Comparing Classifier's Performance Based on Confidence Interval of the ROC

Tobias MALACH, Jitka POMENKOVA

Dept. of Radio Electronics, Brno University of Technology, Technicka 12, 612 00 Brno, Czech republic

tmalach@phd.feec.vutbr.cz, pomenkaj@feec.vutbr.cz

Submitted November 6, 2017 / Accepted June 18, 2018

Abstract. *This paper proposes a new methodology for comparing two performance methods based on confidence interval for the ROC curve. The methods performed and compared are two algorithms for face recognition. The novelty of the paper is three-fold: i) designing a methodology for the comparison of decision making algorithms via confidence intervals of ROC curves; ii) investigating how sample sizes influence the properties of the particular methods; iii) recommendations for a general comparison of decision making algorithms via confidence intervals of ROC curves. To support our conclusions we investigate and demonstrate several approaches for constructing parametric confidence intervals on real data. Thus, we present a non-traditional and reliable way of reporting pattern recognition results using ROC curves with confidence intervals.*

Keywords

Confidence interval, ROC curves, face recognition, pattern recognition

1. Introduction

The Receiver Operating Characteristic (ROC) provides a simple and useful instrument for the diagnostic and overview of the overall statistical test performance. The ROC curve is a graphical representation of the relationship between sensitivity and specificity with respect to the cut-off point which runs through the whole range of possible values [1], [2]. ROC curves have been used in natural sciences such as medicine, biology or chemistry [3], [4] as well as in signal processing, machine learning, pattern recognition [6] and other technical sciences [5]. While in engineering it is possible to obtain a large sample of data, in medicine or biology the samples are moderate or small. Generally, statistical tests conducted on small samples can be biased and can lead to incorrect conclusions. Large samples are therefore preferable.

We investigate methods for estimating the ROC curve confidence in the field of machine learning. This research was motivated by the need of a comparison between two decision making algorithms. Their performance is usually presented in the form of ROC curves. Analysing ROC curve confidence enables us to compare classifiers and estimate statistical accuracy of the reported classifier performance. The confidence of an ROC curve is therefore essential for developing classifiers, comparing their performance and optimizing classifier parameters in all disciplines.

In this paper we investigate the performance of two machine vision algorithms for face recognition. We enumerate the confidence of the results reached by each algorithm using several methods. Each method has different properties and thus the enumerated confidence slightly differs. We show that even if a data set contains thousands of samples it is worth investigating the confidence of the results. The construction of confidence intervals (CIs) allows us to decide if the examined face recognition algorithms performs better. This is one of the first in-depth works describing ROC curve confidence in depth in the field of face recognition or pattern recognition. We would like to encourage researchers in the field of pattern recognition and machine learning to adopt this methodology for ROC curve confidence, especially when tests are conducted on non-public data-sets.

2. Contemporary Methods for ROC Comparison

ROC curves have been used as a diagnostic instrument, however, they are also useful for the comparison and performance evaluation of decision making algorithms. There are generally two approaches to make the comparison. The first one is numerical, employing the Area Under the ROC Curve (AUC) [3, 4, 6], the other one is graphical. The latter shows the sample variability of the ROC curve estimate by plotting bounds around the ROC curves [1] and the comparison is performed by evaluating the overlap of CIs.

The AUC represents the probability that a randomly chosen object of group A is (correctly) rated with a higher probability to group A than a randomly chosen object of group B. Thus, a comparison is performed on the comparison of CIs for the AUC value [7–10]. A detailed look into the calculation of AUC uncovers a possible inaccuracy of its enumeration which is usually based on the Mann-Whitney statistic [10]. Further, Westin [8] and Brown and Davis [10]

proposed a calculation of the AUC and its CI. The work of Zou et al. [11] focused on the comparison of multiple AUCs derived from the same group of objects. They presented a method using simultaneous CIs for AUC and investigated a small sample size. The comparison via AUC is a two-step algorithm (the first one is the approximation of AUC, the second one its testing) which can bring additional imperfections and statistical inaccuracies to the enumeration and estimation compared to a one-step approach. As Westin [8] pointed out, a non-significant difference between AUCs for two methods does not imply an equivalence between these methods. A different point of view on AUC was provided by [12] which focused on the fusion of methods for different detectors to improve their individual performance. The authors investigated the maximization of AUC as one of the instruments for optimizing the fusion. Maintaining the aim of this paper we tend to use CIs for ROC curves as they can better visualize the behavior of different decision making algorithms. The comparison of AUCs confidence intervals is not suitable for studying overlaps in detail, thus, we do not investigate it.

Another approach reflected in the literature is a one- or two-dimensional CI for each point of an ROC curve (for sensitivity and specificity separately). Since sensitivity and specificity are both proportions, we can calculate CI using the standard methods for proportions. Two types of 95% CI are generally constructed around the proportion - the asymptotic and the exact 95% one. The exact CI is constructed using binomial distribution while the asymptotic approach assumes a normal approximation of sampling distribution. Asymptotic types of the procedure based on a simple normal approximation of the binomial distribution for CI construction were proposed in [1]. Further, Schäfer [1] presented „Greenhouse and Mantel" confidence bounds based on a statistical test which was 40% smaller than the interval constructed as a combination of separate CIs for sensitivity and specificity derived as a normally distributed relative frequency parameter. Brown and Davis [10] discussed methods for determining CI on various measures. Their paper proposed two expressions of CI of the proportion with the continuity correction term. They considered joint confidence intervals, bootstraping, and regression. Westin [8] summarized the basic approach to the ROC curve construction, its CI and the comparison of two curves via AUCs testing. She provided formulas for standard error of sensitivity and specificity and discussed the influence of sampling.

An alternative approach to the construction of ROC CIs can be performed by using a regression model (and the construction of its confidence band) or a non-parametric kernel estimate. In a similar way to the AUC, the confidence band (CB) for regression curves is a two-step algorithm. First, a regression model of ROC curve is estimated, subsequently, the CB or AUC is calculated. An advantage of such modeling is its usability for the AUC as well as for the CBs calculations. The enumeration of AUC, if the ROC trend is described by the function, could be more precise compared to an approximation in the ROC points. In the case of a sufficiently large sample, it is suitable to use the smoothing of ROC curve points via a non-parametric kernel estimation such as [13], [14]. To apply the kernel method, we need the optimization of its parameters. As Hall [13] proved, the asymptotic method can provide very good performance. Unfortunately, a disadvantage of this approach is the need of a large sample of ROC points and optimum choice of the bandwidth, which is computationally time demanding. As written in [14], the kernel methods usually struggle with the edge problem (when significant parts of a histogram are near zero). Therefore, this approach is not investigated in this paper.

2.1 Paper Contribution and Organization

We have determined that literature does not deal with the following: firstly, how the ROC curve, its CI and AUC can be used for an objective comparison, i.e. to answer the question whether the difference between curves is statistically significant; secondly, what the conditions and recommendations for using such instruments for comparison are. In this paper, ROC curves are used for evaluating methods for face recognition. The task is a reliable identification of a method which makes the whole recognition process more successful. We suggest, for this purpose, a novel methodology based on the CI estimation of ROC curves and the subsequent assessment of possible overlap of CIs. The main contributions of the paper are:

- an algorithm for the comparison of decision making algorithms via CIs of ROC curves,
- a demonstration of several approaches for the parametric CI construction on real data,
- an investigation of how a sample size influences the properties of discussed approaches,
- a recommendation for a general comparison of decision making algorithms via CI of ROC curves.

The demonstration of the proposed algorithm, and consequent results discussion, are shown on the comparison of two machine vision algorithms for face recognition.

3. ROC Curves

The ROC was developed in the field of statistical decision theory [15]. Let us assume two groups of objects. One group of objects satisfies a condition (the positive case) and the other group does not (the negative case). We take a test which denotes an object as positive or negative. Further we will use the following denotation: true positive (TP), true negative (TN), false negative (FN) and false positive (FP). *TP, FN, FP, TN* represent the number of trials resulting in particular outcomes. Thus, according to [4] we can define

TP	outcome positive & test positive
TN	outcome negative & test negative
FP	outcome negative & test positive
FN	outcome positive & test negative

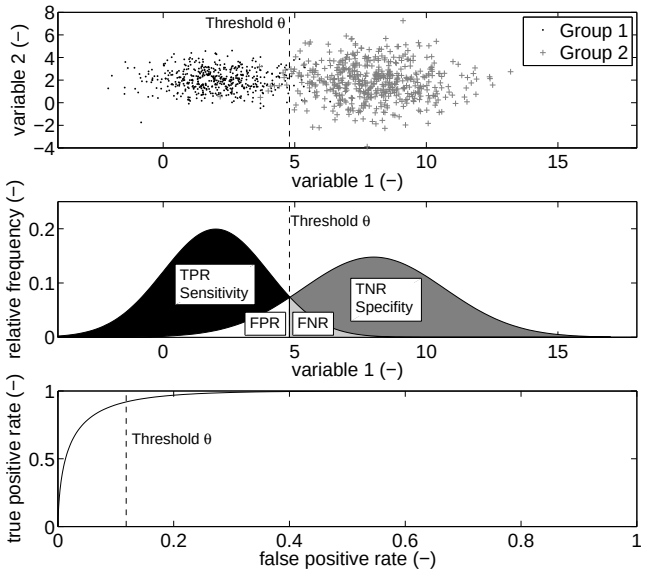

Fig. 1. Illustration of ROC curve construction.

It holds that $TP + FN + FP + TN = q$, where q is the total number of trials. Now, we can estimate Sensitivity as the true positive rate (TPR)

$$TPR = \text{Sensitivity} = \frac{TP}{TP + FN} = \frac{TP}{n^+} \tag{1}$$

and Specificity as the true negative rate (TNR)

$$TNR = \text{Specificity} = \frac{TN}{FP + TN} = \frac{TN}{n^-}. \tag{2}$$

The false positive rate $FPR = 1 - \text{Specificity} = 1 - TNR$. In the area of face recognition TPR is denoted as the Correct Classification Rate and FPR as the False Acceptance Rate.

The resultant ROC curve displays the relationship between Sensitivity (TPR on the y-axis) and 1-Specificity (FPR on the x-axis) for each value of the threshold θ (Fig. 1). As the threshold θ changes, the sizes of surfaces of Sensitivity and Specificity change too. In this way, the ROC performs all possible combinations of the relative frequencies of various kinds of correct and incorrect decisions.

4. Confidence Interval

We distinguish two approaches to the CIs construction. The first one, the parametric CIs, is based on an estimation of the proportion as the population parameter. Consequently, we calculate CI for this proportion. With respect to the assumptions and sample size we correct/adapt the calculation of boundaries. The second approach we suggest is the computation of a CB for the regression model which describes the ROC curve.

4.1 Parametric Confidence Intervals

The investigated normalized characteristics *TPR, TNR* are defined as proportions (i.e. relative frequencies of the algorithm decisions). We propose the use of CI for the relative frequency as presented in [16]. Denote m as the number of positive results (recognition successes) in the case of TPR, $(m = TP)$, the number of negative results in the case of TNR, $(m = TN)$. Let $n = n^+$ be the total number of trials (the number of face images) in the case of TPR, and let $n = n^-$ be the total number of trials in the case of TNR. The lower bound CI_L and the upper bounds CI_U of confidence interval for $X = TPR, TNR$ are defined as in [17]:
if $np > 5$ and $n(1 - p) > 5$

$$(CI_\text{L}, CI_\text{U}) = X \pm u_{1-\alpha/2}\sqrt{\frac{X(1-X)}{n}}, \tag{3}$$

else

$$CI_\text{L} = \frac{m}{(n - m + 1)F_{1-\alpha}(2(n - m + 1), 2m) + m},$$

$$CI_\text{U} = \frac{(m + 1)F_{1-\alpha}(2(m + 1), 2(n - m))}{n - m + (m + 1)F_{1-\alpha}(2(m + 1), 2(n - m))} \tag{4}$$

where $u_{1-\alpha/2}$ is a quantile of the normal distribution and $F_{1-\alpha}$ is a quantile of Fisher-Snedecors's distribution for the risk α. The CI, determined by the lower and upper bounds, can be estimated for each point of the ROC curve. This approach allows different levels of m, n, it is, therefore, more general.

According to [8] we can estimate the standard error (SEE) for a sufficiently large sample for Sensitivity and for Specificity as

$$SEE\,(TPR) = \sqrt{TPR(1 - TPR)/n^+}, \tag{5}$$

$$SEE\,(TNR) = \sqrt{TNR(1 - TNR)/n^-}. \tag{6}$$

Thus, the CI can be constructed as

$$(CI_\text{L}, CI_\text{U}) = TPR \pm t_{1-\alpha/2}(n^+ - 1)SEE\,(TPR), \tag{7}$$

$$(CI_\text{L}, CI_\text{U}) = TNR \pm t_{1-\alpha/2}(n^- - 1)SEE\,(TNR). \tag{8}$$

The univariate CI proposed by Brown and Davis works with two formulas [10]. Let both TPR and TNR be proportions having a binomial distribution. If the number of events is reasonably large, the binomial distribution can be approximated by the normal distribution. Thus, the CIs on estimation proportions X are similar to (3) adapted with the continuity correction term $1/2n$. It yields

$$(CI_\text{L}, CI_\text{U}) = X \pm \left[u_{\alpha/2}\sqrt{\frac{X(1-X)}{n}} + \frac{1}{2n}\right] \tag{9}$$

where $X = TPR$ is the true positive rate, $n = n^+$; and for $X = TNR$ the number of negative results, $n = n^-$; $u_{1-\alpha/2}$ is a quantile of the normal distribution; α is the risk. If $n \cdot X(1 - X) < 10$, the formula can be extremely erratic and inaccurate.

Brown and Davis [10] proposed another formula, the Wilsons interval, which has better statistical properties. Such a CI is reasonable for any $n \cdot X(1-X)$, n is the sample size of the corresponding proportion. Thus

$$(CI_{\mathrm{L}}, CI_{\mathrm{U}}) = \frac{n}{n+u_{\alpha/2}} \left[X + \frac{u_{\alpha/2}}{2n} \pm u_{1-\alpha} \sqrt{\frac{X(1-X)}{n} + \frac{u^2_{\alpha/2}}{4n^2}} \right]. \tag{10}$$

The corresponding formulas for the FPR is easily derived from the relation $FPR = 1 - TNR$.

4.2 Confidence Band around Regression Model

Let us consider the number of given ROC curve points constructed as pairs $\{(TPR_i, FPR_i)|i = 1, \ldots N\}$. The ROC curve points can be described by a regression model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon, \tag{11}$$

where $\mathbf{Y}$ is the dependent variable, $\mathbf{X} = (x_{ij})$ is the matrix of inputs, β is the vector of unknown parameters and ϵ is the error terms vector. An estimate of the unknown parameter β, denoted further as $\mathbf{b}$, can be found via the ordinary least square method [18].

Considering the shape of an expected ROC curve, we pre-selected the power and the polynomial regression model. Thus the regression model for the power function [18] is

$$TPR_i = \beta_0 \cdot FPR_i^{\beta_1} + \epsilon_i,\ i = 1, \ldots, N \tag{12}$$

where TPR_i is the observation of a dependent variable, FPR_i is the observation of an independent variable. To simplify the calculations from the non-linear to linear form, we use

$$\ln(TPR_i) = \ln(\beta_0) + \beta_1 \ln(FPR_i) + \ln(\epsilon_i), i = 1, \ldots, N. \tag{13}$$

The CB around the regression power model is given by the area where empirical values are dispersed around the regression function representing the theoretical values. Denote the lower bound CB_{L} and the upper bound CB_{U}. Then

$$(CB_{\mathrm{L}}, CB_{\mathrm{U}}) = \exp\left\{ \widehat{Y}_i \pm t_{1-\alpha/2}(N-k) \cdot \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}{N-k}} \right\} \tag{14}$$

where N is the sample size, k is the number of regression parameters including the absolute member and $\widehat{Y}_i = \ln(b_0) + b_1 \ln(FPR_i)$ is the regression model estimate.

The regression model for a polynomial function of order p is [18]

$$TPR_i = \beta_0 + \beta_1 \cdot FPR_i^p + \ldots + \beta_p \cdot FPR_i^p + \epsilon_i, \tag{15}$$

where TPR_i is the observation of a dependent variable, FPR_i is the observation of a independent variable. The corresponding CBs $(CB_{\mathrm{L}}, CB_{\mathrm{U}})$ around the polynomial regression model of order p is

$$(CB_{\mathrm{L}}, CB_{\mathrm{U}}) = \widehat{Y}_i \pm t_{1-\alpha/2}(N-k) \cdot \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \widehat{Y}_i)^2}{N-k}} \tag{16}$$

where N is the sample size, k is the number of regression parameters including the absolute member and $\widehat{Y}_i = b_0 + b_1 FPR_i + b_2 FPR_i^2 + \ldots + b_p FPR_i^p$ is the regression model estimate.

Let us denote the regression model estimate $\widehat{Y} = \widehat{TPR}$ and $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} TPR$. For measuring how well the regression function fits, we can use the index of determination

$$R^2 = \frac{ESS}{TSS} = \frac{= \sum_{i=1}^{N}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2} \tag{17}$$

where N is the number of ROC curve points. The closer R^2 is to 1, the better data description we obtain.

5. Application

This section presents the application of the proposed methods for the ROC confidence estimation on the performance comparison of two face recognition algorithms. It also briefly introduces face recognition algorithms and datasets used for testing.

5.1 Face Recognition Algorithms

Face recognition consists of face detection (i.e. the localization of a face in an image) and recognition. The method we used for face detection is the Viola-Jones face detector [19]. Face location and rotation was refined by an algorithm based on the detection of facial features - eyes, nose and mouth. The detected and aligned face was described for the purpose of recognition by Local Binary Pattern Histograms (LBPH) [20]. The face image represented by LBPH features was classified by the Nearest Neighbor (NN) classifier. This basic recognition framework is common for both of the algorithms being compared. the difference is the creation of reference models (also called templates) for the NN classifier. We used two approaches to model creation: Centroid method (Centroid) and the Higher Quantile Method (HQM) [21].

The Centroid constructs a face model by computing a centroid of a cluster formed by LBPH feature vectors extracted from training face images [21]. The centroid represents the cluster and forms the face model. The centroids are computed for each individual separately, thus a set of centroids is used as a reference for NN classifiers. HQM utilizes the statistical estimation of a cluster formed by training feature vectors. An Extreme Value Distribution is fitted to a histogram of feature values in each dimension. Then two quantiles are used as borders of the space that represents a face model.

Algorithms for face recognition differ in the face model creation process sometimes called a training or learning stage. The recognition performance of algorithm thus differs based on the method used: the Centroid or the HQM. We will show that the difference of the performance is up to 10% of TPR (Specificity). However, for the following steps, we need to evaluate if the difference of the performance is statistically significant or not.

5.2 Test Data

The face recognition algorithms were tested on the face image database called IFaViD [22]. We created IFaViD in order to examine and investigate the properties and performance of face recognition algorithms in surveillance camera systems. IFaViD consists of two subsets based on individuals actions performed in the view of a surveillance camera, we call them scenarios.

- Scenario A: a person walking through a door frame or a corridor (significant variability in face pose); 8731 images in total.
- Scenario B: a person requesting access at a closed door or a gateway access via an identification device (significant variability in illumination); 7711 images in total.

Scenario A images differ from Scenario B images in face dimensions, face pose, mutual face-camera position and face illumination. See [22], [21] for sample images. Each scenario based subset contains images of individuals that have created face models - internal persons and individuals without face models - external persons also called impostors. The IFaviD database is one of the larger databases. Despite this, we will show that the confidence of the results expressed by ROC curves is limited.

5.3 Short Demonstration of Selected Face Recognition Algorithms

Figure 2 shows ROC curves for different face recognition algorithms. We compare the Centroid Method, the Weighted Centroid Method and a method estimating centroids using the Gaussian Mixture Model (GMM) on both scenarios, the Baseline stands for the Most Similar Face method (for more details see [21]). As it was proved in paper [21] the Centroid Method performs best. Therefore, in the next part we will demonstrate the comparison of HQM method with the Centroid using CI assessment.

5.4 Confidence Interval Assessment

In this part, we calculate the CI of ROC curves representing the performance of algorithms described above. We use the following notation for the calculated CIs. The parametric CI calculated according to Stehlikova [17] and equations (3), (4) are denoted as „Stehlikova CI“; according to Westin [8] and equation (7) is denoted as „Westin CI“ and according to Brown and Davis [10] and equation (10) is denoted as „Brown CI“. We also calculate the CB around the regression model of ROC curve according to (14) or (16) and denote it „RM CB“.

With respect to the expected shape of ROC curve we pre-selected the polynomial and the power regression model. The final selection of the model was done with respect to the shape of the curve and the level of $R^2 > 0.99$ see (17). Therefore, we use the power model for both methods (the Centroid and HQM) in Scenario A. We use the polynomial model ($p = 6$) for the Centroid in Scenario B; we use the polynomial model ($p = 5$) for HQM in Scenario B. The summary is shown in Tab. 1.

Let us focus on the comparison of the performance of face recognition algorithms via CIs. We will assess an overlap of CIs for both methods. Generally, if the CIs for HQM and Centroid do not overlap, the method with a higher TPR (for the given FPR) performs better. If the CIs overlap, we identify the regions of ROC points with overlap and assess the extent of overlap. We keep in mind that a wider CI gives a higher uncertainty for the estimated proportion. This is considered as an advantage in the following sense. If wider CIs do not overlap, there is a strong possibility that one method provides better results. The estimates of CI for Scenario A are given in Fig. 3, for Scenario B in Fig. 4.

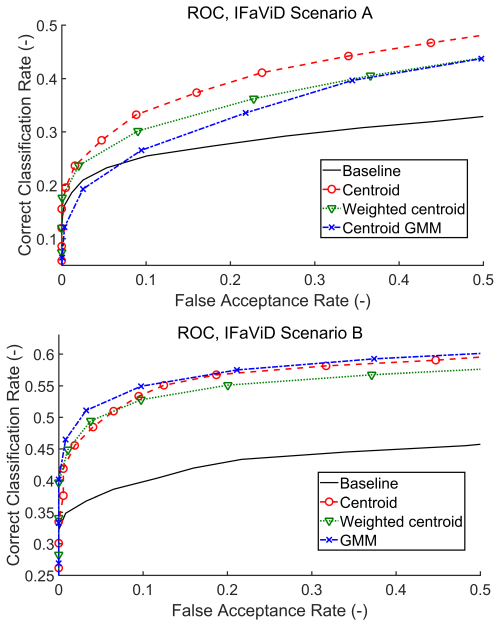

Fig. 2. ROC curves describing face recognition system performance in two scenarios from the IFaViD database. (Baseline: solid line; Centroid: dashed line; Weighted centroid: dotted line; GMM: dashed dotted line). The y-axis is displayed in the range $\langle 0.1, 0.5\rangle$ for Scenario A, and in $\langle 0.25, 0.6\rangle$ for Scenario B.

Scenarios A		
Regression model	Centroid ($N = 43$)	HQM ($N = 29$)
Polynomial, $p = 6$	0.9992	0.9890
Polynomial, $p = 5$	0.9987	0.9839
Power model	0.9969	0.9966
Scenarios B		
Regression model	Centroid ($N = 43$)	HQM ($N = 32$)
Polynomial, $p = 6$	0.9914	0.9988
Polynomial, $p = 5$	0.9885	0.9981
Power model	0.9828	0.9663

Tab. 1. A regression model and its R^2 (N is number of ROC curve points).

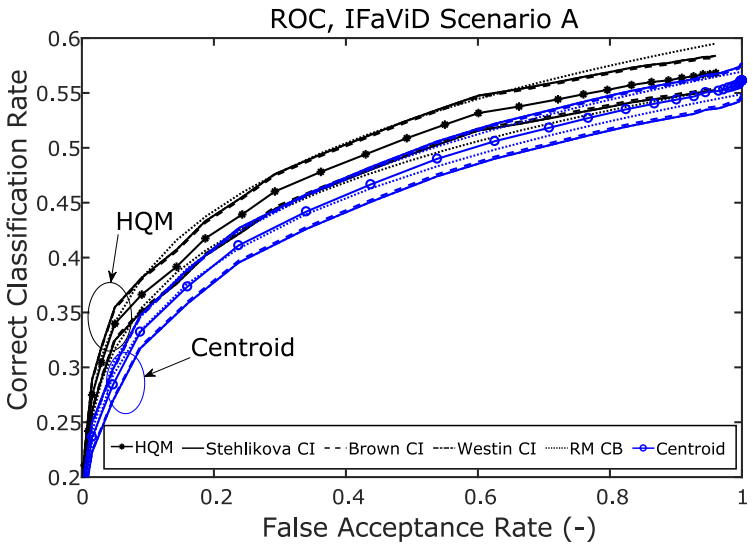

Fig. 3. Confidence intervals for scenario A (Stehlikova CI: solid line; Brown CI: dashed line; Westin CI: dashed dotted line; RM CB: dotted line). The y-axis is displayed in the range $\langle 0.2, 0.6\rangle$ to enable better reading. Westin CI and Stehlikova CI show similar results, therefore we cannot see the difference in the figure.

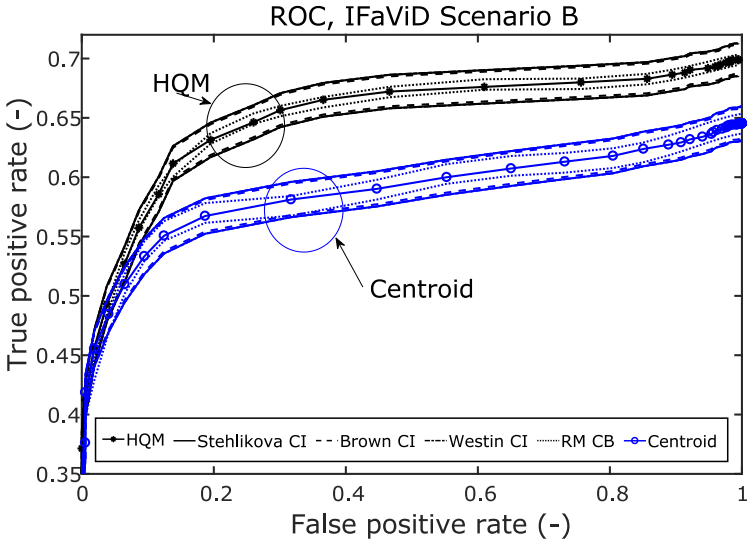

Fig. 4. Confidence intervals for scenario B (Stehlikova CI: solid line; Brown CI: dashed line; Westin CI: dashed dotted line; RM CB: dotted line). The y-axis is displayed in the range $\langle 0.35, 0.73\rangle$ to enable better reading. Westin CI and Stehlikova CI show similar results, therefore we cannot see the difference in the figure.

Let us focus on Scenario A (Fig. 3). We can see that the HQM method performs better than the Centroid. The CI of the respective ROC curves are very tight and they overlap in some areas. If we make an in-depth analysis of CI for both methods, we can find three basic areas - not overlapping CIs, tight CIs and overlapping CIs. In the range of FPR $\langle 0, 0.1\rangle$, no CI or band overlaps. In the range $\langle 0.1, 0.6\rangle$, the CIs are very tight, and in the range $\langle 0.6, 1\rangle$, they overlap. Hence, we conclude that HQM can provide better results than the Centroid for low values of FPR; the improvement is not so significant for higher FPR values.

In the case of Scenario B (Fig. 4), we can see that the HQM method outperforms the Centroid. In the initial range $\langle 0, 0.1\rangle$ of the FPR, both ROC curves are close to each other. Both CIs overlap, so in this range the difference of the performance between methods is considered insignificant. Furthermore,in the range of FPR $\langle 0.1, 1\rangle$, no intervals or bands overlap. Therefore, with respect to the 90% of non overlapping CIs, we conclude that the HQM method performs better than the Centroid.

Focusing on the results of the CIs calculated according to [17], [8] and [10] for both scenarios (Fig. 3, 4) we do not observe any significant differences. The proximity of the results is caused by the large sample size ($n > 5000$) for estimating ROC points. Therefore, in such a case, we recommend calculating the CI according to Westin ([8] and (7)) due to its simplicity. Next we investigate the influence of sample size on the estimated confidence of ROC curve.

5.5 Sample Size Investigation

The influence of the sample size on the CI width is natural. Therefore, considering the calculation of CI which can be generally written as estimated statistic ± permissible error, we can also investigate the level of permissible error with respect to the sample size. Firstly, we investigate the influence of sample size reduction on the CI width. The reduction is done by establishing the percentage of the complete data-sets. The percentage was established on the 100, 50, 20, 10, 8, 1 % setting value which corresponds to the sample size n = 6549, 3275, 1310, 655, 524, 65. The illustration of the Brown CI for HQM is presented in Fig. 5. The graphs of the other CIs, i.e. for HQM according to Stehlikova and Westin, generally look the same and give the same conclusion. That is, the smaller the sample size, the wider the CI. Therefore, we do not present them.

Secondly, we investigate the influence of sample size on the permissible error of CI for a selected ROC point. Figure. 6 provides the illustration for Stehlikova, Westin and Brown CIs for the range of sample size 20–6549. The calculation follows two basic steps. First, we take the value of TPR for which we then calculate the permissible error of CI for a changing sample size. then we start with a sample size value n = 20, we establish a step of 10 samples and end with the sample size n = 6549. This calculation yields one of the curves in Fig. 6.

The curves in Fig. 6 are calculated for the pre-selected TPR = 50, 80, 95. We show the results of the sample size in the range 20 to 300, because from a sample size $n > 300$ all curves have the same tendency and converge slowly to the horizontal axis. The bottom group of curves corresponds to the ROC point for which TPR = 0.95, the middle group of curves corresponds to TPR = 80 and the upper group of curves corresponds to the ROC point for which TPR = 50. Hence, we simulate the influence of sample size for several points of ROC curve according to TPR.

Based on Fig. 6 we can formulate several conclusions: i) the biggest differences among the CI estimation methods are for the sample size in (0, 150); ii) for a growing sample size $n > 200$, the CIs converge to each other and to the horizontal axis (i.e. the permissible error decreases); iii) the higher the position of ROC (the higher TPR) for a non-zero value of FPR, the lower the permissible error; iv) for every CI estimation method, a higher TPR yields a lower permissible error in the whole range of the sample size.

As for comparing ROC curves, based on the first conclusion in the previous paragraph denoted as i), we recommend using Stehlikova and Brown CIs in the case of a small sample size ranging (0, 150). The estimation of CB around the regression model is not considered because the permissible error of such a CB depends on the number of ROC points and on the estimated regression model. The permissible error does not depend on the sample size used for estimating of each ROC curve. We recommend using the regression modeling of ROC points when the calculation of AUC is investigated. This, however, falls out of the scope of this paper.

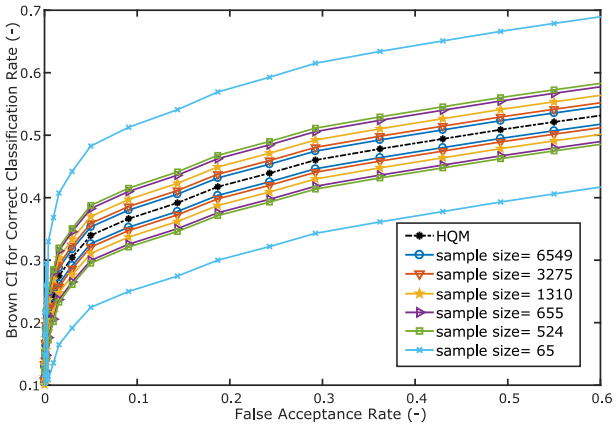

Fig. 5. Brown CI for different sample size. The sample size reduction causes the reduction of the CI's width. The order of the legend corresponds with the reduction of the CIs.

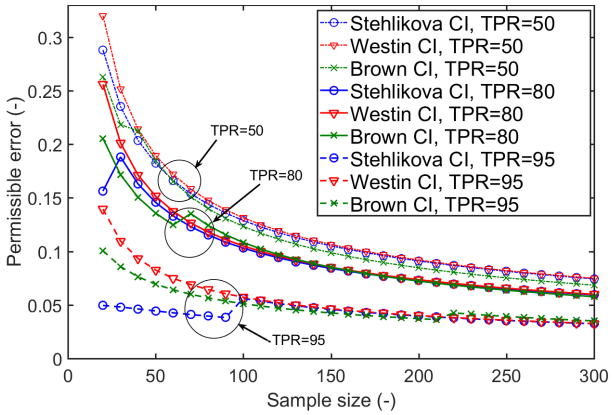

Fig. 6. Dependence of permissible error on sample size.

6. Conclusion

The aim of the paper was to propose a methodology for comparing decision making processes on the basis of CIs for ROC curves and to investigate how a sample size influences the applicability of this methodology. We studied CIs for ROC curves and the overlap of CIs for the compared ROC curves. We kept in mind that a wider CI gives a higher uncertainty for the estimated ROC point. If a wider CI does not overlap, we can claim that one method provides better results with a stronger probability. The application was demonstrated on face recognition in order to identify a method with a higher recognition performance.

The sample size investigation and the application on real data revealed the following conclusions and recommendations. In cases of large samples ($n > 1000$) we recommend comparing ROC curves via the calculation of CI according to Westin because of its simplicity and the proximity of all discussed intervals. For a growing sample size $n \in (200, 1000)$, all parametric CIs provide similar confidence and permissible errors. Therefore, there is no preferred approach. In the case of small samples $n \in (0, 150)$, we recommend using both Stehlikova and Brown approaches for comparing ROC curves, i.e. of decision making algorithms.

We recommend calculating of CBs, via regression modeling of the ROC curve trend, if the aim of an analysis is i) comparing each of ROC point separately via CIs, and/or ii) calculating of AUC. In such a case the CI depends on the number of ROC points and on the quality of the fitted model.

Focusing on real data, we can state the following. In the case of Scenario A, HQM provided better results than the Centroid for a lower level of FPR; for a higher level of the FPR, the improvement was not so significant. In Scenario B, with respect to 90% of non overlapping CIs we claim that HQM provided a better face template than the Centroid.

Acknowledgments

The research described in this paper was financed by the Czech Ministry of Education in the frame of the National Sustainability Program under Grant LO1401. For research, the infrastructure of the SIX Center was used.

References

[1] SCHAFER, H. Efficient confidence bounds for ROC curves. *Statistics in Medicine*, 1994, vol. 13, no. 15, p. 551–1561. DOI: 10.1002/sim.4780131506

[2] LOPEZ-RATON, M., RODRIGUEZ-ALVAREZ, A. X., CADARSO-SUAREZ, C., GUEDA-SAMPEDRO, F. OptimalCutpoints: An R package for selecting optimal cutpoints in diagnostic tests. *Journal of Statistical Software*, 2014, vol. 61, no. 8, p. 1–36. DOI: 10.18637/jss.v061.i08

[3] VENKATAKRISHNAN, P., SANGEETHA, S. Singularity detection in human EEG signal using wavelet leaders. *Biomedical Signal Processing and Control*, 2014, vol. 13, no. 1, p. 282–294. DOI: 10.1016/j.bspc.2014.06.002

[4] XIA, J., BROADHURST, D., I. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*, 2013, vol. 9, no. 2, p. 280–299. DOI: 10.1007/s11306-012-0482-9

[5] VERGARA, L., SORIANO, A., SAFON, G., SALAZAR, A. On the fusion of non-independent detectors. *Digital Signal Processing*, 2016, vol. 50, p. 24–33. DOI: 10.1016/j.dsp.2015.11.009

[6] BOASHAS, B., AZEMI, G., ALI KHAN, N. Principles of time-frequency feature extraction for change detection in non-stationary signals: Applications to newborn EEG abnormality detection. *Pattern Recognition*, 2015, vol. 48, no. 43, p. 616–627. DOI: 10.1016/j.patcog.2014.08.016

[7] HANLEY, J. A., MCNEIL, B. J. A method of comparing the areas under receiver operating characteristic curve derived from the same cases. *Radiology*, 1983, vol. 148, no. 3, p. 839–843. DOI: 10.1148/radiology.148.3.6878708

[8] WESTIN, L. K. *Receiver Operating Characteristic (ROC) Analysis. Evaluating Discriminance Effects among Decision Support Systems*. 28 pages. [Online] Cited 2015-02-11. Available at: http://nutkin.cs.umu.se/research/reports/2001/018/part1.pdf

[9] BUI, D. T., TUAN, T. A., KLEMPE, H., et al. Spatial prediction models for shallow landslide hazards: A comparative assessment of the efficacy of support vector machines, artificial neural networks, kernel logistic regression, and logistic model tree. *Landslides*, 2016, vol. 13, no. 2, p. 361–378. DOI: 10.1007/s10346-015-0557-6

[10] BROWN, C. D., DAVIS, H. T. Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 2006, vol. 80, no. 1, p. 24–38. DOI: 10.1016/j.chemolab.2005.05.004

[11] ZOU, G. Y., YUE, L. Using confidence intervals to compare several correlated areas under the receiver operating characteristic curves. *Statistics in Medicine*, 2013, vol. 32, no. 29, p. 5077–5090. DOI: 10.1002/sim.5889

[12] SORIANO, A., VERGARA, L., BOUYIANE, A., SALAZAR, A. Fusion of scores in a detection context based on alpha integration. *Neural Computation*, 2015, vol. 27, no. 9, p. 1983–2010. DOI: 10.1162/NECO_a_00766

[13] HALL, P., HYNDMAN, R. J., FAN, Y. Nonparametric confidence intervals for receiver operating characteristic curves. *Biometrika*, 2004, vol. 91, no. 3, p. 743–750. DOI: 10.1093/biomet/91.3.743

[14] ZOU, K. H., HALL, W. J., SHAPIRO, D. E. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 1997, vol. 16, no. 19, p. 2143–2156. DOI: 10.1002/(SICI)1097-0258(19971015)16:19<2143::AID-SIM655>3.0.CO;2-3

[15] FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, vol. 27, no. 8, p. 861–874. DOI: 10.1016/j.patrec.2005.10.010

[16] MALACH, T., POMENKOVA, J. Confidence assessment of face recognition results. In *Proceedings of the 25th International Conference Radioelektronika*. Pardubice (Czech Republic), 2015, p. 176–179. DOI: 10.1109/RADIOELEK.2015.7129002

[17] STEHLIKOVA, B., TIRPAKOVA, A., POMENKOVA, J., et al. *Research Methodology and Statistical Inference.* 1st ed. Brno (Czech Republic): Mendel University, 2009. ISBN 978-80-7375-362-7

[18] GREEN, W. H. *Econometric Analysis.* USA: Prentice Hall, 2012. ISBN 978-01-3139-538-1

[19] VIOLA, P., JONES, M. Robust real-time object detection. *International Journal of Computer Vision*, 2001, vol. 57, no. 2, p. 1–25.

[20] AHONEN, A., HADID, A., PETIKAINEN, M. Face description with local binary patterns: application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, vol. 28, no 12, p. 2037–2041. DOI: 10.1109/TPAMI.2006.244

[21] MALACH, T., POMENKOVA, J. Face template creation: Is centroid method a suitable approach? In *Proceedings of the 24th International Conference Radioelektronika*. Bratislava (Slovakia), 2014, p. 724–729. DOI: 10.1109/Radioelek.2014.6828428

[22] BAMBUCH, P., MALACH, T., MALACH, J. Video database for face recognition. In *Proceeding of Technical Computing*. Bratislava (Slovakia), 2012, p. 1–7.

About Authors...

Tobiáš MALACH graduated from the Faculty of Electrical Engineering and Communication, Brno University of Technology in 2013. Currently he is a Ph.D. student at the DREL, Brno University of Technology. He has been devoted to computer vision application to surveillance and security systems since 2011.

Jitka POMĚNKOVÁ received a Ph.D. degree in Applied Mathematics at Ostrava University in 2005, a habilitation degree in Econometric and Operational Research at Mendelu in Brno in 2010. From 2011 she is the Senior researcher at the DREL, Brno University of Technology.