Selection Strategies for pAUC-Based Combination of Dichotomizers

Maria Teresa Ricamato, Mario Molinara, and Francesco Tortorella

DAEIMI - Università degli Studi di Cassino
via G. Di Biasio 43, 03043 Cassino, Italy
{mt.ricamato,m.molinara,tortorella}@unicas.it

Abstract. In recent years, classifier combination has been of great interest for the pattern recognition community as a method to improve classification performance. Several combination rules have been proposed based on maximizing the accuracy and the Area under the ROC curve (AUC). Taking into account that there are several applications which focus only on a part of the ROC curve, i.e. the one most relevant for the problem, we recently proposed a new algorithm aimed at finding the linear combination of dichotomizers which maximizes only the interesting part of the AUC. Since the algorithm uses a greedy approach, in this paper we define and evaluate some possible strategies which select the dichotomizers to combine at each step of the greedy approach. An experimental comparison is drawn on a multibiometric database.

Keywords: Classifiers combination, ROC curve, partial AUC.

1 Introduction

Classifier combination has become an established technique for building proficient classification systems. Among the various combination methods proposed up to now, linear classifier combination has been used mainly for its simplicity and effectiveness. In particular, some methods have been designed to increase the Area under the ROC curve (AUC), a more suitable performance measure than the classification accuracy [1], specially for those applications characterized by imprecise environment or imbalanced class priors [2]. AUC resumes in a single quantitative index the performance exhibited by the classifier over all the false positive rate (FPR) values.

However, there are many applications that are interested only to a particular range of FPRs. For example, in a biometric authentication system used to identify people, or to verify the claimed identity of registered users when entering in a protected area, a false positive is considered the most serious error, since it gives unauthorized users access to the systems that expressly are trying to keep them out. In such case, the FPR values considered are the ones that correspond to lower values, and the partial AUC [3] is the most indicate index to use, since it allows us to focus on particular regions of the ROC space. In [4] we have proposed a new method aimed at calculating the weight vector in a

C. Sansone, J. Kittler, and F. Roli (Eds.): MCS 2011, LNCS 6713, pp. 177–186, 2011.

linear combination of $K \geq 2$ dichotomizers, such that the pAUC is maximized. In particular, we have provided an algorithm for finding the optimal weight in a combination of two dichotomizers and then have extended to the combination of $K > 2$ dichotomizers by means of a greedy approach which divides the whole K-combination problem into a series of pairwise combination problems.

In such a case, making the right local choice at each stage is of fundamental importance since it affects the performance of the whole algorithm. For this reason, in this paper we define and evaluate some possible strategies which select the dichotomizers to combine at each step of the greedy approach. The strategies considered are based both on the evaluation of the best single dichotomizer and of the best pair of dichotomizers. Such strategies are then experimentally compared on a biometric database.

The paper is organized as follow. The next section presents the pAUC index and its main properties while section 3 analyzes the combination of two dichotomizers. The combination of $K > 2$ dichotomizers is presented in section 4; in the same section the proposed selection strategies are described and analyzed. The performed experiments and obtained results are shown in section 5, while section 6 concludes the paper.

2 ROC Analysis and Partial Area Under the ROC Curve

Receiver Operating Characteristics (ROC) graphs are useful for visualizing, organizing and selecting classifiers based on their performance. Given a two-class classification model, the ROC curve describes the trade-off between the fraction of correctly classified actually-positive cases (True Positive Rate, TPR) and the fraction of wrongly classified actually-negative cases (False Positive Rate, FPR), giving a description of the performance of the decision rule at different operating points.

In some cases, it is preferable to use the Area under the ROC Curve (AUC) [5] [6], a single metric able to summarize the performance of the classifiers system:

$$AUC = \int_0^1 ROC(\tau)d\tau \tag{1}$$

As said before, some applications do not use all the range of false positive rates: in particular, the most part of biometric and medical applications [7] work on false positive rate close to the zero value. In such cases it is worth to consider a different summary index measuring the area under the part of the ROC curve with FPRs between 0 and a maximal acceptable value t. This index is called *partial AUC* (pAUC) and defined as:

$$pAUC = \int_0^t ROC(\tau)d\tau \tag{2}$$

where the interval $(0, t)$ denotes the false positive rates of interest. Its choice depends on the particular application, and it is related to the involved cost of a false positive diagnosis.

Moreover, the pAUC can be also defined as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one, such that this latter is higher than the $1-t$ quantile[1] q_y^t:

$$pAUC = P\left\{x_i > y_j, y_j > q_y^t\right\} \tag{3}$$

where $x_i = f(\mathbf{p}_i)$ and $y_j = f(\mathbf{n}_j)$ are the outcomes of the dichotomizer f on a positive sample $\mathbf{p}_i \in P$ and a negative sample $\mathbf{n}_j \in N$.

In order to evaluate the pAUC of a dichotomizer avoiding to perform a numerical integration on the ROC curve, we use the non-parametric estimator [3], which is defined as:

$$pAUC = \frac{1}{m_P \cdot m_N} \sum_i^{m_P} \sum_j^{m_N} V_{ij}^{q_y^t} \tag{4}$$

where m_P and m_N are the cardinalities of the positive and negative subsets, respectively, and

$$V_{ij}^{q_y^t} = I\{x_i > y_j, y_j > q_y^t\} = \begin{cases} 1 & \text{if } x_i > y_j \bigwedge y_j > q_y^t; \\ 0.5 & \text{if } x_i = y_j \bigwedge y_j > q_y^t; \\ 0 & \text{if } x_i < y_j \bigwedge y_j > q_y^t. \end{cases} \tag{5}$$

3 Combination of Two Dichotomizers

As a first step, let us consider a set $T = P \cup N$ of samples, and define the outputs of two generic dichotomizers f_h and f_k on two positive and negative samples $\mathbf{p}_i$ and $\mathbf{n}_j$:

$$x_i^h = f_h(\mathbf{p}_i), \qquad x_i^k = f_k(\mathbf{p}_i), \qquad y_j^h = f_h(\mathbf{n}_j), \qquad y_j^k = f_k(\mathbf{n}_j).$$

The pAUCs for the two dichotomizers, considering the FPR interval $(0, t)$, are:

$$pAUC_h = \frac{\sum_{i=1}^{m_P} \sum_{j=1}^{m_N} I\left(x_i^h > y_j^h, y_j^h > q_{y^h}^t\right)}{m_P \cdot m_N}, \quad pAUC_k = \frac{\sum_{i=1}^{m_P} \sum_{j=1}^{m_N} I\left(x_i^k > y_j^k, y_j^k > q_{y^k}^t\right)}{m_P \cdot m_N} \tag{6}$$

It is worth noting that the linear combination of two generic dichotomizers $f_{lc} = \alpha_h f_h + \alpha_k f_k$ can be put as $f_{lc} = f_h + \alpha f_k$ without loss of generalization, with $\alpha = \frac{\alpha_k}{\alpha_h} \in (-\infty, +\infty)$. Therefore, considering the linear combination, the outcomes on $\mathbf{p}_i$ and $\mathbf{n}_j$ are:

$$\xi_i = f_{\text{lc}}(\mathbf{p}_i) = x_i^h + \alpha x_i^k, \qquad \eta_j = f_{\text{lc}}(\mathbf{n}_j) = y_j^h + \alpha y_j^k. \tag{7}$$

and the pAUC is:

$$pAUC_{\text{lc}} = \frac{1}{m_P \cdot m_N} \left(\sum_{i=1}^{m_P} \sum_{j=1}^{m_N} I\left(\xi_i > \eta_j, \left(\eta_j > q_\eta^t\left(\alpha\right)\right)\right) \right) \tag{8}$$

[1] The quantile function returns the value below which random draws from the negative population would fall, $(1-t) \times 100$ percent of the time.

To have an insight into the formulation of the $pAUC_{\mathrm{lc}}$, let us analyze the term $I(\xi_i > \eta_j)$ without considering the constraint on the quantile. In particular, let us consider how it depends on the values of $I(x_i^h, y_j^h)$ and $I(x_i^k, y_j^k)$:

- $I(x_i^h, y_j^h) = 1$ and $I(x_i^k, y_j^k) = 1$. In this case the two samples are correctly ranked by the two dichotomizers, and $I(\xi_i > \eta_j) = 1$.
- $I(x_i^h, y_j^h) = 0$ and $I(x_i^k, y_j^k) = 0$. In this case neither dichotomizer ranks correctly the samples and thus the contribution for the $pAUC$ is 0.
- $I(x_i^h, y_j^h) \operatorname{xor} I(x_i^k, y_j^k) = 1$. Only one dichotomizer ranks correctly the samples while the other one is wrong. In this case the value of $I(\xi_i > \eta_j)$ depends on the weight α.

The subset T can be divided into four subsets: T_{hk}, $T_{h\bar{k}}$, $T_{\bar{h}k}$ and $T_{\bar{h}\bar{k}}$ defined as:

$$\begin{aligned}
T_{hk} &= \{(\mathbf{p}_i, \mathbf{n}_j) | I(x_i^h, y_j^h) = 1 \text{ and } I(x_i^k, y_j^k) = 1\},\\
T_{\bar{h}k} &= \{(\mathbf{p}_i, \mathbf{n}_j) | I(x_i^h, y_j^h) = 0 \text{ and } I(x_i^k, y_j^k) = 1\},\\
T_{h\bar{k}} &= \{(\mathbf{p}_i, \mathbf{n}_j) | I(x_i^h, y_j^h) = 1 \text{ and } I(x_i^k, y_j^k) = 0\},\\
T_{\bar{h}\bar{k}} &= \{(\mathbf{p}_i, \mathbf{n}_j) | I(x_i^h, y_j^h) = 0 \text{ and } I(x_i^k, y_j^k) = 0\}
\end{aligned}$$

Now, let us consider the constraint on the negative samples related to the quantile, and define the following set:

$$\Gamma_\alpha = \{(\mathbf{p}_i, \mathbf{n}_j) \in P \times N | y_j^h + \alpha y_j^k > q_\eta^t\} \tag{9}$$

where q_η^t is the $1 - t$ quantile of η, which depends on the weight α. If we define the sets $T'_{hk}, T'_{\bar{h}k}, T'_{h\bar{k}}, T'_{\bar{h}\bar{k}}$ as:

$$\begin{aligned}
T'_{hk} &= T_{hk} \cap \Gamma_\alpha, \quad T'_{\bar{h}k} = T_{\bar{h}k} \cap \Gamma_\alpha,\\
T'_{h\bar{k}} &= T_{h\bar{k}} \cap \Gamma_\alpha, \quad T'_{\bar{h}\bar{k}} = T_{\bar{h}\bar{k}} \cap \Gamma_\alpha,
\end{aligned}$$

the expression for $pAUC_{\mathrm{lc}}$ in equation 8 can be written as:

$$pAUC_{\mathrm{lc}} = \frac{1}{m_P \cdot m_N} \left(\sum_{(\mathbf{p}_i, \mathbf{n}_j) \in T'_{\bar{h}\bar{k}}} I(\xi_i > \eta_j) + \sum_{(\mathbf{p}_i, \mathbf{n}_j) \in T'_{hk}} I(\xi_i > \eta_j) \quad + \sum_{(\mathbf{p}_i, \mathbf{n}_j) \in T'_{h\bar{k}} \cup T'_{\bar{h}k}} I(\xi_i > \eta_j) \right).$$

Starting from equation above, the value of α which maximizes $pAUC_{\mathrm{lc}}$ can be found by means of a linear search; the details are described in [4].

4 Combination of $K > 2$ Dichotomizers

Let us now consider the linear combination of $K > 2$ dichotomizers which is defined as:

$$f_{lc}(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + ... + \alpha_K f_K(x) = \sum_{i=1}^{K} \alpha_i f_i(x) \tag{10}$$

In order to find the weight vector $\alpha_{opt} = (\alpha_1, ..., \alpha_K)$ that maximizes the pAUC associated to $f_{lc}(x)$, we can consider a greedy approach which divides the whole K-combination problem into a series of pairwise combination problems. Even though suboptimal, such approach provides a computationally feasible algorithm for the problem of combining of K dichotomizers which would be intractable if tackled directly. In particular, the first step of the algorithm chooses the two "most promising" dichotomizers and combines them so that the number of dichotomizers decreases from K to $K-1$. This procedure is repeated until all the dichotomizers have been combined.

The choice of the dichotomizers that should be combined in each iteration plays an important role since it affects the performance of the algorithm. To this aim, we have considered various selection strategies that differ in the way the greedy approach is accomplished: in the *Single Classifier Selection* the best candidate dichotomizer is chosen, while with the *Pair Selection* the best candidate pair of dichotomizers is taken.

4.1 Single Classifier Selection

This is the most immediate approach, that selects the dichotomizer with the best performance index. The related implementation first sorts the K dichotomizers into decreasing order of an individual performance measure, and the first two classifiers are combined. The remaining dichotomizers are then singularly added to the group of the combined dichotomizers.

pAUC based selection. A first way to implement this strategy is to use the pAUC of the dichotomizers as single performance measure. In this way, one looks at the behavior of the single classifier in the range of interest of the false positives assuming that it could be a sufficiently good estimate of how the classifier contributes to the combination. In other words, we are assuming that it is sufficient to take into account the performance of the dichotomizer, say f_h, on the set of the negative samples $N_h = \{\mathbf{n}_j \in N | y_j^h > q_h^t\}$ even though this set likely does not coincide with the set of negative samples contributing to the value of $pAUC_{lc}$.

AUC based selection. A second way is to employ the whole AUC of the dichotomizer. In this case we don't assume that the samples in N_h are sufficient to predict the performance of the combination and thus consider the behavior of the dichotomizer in the whole FPR range.

4.2 Pair Selection

This approach is based on the estimation of the joint characteristics of a pair of dichotomizers, so as to predict how proficient is their combination. In particular, we rely on the idea that combining dichotomizers with different characteristics should provide good performance. Therefore, a pairwise measure is evaluated which estimates the diversity of the dichotomizers; at each step, the pair of classifiers exhibiting the best index is combined. The procedure is repeated $K-1$ times until a single classifier is obtained. Two different measures are considered which estimate the diversity in the ranking capabilities between two dichotomizers.

Kendall Rank Coefficient. The first index we consider is the *Kendall rank correlation coefficient* [8] that evaluates the degree of agreement between two sets of ranks with respect to the relative ordering of all possible pairs of objects. Given K the sum of concordant pairs and l the number of considered items, the Kendall rank correlation coefficient is defined as:

$$\frac{2K}{\frac{1}{2}l(l-1)} - 1 \tag{11}$$

where $\frac{1}{2}l(l-1) = \binom{l}{2}$ is the total amount of pairs.

In our case the subsets are the ones defined in the previous section and thus the correlation coefficient can be redefined as:

$$\tau' = \frac{2\left(|T'_{hk}| + |T'_{\bar{h}\bar{k}}|\right)}{t \cdot m_P \cdot m_N} - 1 \tag{12}$$

However, to evaluate τ' we should previously know the optimal value of the coefficient α chosen for the combination, but this would be computationally heavy. Therefore we use a surrogate index τ defined as

$$\tau = \frac{2\left(|T_{hk}| + |T_{\bar{h}\bar{k}}|\right)}{m_P \cdot m_N} - 1 \tag{13}$$

which is an upper bound for τ' since $|T_{hk} \geq |T'_{hk}|$ and $|T_{\bar{h}\bar{k}}| \geq |T'_{\bar{h}\bar{k}}|$. At each step the pair with the lowest τ is chosen.

Ranking Double Fault. The second index comes from an analysis of the expression of the $pAUC_{lc}$ given in eq. 10. The maximum allowable pAUC of a linear combination depends on the cardinality of the subsets T'_{hk}, $T'_{h\bar{k}}$ and $T'_{\bar{h}k}$. Since the number of pairs in each subset depends on the value of the quantile, it is not possible to compute a priori the value of $pAUC_{lc}^{max}$. However, it is feasible to obtain a lower bound for it. To this aim, let us consider the relation between the number of pairs of positive and negative samples obtained without using the quantile, and its reduction after using the quantile. The $(1-t)$ quantile is the value which divides a set of samples such that there is the given proportion $(1-t)$ of observations below it. Therefore, when the quantile is applied, the number of the considered negative samples decreases, with the consequent change of the number of the total pairs:

$$\begin{aligned} m'_{tot} &= m_{tot} - |\{(\mathbf{p}_i, \mathbf{n}_j) | \eta_j < q_\eta^t(\alpha)\}| \\ &= m_P \cdot m_N - m_P[(1-t)m_N] = t \cdot m_P \cdot m_N = t \cdot m_{tot} \end{aligned} \tag{14}$$

where $m_{tot} = m_P \cdot m_N$ is the number of pairs considered without the constraint of the quantile.

Therefore, the $pAUC_{lc}^{max}$ can be rewritten as:

$$pAUC_{lc}^{max} = \frac{1}{m_P \cdot m_N}\left(m'_{tot} - |T'_{\bar{h}\bar{k}}|\right) = \frac{1}{m_P \cdot m_N}\left(t \cdot m_P \cdot m_N - |T'_{\bar{h}\bar{k}}|\right) \quad (15)$$

It is obvious that :

$$|T'_{\bar{h}\bar{k}}| \leq |T_{\bar{h}\bar{k}}| \quad (16)$$

since the number of the considered negative samples decreases.

Therefore:

$$pAUC_{lc}^{max} \geq \frac{1}{m_P \cdot m_N}\left(t \cdot m_P \cdot m_N - |T_{\bar{h}\bar{k}}|\right) \quad (17)$$

In particular, the lower bound for $pAUC_{lc}^{max}$ is high when we have a low number of pairs that have been misranked by both the classifiers. This quantity can be interestingly related to the *double fault* measure [9] that is used to evaluate the diversity between classifiers. For this reason we define *Ranking Double Fault* the index:

$$DF = \frac{|T_{\bar{h}\bar{k}}|}{m_P \cdot m_N} \quad (18)$$

and adopt it as second diversity index. Also in this case, the pair with the lowest DF is chosen at each step.

5 Experimental Results

In order to compare the selection strategies proposed in the previous section, some experiments have been performed on the public-domain biometric dataset XM2VTS [10], characterized by 8 matchers. We used the partition of the scores into training and test set proposed in [10] and shown in table 1. The XM2VTS is a multimodal database containing video sequences and speech data of 295 subjects recorded in four sessions in a period of 1 month. In order to assess its performance the Lausanne protocol has been used to randomly divide all the subjects into positive and negative classes: 200 positive, 25 evaluation negatives and 70 test negatives. All the details about the procedure used to obtain the final dichotomizers are described in [10].

For each considered strategy the vector of coefficients for the linear combination is evaluated on the validation set, and then applied to the test set. The results are analyzed in term of partial AUC, considering the false positive ranges: $FPR_{0.1} = (0, 0.1)$, $FPR_{0.05} = (0, 0.05)$ and $FPR_{0.01} = (0, 0.01)$. For the sake

Table 1. XM2VTS database properties

	# Sample	# Positive	# Negative
Validation Set	40600	600	40000
Test Set	112200	400	111800

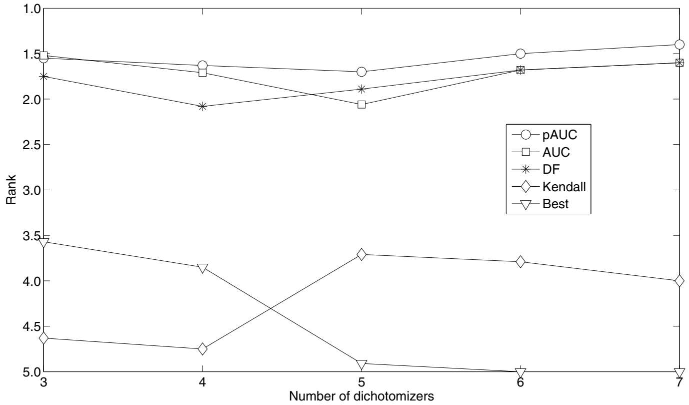

Fig. 1. Mean rank of the selection strategies for t=0.1

of comparison, we consider, besides the combinations obtained with the four selection strategies described in the previous section, also the best single classifier chosen by looking at the highest pAUC value on the validation set.

The number of combined dichotomizers varies from 2 to 7. For each of those experiments we obtain different number of possible combinations that are independent from each other. Therefore, we use an approach based on giving a rank to each method compared to the others, for each independent experiment. Let us consider the pAUC values $\{pAUC_{ij}\}_{M \times L}$, for $i = 1, \ldots, M$ with M the number of combinations, and for $j = 1, \ldots, L$ with L number of the strategies compared. For each row we assign a rank value r_j^i from 1 to L to each column depending on the pAUC values: the highest pAUC gets rank 1, the second highest the rank 2, and so on until L (in our case $L = 5$). If there are tied pAUCs, the average of the ranks involved is assigned to all pAUCs tied for a given rank. Only in this case it is appropriate to average the obtained ranks on the number of combinations:

$$\bar{r}_j = \frac{1}{M} \sum_{i=1}^{M} r_j^i \tag{19}$$

Figures 1-3 show the results obtained varying the FPR ranges. The higher the curve, i.e. the lower the average rank value, the better the related method.

The results show a clear predominance of the strategies based on single classifier selection. In particular, only the pair choice selection based on the DF index is comparable with pAUC and AUC based selection, specially when the number of the dichotomizers to be combined grows. A probable reason for such outcome is that in the pair selection strategies we use upper bound surrogates of the actual indices and this could sensibly affect the effectiveness of the selection strategy.

A comparison between the two single classifier selection strategies reveals how, even though pAUC is almost always better than AUC, the difference is

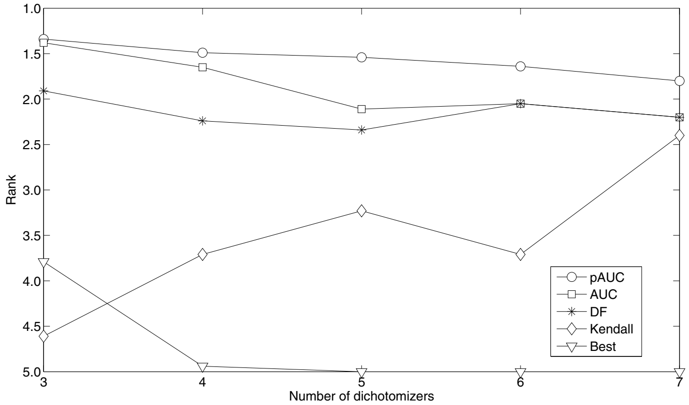

Fig. 2. Mean rank of the selection strategies for t=0.05

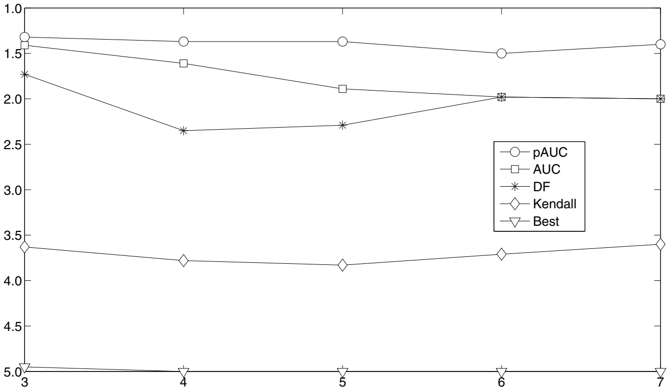

Fig. 3. Mean rank of the selection strategies for t=0.01

not so high for $t = 0.1$. As the FPR range becomes smaller and smaller, the pAUC based strategy clearly outperforms the AUC counterpart and this trend becomes more evident when the number of dichotomizers grows. This suggests that the behavior of the single classifier in the FPR range of interest provides a sufficiently good estimate of how the classifier contributes to the combination.

In summary, a selection strategy which chooses the single dichotomizer with the highest pAUC at each step of the greedy approach described in Sect. 4 seems to ensure the best results on a large extent of situations.

6 Conclusions

In this paper, we have defined and evaluated some strategies which select the dichotomizers to combine at each step of a pAUC combination method based on a greedy approach. The strategies considered are based both on the evaluation of the best single dichotomizer and of the best pair of dichotomizers. Such strategies have been experimentally compared on a biometric database, i.e. an application for which the use of the pAUC is particularly important. The results obtained have shown clearly that the single classifier selection strategies seem the most proficient ones, in particular the selection based on the pAUC of the single dichotomizer. However, it should be taken into account that the indices used by the pair strategies are actually replaced by computationally feasible approximations. Future investigations will be aimed at verifying if more tight (and hopefully more effective) approximations are attainable.

References

1. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. on Knowledge and Data Engineering 17, 299–310 (2005)
2. Cortes, C., Mohri, M.: AUC optimization vs. error rate minimization advances. In: Neural Information Processing Systems. MIT Press, Cambridge (2003)
3. Dodd, L.E., Pepe, M.S.: Partial AUC estimation and regression. Biometrics 59, 614–623 (2003)
4. Ricamato, M.T., Tortorella, F.: Combination of Dichotomizers for Maximizing the Partial Area under the ROC Curve. In: Hancock, E.R., Wilson, R.C., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SSPR&SPR 2010. LNCS, vol. 6218, pp. 660–669. Springer, Heidelberg (2010)
5. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Patt. Recogn. 30, 1145–1159 (1997)
6. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36 (1982)
7. Nandakumar, K., Dass, S.C., Jain, A.K.: Likelihood ratio-based biometric score fusion. IEEE Trans. on Patt. Anal. and Mach. Intell. 30, 342–347 (2008)
8. Kendall, M.G.: A new measure of rank correlation. Biometrika 30, 81–93 (1938)
9. Giacinto, G., Roli, F.: Design of effective neural network ensembles for image classification processes. Image and Vision Computing 19, 699–707 (2001)
10. Poh, N., Bengio, S.: Database, protocol and tools for evaluating score-level fusion algorithms in biometric authentication. Patt. Recogn. 39, 223–233 (2006)