) + \underset{(x,-1)\sim S}{P}(f(x) = -1)\right) \\ &= \frac{1}{2n_1}\sum_{\{i|y_i=1\}} I(f(x_i) = 1) + \\ &\frac{1}{2n_2}\sum_{\{j|y_j=-1\}} I(f(x_j) = -1),\end{aligned} \tag{12}$$

where I is an indicator function,

$$I(a) \equiv \begin{cases} 1 & \text{if } a = true, \\ 0 & \text{if } a = false. \end{cases} \tag{13}$$

We can consider the prediction on the test examples as a series of Bernoulli experiments. Let q_1 and q_2 be the true accuracies of a classifier for each of the two classes. Getting a certain number of correct predictions for each of the two classes can be considered as two independent events. Therefore the probability of making k_1 correct predictions on the n_1 examples in class 1 and at the same time making k_2 correct predictions on the n_2 examples in class -1 is given by a product of two binomial distributions

$$\begin{aligned}&\pi(k_1, k_2|q_1, n_1, q_2, n_2) = \pi(k_1|q_1, n_1) \cdot \pi(k_2|q_2, n_2) \\ &= \binom{n_1}{k_1} {q_1}^{k_1}(1 - q_1)^{n_1-k_1}\binom{n_2}{k_2}{q_2}^{k_2}(1 - q_2)^{n_2-k_2}\end{aligned} \tag{14}$$

Then the probability of the event that the classifier obtains an empirical balanced accuracy of at least $\hat{\eta}_S$ on a test set with n_1 positively and n_2 negatively labeled examples is calculated as

$$\begin{aligned}&\pi(\hat{\eta} \geq \hat{\eta}_S|q_1, q_2, n_1, n_2) \\ &= \sum_{k_1=0}^{n_1}\sum_{k_2=0}^{n_2} I\left(\frac{1}{2}(\frac{k_1}{n_1} + \frac{k_2}{n_2}) \geq \hat{\eta}_S\right)\pi(k_1, k_2|q_1, n_1, q_2, n_2)\end{aligned} \tag{15}$$

where the indicator function $I(\cdot)$ ensures that only terms with $\frac{1}{2}(\frac{k_1}{n_1} + \frac{k_2}{n_2}) \geq \hat{\eta}_S$ are included in the summation.

Definition III.3. *(Likelihood) The Likelihood $L_{\hat{\eta}_S,n_1,n_2}(q_1, q_2)$ is a function of (q_1, q_2) that maps (q_1, q_2) to the probability that the event $\hat{\eta} \geq \hat{\eta}_S$ occurs, if drawing from the distribution $\pi(k_1, k_2|q_1, n_1, q_2, n_2)$.*

$$\begin{aligned}L_{\hat{\eta}_S,n_1,n_2} \quad &: \quad [0, 1] \times [0, 1] \to [0, 1] \\ &(q_1, q_2) \mapsto \pi(\hat{\eta} \geq \hat{\eta}_S|q_1, q_2, n_1, n_2)\end{aligned} \tag{16}$$

On a test data set sampled from D, which contains n_1 examples from class 1 and n_2 examples from class -1, the classifier f will reach an empirical balanced accuracy of at least $\hat{\eta}_S$ with probability $L_{\hat{\eta}_S,n_1,n_2}$, if the true accuracies for class 1 and class -1 are q_1 and q_2, respectively. Examples for the likelihood function for three different empirical balanced accuracy are shown in Fig. 1.

Definition III.4. *(Delta Superlevel Set) The delta superlevel set of $L_{\hat{\eta}_S,n_1,n_2}$ at δ is defined as*

$$\Gamma^+_\delta(L_{\hat{\eta}_S,n_1,n_2}) \equiv \{(q_1, q_2)|L_{\hat{\eta}_S,n_1,n_2}(q_1, q_2) \geq \delta\} \tag{17}$$

The delta superlevel set $\Gamma^+_\delta(L_{\hat{\eta}_S,n_1,n_2})$ is the set of pairs of true class-based accuracies (q_1, q_2) for which the probability of having $\hat{\eta} \geq \hat{\eta}_S$ is at least δ.

Definition III.5. *(Delta Level Set) The delta level set of $L_{\hat{\eta}_S,n_1,n_2}$ is defined as the set of all (q_1, q_2) where the function $L_{\hat{\eta}_S,n_1,n_2}$ takes on the value $\delta \in (0, 1]$:*

$$\Gamma_\delta(L_{\hat{\eta}_S,n_1,n_2}) \equiv \{(q_1, q_2)|L_{\hat{\eta}_S,n_1,n_2}(q_1, q_2) = \delta\} \tag{18}$$

The delta level set $\Gamma_\delta(L_{\hat{\eta}_S,n_1,n_2})$ is an algebraic curve in the $[0, 1] \times [0, 1]$ plane that forms the lower boundary of $\Gamma^+_\delta(L_{\hat{\eta}_S,n_1,n_2})$.

Theorem III.1. *(Test Set Bound for Imbalanced Data) For all classifiers f, all distributions D, all data sets S with n_1 positive labels and n_2 negative labels sampled from D, and for all $\delta \in (0, 1]$*

$$\underset{S\sim D}{P}((q_1, q_2) \in \Gamma^+_\delta(L_{\hat{\eta}_S,n_1,n_2})) \geq 1 - \delta. \tag{19}$$

Proof: Irrespective of the true combination of (q_1, q_2), the observation η_S will not fall into the tail of size δ of the distribution $\pi(\hat{\eta} \geq \hat{\eta}_S|q_1, q_2, n_1, n_2)$ (15) with probability $1 - \delta$. Therefore the true values of (q_1, q_2) have to be above or on the curve of $\Gamma_\delta(L_{\hat{\eta}_S,n_1,n_2})$ with confidence $1 - \delta$. ■

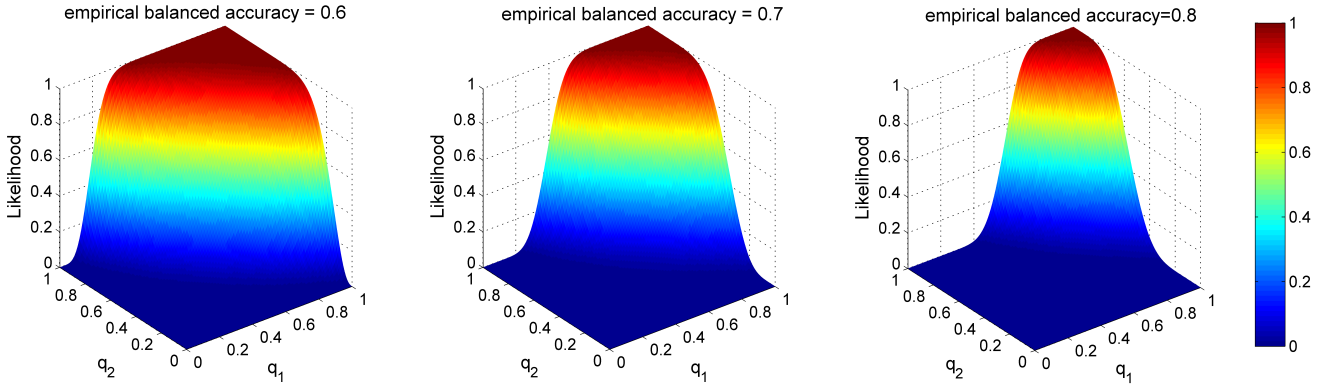

Fig. 1. The Likelihood $L_{\hat{\eta}_S,n_1=35,n_2=15}$ as function of q_1 and q_2, for three different values of the empirical balanced accuracy $\hat{\eta}_S$.

2) Significance test for the balanced accuracy: In order to test the significance of the observed empirical balanced accuracy, we need to test whether the null hypothesis that the true balanced accuracy is lower than chance level, i.e. $\eta \leq 0.5$ can be rejected at level δ. In the space $q_1 \times q_2$, the condition $\eta = 0.5(q_1 + q_2) = 0.5$ corresponds to the line $q_1 + q_2 = 1$. All points (q_1, q_2) on and below that line mark the values of q_1, q_2 corresponding to the above null hypothesis.

Note that testing whether the balanced accuracy is above chance level is not equivalent to testing whether the predictions of each of the classes are Bernoulli unbiased. The balanced accuracy η is defined as the average of the class-based accuracies q_1 and q_2, i.e. $\eta = 0.5(q_1+q_2)$. Even if q_1 and q_2 are Bernoulli biased (i.e. different from 0.5), the balanced accuracy can still be at chance level ($\eta = 0.5$). In fact, all combinations of q_1 and q_2 along the line $q_1 + q_2 = 1$ correspond to an unbiased balanced accuracy. Therefore, it is not sufficient to test both q_1 and q_2 individually for bias. Instead, one has to test for the combination η, which makes the problem non-trivial.

Because $L_{\hat{\eta}_S,n_1,n_2}(q_1, q_2)$ is monotonously increasing with q_1 and q_2, it is sufficient to test whether any of the points on the diagonal $q_1 + q_2 = 1$ lie within $\Gamma^+_\delta(L_{\hat{\eta}_S,n_1,n_2})$ (Fig. 2). This can be done by finding the maximum of $L_{\hat{\eta}_S,n_1,n_2}(q_1, q_2)$ along the line $q_2 = 1 - q_1$ (Fig.3), and checking whether it is larger or equal to δ. In that case, the null hypothesis cannot be rejected at level δ.

Thus, the p-value is obtained as

$$p = \max_{q_1} L_{\hat{\eta}_S,n_1,n_2}(q_1, 1 - q_1). \tag{20}$$

With eqns. (15) and (14) we get

$$p = \max_{q_1} \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} I\left(\frac{1}{2}(\frac{k_1}{n_1} + \frac{k_2}{n_2}) \geq \hat{\eta}_S\right) \binom{n_1}{k_1} {q_1}^{k_1} (1 - q_1)^{n_1-k_1} \binom{n_2}{k_2} (1 - q_1)^{k_2} (q_1)^{n_2-k_2}. \tag{21}$$

This can also be interpreted as the maximum over all values q_1 of the probability that a reasonable "verifier" obtains a balanced accuracy as high as the one observed by flipping a coin with bias q_1 repeatedly, n_1 times for the positive class and n_2 times for the negative class.

Summarizing the factors, we can obtain the p-value for the prediction theoretic significance test in the case of imbalanced class-labels as

$$p = \max_{q_1} \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} I\left(\frac{1}{2}(\frac{k_1}{n_1} + \frac{k_2}{n_2}) \geq \hat{\eta}_S\right) \binom{n_1}{k_1} \binom{n_2}{k_2} {q_1}^{k_1+n_2-k_2} (1 - q_1)^{k_2+n_1-k_1}. \tag{22}$$

IV. Evaluation

We conducted simulation studies to assess the type I error rate and the power of the prediction theoretic test and the Bayesian posterior based test at different significance levels. The power corresponds to the probability that the test correctly rejects the null hypothesis. If the probability for a type II error is given by the false negative rate β, then the power is $1 - \beta$. Since the power depends on the relative amount of noise, we investigated the power for three different relative noise levels. It is important to strictly bound the probability of finding a spurious association, i.e. the type I error rate, by setting the significance level δ. Therefore the type I error rate should always be lower than or equal to the chosen significance level. We investigated whether this strictly holds for the two tests.

The simulations were conducted as follows. For each scenario we generated 100,000 data sets describing a 2-class problem. For each class i, we randomly sampled the data from a 20-dimensional isotropic Gaussian distribution with class center $\mathbf{c}_i$ and variance $\sigma^2 = 100$. The class centers were at positions $\mathbf{c}_1 = (\nu, \nu, \ldots, \nu)$ for the positive class, and $\mathbf{c}_2 = (-\nu, -\nu, \ldots, -\nu)$ for the negative class. Both the training and test each contained $m = 200$ data points, m_1 for class 1 and m_2 for class -1. We investigated four different levels of class imbalance, with $(m_1, m_2) \in \{(100, 100), (140, 60), (160, 40), (180, 20)\}$. The first case corresponds to a balanced data set, the last case to a heavily imbalanced one.

The relative noise level of the data was varied by changing the class separation, i.e. the position of the class centers, by

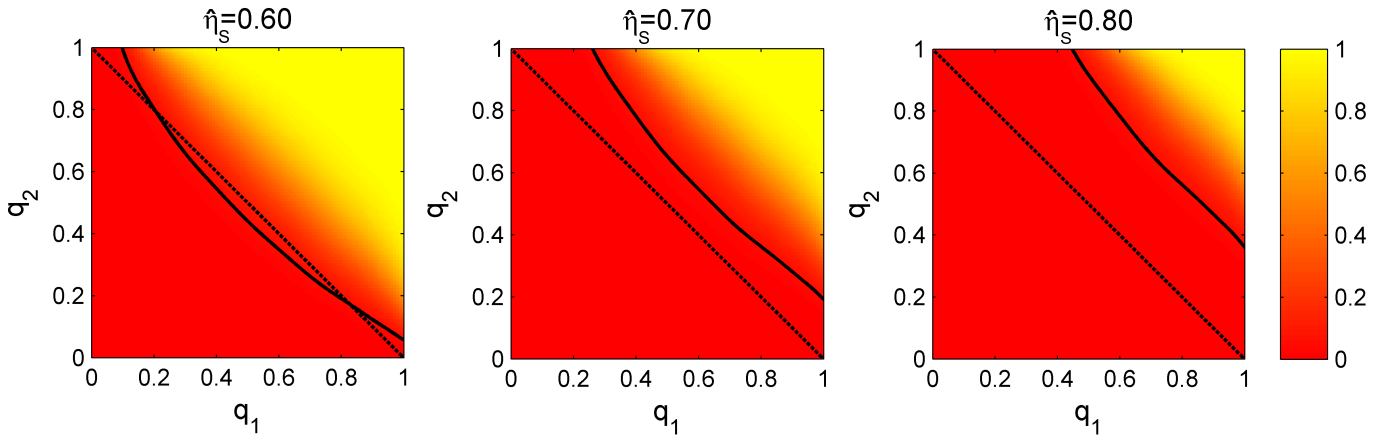

Fig. 2. This figure shows $L_{\hat{\eta}_S,n_1=35,n_2=15}(q_1,q_2)$ in $[0,1]\times[0,1]$ for three different values of the empirical balanced accuracy $\hat{\eta}_S$. The solid curve denotes the delta level set at $\delta=0.05$. The dotted curve denotes the line $(1/2)(q_1+q_2)=0.5$, where the balanced accuracy is at chance level. For $\hat{\eta}_S=0.6$ the two curves cross, which means that the null hypothesis that the balanced accuracy of the classifier is at or below chance level cannot be rejected.

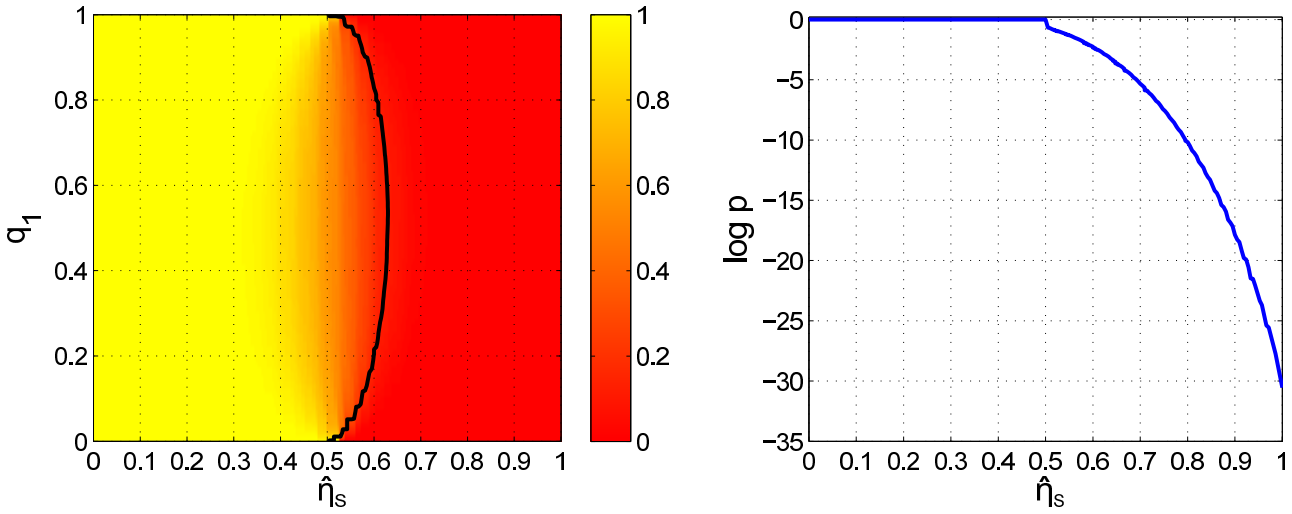

Fig. 3. Left: The function $L_{\hat{\eta}_S,n_1=35,n_2=15}(q_1,q_2)$ for all points on the line $q_2=1-q_1, q_1\in[0,1]$ for different values of the empirical balanced accuracy $\hat{\eta}_S$. The black curve shows the delta level set at $\delta=0.05$. Right: The logarithm of the p-value that corresponds to the maximum of the function $L_{\hat{\eta}_S,n_1=35,n_2=15}(q_1,q_2)$ over the line $q_2=1-q_1, q_1\in[0,1]$ for different values of $\hat{\eta}_S$.

varying $\nu\in\{1,2,3\}$. The distance between the class centers is given by $\Delta=\sqrt{20(2\nu)^2}=\sqrt{80}\nu$. Thus increasing ν decreased the ratio between within class variance and class separation, i.e. the relative noise level.

In order to analyze the type I error rate of the tests, both classes were sampled from the same 20-dimensional isotropic Gaussian distribution with variance $\sigma^2=100$ distribution centered at the origin, i.e. $\mathbf{c}_1=\mathbf{c}_2=0$.

As classifier we employed a C-SVM with linear kernel and fixed $C=0.01$. Each classifier was learned on the training set and used on the test set for prediction. Then both, the prediction theoretic tests and the test based on the Bayesian posterior, were applied at different significance levels[1].

For each data set there was either a true association (for $\nu\in\{1,2,3\}$), or there was no association (for $\nu=0$). The type I error rate was calculated for each significance level δ as the proportion of the data sets with no association for which the null hypothesis was (falsely) rejected at level δ. The power was calculated at different noise levels $\nu\in\{1,2,3\}$ and for each significance level δ as the proportion of the data sets for which the null hypothesis could not be rejected at level δ. The results for both type I error rate and power of the classifiers are reported in Table I for the test using the Bayesian posterior

[1]For the test based on the Bayesian posterior, we used the Matlab-code available at http://people.inf.ethz.ch/bkay/downloads.html

and in Table II for the prediction theoretic test. Type I error rates that were correctly below the significance level δ are marked in bold font.

For the prediction theoretic test, the type I error rate never exceeded the specified significance level. In contrast, the test based on the Bayesian posterior displayed in many cases slightly inflated type I error rates that exceeded the significance level δ. However, the test based on the Bayesian posterior had in many cases slightly more power than the prediction theoretic test. The difference in power of two tests increased with larger class imbalance and with higher noise level. In summary, these results show that the prediction theoretic test is more conservative than the posterior based test.

V. Discussion

Classifier-based association tests are employed to establish the existence of a significant group difference on data consisting of multidimensional vectors, such as functional magnetic resonance imaging or gene expression patterns. In this paper, we first gave a brief review of two association tests for balanced data. The main focus was, however, on the case of imbalanced data, which often arises from observational studies. The main difference to the balanced case is that it requires the use of a special performance measure, such as the balanced accuracy. The main contribution of this paper is a classifier-based association test for imbalanced data that was derived in the framework of prediction theory.

As a first step, we extended the test set bound for binary classifiers [6] to the case of imbalanced data. This bound was then used to derive a hypothesis test that allows to assess the significance of the balanced accuracy obtained by a classifier. We compared this prediction theoretic test to a previously derived test based on the Bayesian posterior of the true balanced accuracy by evaluating the type I error rate and power at different levels of relative noise. The analysis showed that the type I error rate of the prediction theoretic test never exceeded the significance level, whereas the posterior-based test sometimes showed a slight inflation of type I error rates. However, while not guaranteeing strict type I error rates, the Bayesian posterior based test had slightly more power than the test derived from prediction theory. So if strict type I error control is desired, the more conservative prediction theoretic test might be preferred, at the cost of slight losses in power.

For classifier-based tests the power will depend on the suitability of the classification model, the performance of the learning algorithm, the size of the training and test data sets, and the relative noise level in the data. If the classification model is not flexible enough to capture the underlying dependency, than it is unlikely that the null hypothesis can be rejected, and the power of the test will be low. If the model is too flexible, it is likely to adapt to any noise in the data, and again, the power of the test will be reduced. Usually, if given some arbitrary data set, it is not a priori known what a suitable classification model will be. Therefore a good strategy would be to start with a simple model, for example a linear classifier, and test whether the null hypothesis can be rejected. If this fails one can try a more flexible classification model. Multiple testing issues resulting from such a procedure can be taken into account by family-wise error correction or by controlling the false discovery rate. Alternatively, one could select the model complexity using an inner cross-validation loop.

Both tests assume that the Bernoulli trials are independent, which is only the case if the test set is completely independent from the training set. They cannot be applied to cross-validation results, since the overlap of the training sets and the fact that the test points of one fold are contained in the training sets of the other folds induce a dependency structure that increases the variance of the point estimates [3], [10].

Acknowledgment

This work was supported by the BMBF (grant numbers 01GQ0911 and 01ZX1311D).

References:

[1] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. Mcshane, "Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification." *J Natl Cancer Inst*, vol. 95, no. 1, pp. 14–18, Jan. 2003.

[2] U. W. Bolin, H. Goransson, M. Fryknas, M. Gustafsson, and A. Isaksson, "Improved variance estimation of classification performance via reduction of bias caused by small sample size," *BMC Bioinformatics*, vol. 7, no. 1, p. 127 pp., Mar. 2006. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-7-127

[3] A. Isaksson, M. Wallman, H. Goransson, and M. Gustafsson, "Cross-validation and bootstrapping are unreliable in small sample classification," *Pattern Recognition Letters*, vol. 29, no. 14, pp. 1960–1965, Oct. 2008. [Online]. Available: http://dx.doi.org/10.1016/j.patrec.2008.06.018

[4] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: A tutorial overview," *NeuroImage*, 2009.

[5] K. H. Brodersen, T. M. Schofield, A. P. Leff, C. S. Ong, E. I. Lomakina, J. M. Buhmann, and K. E. Stephan, "Generative embedding for model-based classification of fmri data." *PLoS Computational Biology*, vol. 7, no. 6, 2011.

[6] J. Langford, "Tutorial on practical prediction theory for classification," *Journal of Machine Learning Research*, vol. 6, pp. 273–306, 2005.

[7] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," in *2010 International Conference on Pattern Recognition*. IEEE computer society, 2010, pp. 3121–3124.

[8] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[9] M. Kääriäinen and J. Langford, "A comparison of tight generalization error bounds," in *Proceedings of the 22nd international conference on Machine learning*, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 409–416. [Online]. Available: http://doi.acm.org/10.1145/1102351.1102403

[10] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *J. Mach. Learn. Res.*, vol. 5, pp. 1089–1105, Dec. 2004. [Online]. Available: http://dl.acm.org/citation.cfm?id=1005332.1044695

TABLE I
TEST BASED ON BAYESIAN POSTERIOR

m_1/m_2			δ				
			0.01	0.02	0.03	0.04	0.05
100/100	Type I error rate		**0.0100**	0.0202	**0.0289**	**0.0394**	0.0522
	Power	$\nu = 1$	0.9131	0.9474	0.9605	0.9703	0.9778
		$\nu = 2$	1	1	1	1	1
		$\nu = 3$	1	1	1	1	1
140/60	Type I error rate		0.0104	0.0210	0.0308	0.0410	0.0516
	Power	$\nu = 1$	0.8268	0.8865	0.9144	0.9311	0.9438
		$\nu = 2$	1	1	1	1	1
		$\nu = 3$	1	1	1	1	1
160/40	Type I error rate		0.0108	0.0220	0.0318	0.0424	0.0522
	Power	$\nu = 1$	0.6423	0.7337	0.7819	0.8144	0.8401
		$\nu = 2$	0.9999	0.9999	1	1	1
		$\nu = 3$	1	1	1	1	1
180/20	Type I error rate		0.0114	0.0219	0.0329	0.0436	0.05488
	Power	$\nu = 1$	0.2818	0.3649	0.4299	0.4771	0.5175
		$\nu = 2$	0.9606	0.9766	0.9838	0.9875	0.9900
		$\nu = 3$	0.9998	1	1	1	1

TABLE II
TEST BASED ON PREDICTION THEORY

m_1/m_2			δ				
			0.01	0.02	0.03	0.04	0.05
100/100	Type I error rate		**0.0099**	**0.0142**	**0.0286**	**0.0390**	**0.0390**
	Power	$\nu = 1$	0.9126	0.9318	0.9604	0.9701	0.9701
		$\nu = 2$	1	1	1	1	1
		$\nu = 3$	1	1	1	1	1
140/60	Type I error rate		**0.0100**	**0.0188**	**0.0283**	**0.0374**	**0.0490**
	Power	$\nu = 1$	0.8179	0.8757	0.9065	0.9241	0.9387
		$\nu = 2$	1	1	1	1	1
		$\nu = 3$	1	1	1	1	1
160/40	Type I error rate		**0.0087**	**0.0180**	**0.0275**	**0.0353**	**0.0474**
	Power	$\nu = 1$	0.6008	0.6979	0.7505	0.7836	0.8195
		$\nu = 2$	0.9997	0.9999	0.9999	0.9999	1
		$\nu = 3$	1	1	1	1	1
180/20	Type I error rate		**0.0090**	**0.0167**	**0.0266**	**0.0361**	**0.0458**
	Power	$\nu = 1$	0.2304	0.3007	0.3651	0.4152	0.4525
		$\nu = 2$	0.9112	0.9413	0.9583	0.9682	0.9738
		$\nu = 3$	0.9972	0.9986	0.9992	0.9995	0.9996