AUC-based combination of dichotomizers: is whole maximization also effective for partial maximization?

Maria Teresa Ricamato, Francesco Tortorella
DAEIMI
Università degli Studi di Cassino
Cassino (FR), Italy
{mt.ricamato, tortorella}@unicas.it

Abstract—The combination of classifiers is an established technique to improve the classification performance. When dealing with two-class classification problems, a frequently used performance measure is the Area under the ROC curve (AUC) since it is more effective than accuracy. However, in many applications, like medical or biometric ones, tests with false positive rate over a given value are of no practical use and thus irrelevant for evaluating the performance of the system. In these cases, the performance should be measured by looking only at the interesting part of the ROC curve. Consequently, the optimization goal is to maximize only a part of the AUC instead of the whole area. In this paper we propose a method tailored for these situations which builds a linear combination of two dichotomizers maximizing the partial AUC (pAUC). Another aim of the paper is to understand if methods that maximize the AUC can maximize also the pAUC. An empirical comparison drawn between algorithms maximizing the AUC and the proposed method shows that this latter is more effective for the pAUC maximization than methods designed to globally optimize the AUC.

Keywords—Area under the ROC Curve; Combination of Classifiers

I. INTRODUCTION

The idea of combining classifiers has received an increasing attention in recent years. It has been shown that the classification performance can often be improved by extending one classifier to a combination of classifiers. Different classifier models offer potentially complementary information about the pattern to be classified [1]. In this way, all the classifier models are used for decision making by combining their individual outputs to derive a final decision. To this aim, several combination rules have been proposed in literature. Most of them are based on the maximization of the Area under the ROC curve (AUC) [2] [3], which is a more suitable performance measure than the classification accuracy [4], particularly in applications where dissymmetrical costs or imbalanced class priors are involved [5] [6].

The AUC summarizes test performance on all the possible values of false positive rates (FPRs). However, there are applications which, instead of focusing on the whole ROC curve, consider only those regions where data have been observed or which correspond to relevant values for the problem. In medical screening, for example, high FPRs values generate large monetary costs, since such errors involve more sophisticated and expensive tests in order to be sorted out. Thus, a decision maker has a maximum FPR threshold which he or she is not willing to exceed. In these cases, the region of interest is the one that corresponds only to acceptable low FPRs, and the partial AUC [7] is the most indicate index to use, since it allows us to focus on particular regions of the ROC space. Therefore, restricting attention to the ROC curve over a practically relevant range of false positive rates is appealing.

Although the partial AUC is a diffused index in Medical Statistics [8] [9] and has gained popularity particularly for evaluating screening techniques, little attention has been given to the use of pAUC in machine learning and specifically as a performance measure in combining classifiers.

In this paper we propose a new method based on the maximization of the pAUC for a linear combination of two dichotomizers. In particular, the algorithm aims at finding the optimal weight in order to achieve the maximum pAUC for the classifiers combination. Furthermore, we empirically show that algorithms designed to maximize the AUC are not as much effective as methods designed for pAUC maximization. The paper is organized as follow: the next section presents a brief introduction to the partial AUC. Section 3 describes the algorithm to maximize the pAUC in a combination of two dichotomizers. Experimental Results performed on the biometric dataset XM2VTS [10] are shown in Section 4, while some conclusion and future work are drawn in the last section.

II. THE ROC CURVE AND THE PARTIAL AREA UNDER THE ROC CURVE

The ROC curve of a two-class classification model describes the trade-off between the fraction of positive samples correctly classified as positive (True Positive Rate, TPR) and the fraction of negative sample incorrectly classified as positive (False Positive Rate, FPR), and gives a description of the performance of the decision rule at different operating points. Although the ROC curve provides a comprehensive evaluation of the decision rule, it is often preferable to employ the Area under the ROC Curve (AUC) [11] [12], which
is a single value measure that resumes the performance of the rule, and it is defined as:

$$AUC = \int_0^1 \text{ROC}(t)dt$$  \hspace{1cm} (1)$$

As previously mentioned, some applications consider only a particular range of false positive rates, hence, tests performed with values of false positive rate outside of a particular domain are not useful for evaluating the accuracy of the test. The summary index that takes into account all the points of the ROC curve in the range \((t_0, t_1)\) of FPRs of interest, is the partial AUC (pAUC), defined as:

$$pAUC = AUC(t_0, t_1) = \int_{t_0}^{t_1} \text{ROC}(t)dt$$  \hspace{1cm} (2)$$

Let us consider \(x_i = f(p_i)\) and \(y_j = f(n_j)\) the outputs of the classifier \(f\) on a positive sample \(p_i\) and a negative sample \(n_j\).

Then, the pAUC is also defined as the probability of \(x_i > y_j\) such that \(y_j\) is in the range of the \(1 - t_k\) quantiles \(q_y^{t_k}\) of the negative samples:

$$pAUC = P\{x_i > y_j, y_j \in \{q_y^{t_1}, q_y^{t_0}\}\}$$  \hspace{1cm} (3)$$

The partial AUC is simply the area under the ROC curve between the FPRs \(t_0\) and \(t_1\).

Selecting the interval \((t_0, t_1)\) is an important practical issue. The choice depends on the particular setting and should depend on the costs of a false positive diagnosis. Without loss of generality, and since most of the applications focus the attention on lower values of FPRs [13] and, in particular, on the range close to the zero value, in the next analysis we consider \(t_0 = 0\).

In order to evaluate the pAUC of a classifier, we use the non-parametric estimator [7], which is defined as:

$$pAUC = \frac{1}{m_pm_N} \sum_i^{m_p} \sum_j^{m_N} \Phi_{ij}^{q_y^{t_0}, q_y^{t_1}}$$  \hspace{1cm} (4)$$

with \(\Phi_{ij}^{q_y^{t_0}, q_y^{t_1}} = I\{x_i > y_j, y_j > q_y^{t_1}\}\), where \(m_p\) and \(m_N\) are the cardinalities of the positive and negative subsets, respectively, and:

\[
I\{x_i > y_j, y_j > q_y^{t_1}\} = \begin{cases} 
1, & \text{if } x_i > y_j \land y_j > q_y^{t_1}; \\
0.5, & \text{if } x_i = y_j \land y_j > q_y^{t_1}; \\
0, & \text{if } x_i < y_j \land y_j > q_y^{t_1}.
\end{cases}
\]

Using this non-parametric estimator avoids to plot the ROC curve and then perform a numerical integration on it in order to evaluate the pAUC.

### III. Combination of two dichotomizers maximizing the pAUC

Let us consider a set \(T\) of samples, and define the outputs of two classifiers \(f_0\) and \(f_1\) on two generic positive and negative samples \(p_i\) and \(n_j\) as shown in [2]:

\[
x_i^0 = f_0(p_i), \quad y_j^0 = f_0(n_j),
\]

\[
x_i^1 = f_1(p_i), \quad y_j^1 = f_1(n_j).
\]

Considering the interval \((0, t_1)\) as the FPRs range, the pAUCs for the two classifiers, evaluated according to the parametric estimator (eq. 4), are:

\[
pAUC_0 = \frac{\sum_i^{m_p} \sum_j^{m_N} I\{x_i^0 > y_j^0, y_j > q_y^{t_1}\}}{m_pm_N}
\]

\[
pAUC_1 = \frac{\sum_i^{m_p} \sum_j^{m_N} I\{x_i^1 > y_j^1, y_j > q_y^{t_1}\}}{m_pm_N}
\]

In a two-class classification problem with two classifiers outputs, \(f_0\) and \(f_1\), finding the linear combination \(\alpha_0f_0 + \alpha_1f_1\) which maximizes the pAUC, is equivalent to find the value \(\alpha = \frac{\alpha_0}{\alpha_1} \in (-\infty, +\infty)\) such that \(f_{lc} = f_0 + \alpha f_1\) maximizes the pAUC. The outputs of \(f_{lc}\) on \(p_i\) and \(n_j\) are as follow:

\[
\xi_i = f_{lc}(p_i) = x_i^0 + \alpha x_i^1,
\]

\[
\eta_j = f_{lc}(n_j) = y_j^0 + \alpha y_j^1.
\]

and the pAUC is the following:

\[
pAUC_{lc} = \frac{1}{m_pm_N} \left( \sum_i^{m_p} \sum_j^{m_N} I\{\xi_i > \eta_j, (\eta_j > q_\eta^{t_1}(\alpha))\} \right)
\]

To find the value \(\alpha_{opt}\) which maximizes \(pAUC_{lc}\), let us analyze the term \(I(\xi_i > \eta_j)\) and in particular let us remind from [2] how it depends on the values of \(I(x_i^0, y_j^0)\) and \(I(x_i^1, y_j^1)\):\(^1\)

- \(I(x_i^0, y_j^0) = 1\) and \(I(x_i^1, y_j^1) = 1\). In this case both the classifiers rank correctly the two samples, pAUC only depends on the value of the quantile (which depends on \(\alpha\)).
- \(I(x_i^0, y_j^0) = 0\) and \(I(x_i^1, y_j^1) = 0\). In this case neither classifier ranks correctly the samples and thus the contribution for the pAUC is 0.
- \(I(x_i^0, y_j^0) \text{xor } I(x_i^1, y_j^1) = 1\). Only one classifier ranks correctly the samples while the other one is wrong. In

\(^1\)For classifiers with continuous output a tie occurs very seldom and thus we assume \(I(x, y) = 0\) when \(x = y\) as well. This implies a negligible underestimate of the empirical pAUC, but it simplifies the analysis.
this case the value of $I(\xi_i > \eta_j)$ depends on the weight $\alpha$.

We can split the subset $T$ in four subsets: $T_{00}$, $T_{01}$, $T_{10}$ and $T_{11}$ defined as:

$$T_{hk} = \{(p_i, n_j) | I(x_i^0, y_j^0) = h \text{ and } I(x_i^1, y_j^1) = k\} \quad (9)$$

Considering the dependence from the quantiles, we define other four subsets: $T'_{00}$, $T'_{01}$, $T'_{10}$ and $T'_{11}$ as follow:

$$T'_{hk} = \{(p_i, n_j) \in T_{hk} | y_j^0 + \alpha y_j^1 > q_n^i(\alpha)\} \quad (10)$$

where $q_n^i$ is the $1 - t_1$ of $\eta$, which depends on the weight $\alpha$.

Therefore, the expression for $pAUC_k$ in eq.(8) can be written as:

$$pAUC_k = \frac{1}{m_p m_N} \left( \sum_{(p_i, n_j) \in T'_{00}} I(\xi_i > \eta_j) + \sum_{(p_i, n_j) \in T'_{01}} I(\xi_i > \eta_j) \right)$$

$$= \frac{0 + \gamma(\alpha) + \nu(\alpha)}{m_p m_N}. \quad (11)$$

and the optimal weight is given by:

$$\alpha_{opt} = \arg \max_{\alpha} (\gamma(\alpha) + \nu(\alpha)). \quad (12)$$

In order to find $\alpha_{opt}$, let us recall that $I(\xi_i > \eta_j) = 1$ only if:

$$x_i^0 - y_j^0 + \alpha(x_i^1 - y_j^1) > 0 \quad (13)$$

such that: $y_j^0 + \alpha y_j^1 > q_n^i(\alpha)$.

Depending on which of the three sets $T'_{11}, T'_{10}, T'_{01}$ are considered, there are three different constraints:

$$\alpha < -\frac{x_i^0 - y_j^0}{x_i^1 - y_j^1} \quad \text{if } (p_i, n_j) \in T'_{01},$$

$$\alpha > -\frac{x_i^0 - y_j^0}{x_i^1 - y_j^1} \quad \text{if } (p_i, n_j) \in T'_{01},$$

$$\alpha > -\frac{x_i^0 - y_j^0}{x_i^1 - y_j^1} \quad \text{if } (p_i, n_j) \in T'_{11}. $$

By maximizing the number of pairs satisfying the previous constraints, the pAUC is maximized. To this intent, we introduce the cumulative functions:

$$F_{10}^1(\alpha) = \text{card} \left( (p_i, n_j) \in T'_{10} | -\frac{\Delta_0}{\Delta_{ij}} > \alpha \right) \quad (15a)$$

$$F_{01}^1(\alpha) = \text{card} \left( (p_i, n_j) \in T'_{01} | -\frac{\Delta_0}{\Delta_{ij}} < \alpha \right) \quad (15b)$$

$$F_{11}^1(\alpha) = \text{card} \left( (p_i, n_j) \in T'_{11} | -\frac{\Delta_0}{\Delta_{ij}} < \alpha \right) \quad (15c)$$

where $\Delta_{ij}^0 = x_i^0 - y_j^0$ and $\Delta_{ij}^1 = x_i^1 - y_j^1$. The function to be maximized can be defined as

$$\gamma(\alpha) + \nu(\alpha) = F_{10}^1(\alpha) + F_{01}^1(\alpha) + F_{11}^1(\alpha) \quad (16)$$

and the optimal value is given by

$$\alpha_{opt} = \arg \max_{\alpha} (F_{10}^1(\alpha) + F_{01}^1(\alpha) + F_{11}^1(\alpha)). \quad (17)$$

that can be easily found with linear search.

IV. EXPERIMENTAL RESULTS

In this section, we show the results obtained on a public-domain biometric dataset namely XM2VTS [10], characterized by the partition in training set and test set given in Table I. The dataset contains 8 matchers, since we are considering combinations of two classifiers at time, we have in total 28 combinations.

The combination rule proposed (pROC) is compared with other two algorithms already described in literature: a method proposed in [3] that provides a linear combination to maximize the AUC under the multivariate normal distribution model, which we are referring as SuLiu, and the DROC method proposed in [2] that maximizes the AUC analyzing the dependence of the AUC of a linear combiner on the weight $\alpha$. It is worth to note that both of the algorithms considered for the comparison evaluate a weight such that the combination maximizes the AUC.

For each considered method the weight for the linear combination is evaluated on the validation set, and then applied to the test set. The pAUC is computed using the non-parametric estimator (eq. 4). The ranges of false positive rates considered are: $FPR_{0.01} = (0, 0.1)$, $FPR_{0.05} = (0, 0.05)$ and $FPR_{0.01} = (0, 0.01)$.

The results are analyzed in term of partial AUC. In order to obtain a synthesis and to evaluate the performance on each method, and for each range of false positive rate considered, we have used relative rank values for all the classifier combinations. For each combination we calculate a rank value, related to each method, from 1 to 3, the highest pAUC gets rank 1, the second highest the rank 2, and so on. If there are tied pAUCs, the average of the ranks involved is assigned to all pAUCs tied for a given rank. Then we calculate the average of the rank for each method on the 28 combinations considered, that is shown on validation set in tables II(a)-IV(a) and on test set in tables II(b)-IV(b), varying the FPR. The lower the value, the better the related method.

Analyzing the results, on both validation and test sets, it is evident that methods with the purpose of maximizing

<table>
<thead>
<tr>
<th>Validation Set</th>
<th># Sample</th>
<th># Positive</th>
<th># Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Set</td>
<td>112200</td>
<td>400</td>
<td>111800</td>
</tr>
</tbody>
</table>

Table I

XM2VTS DATABASE PROPERTIES.
the AUC of a combination of classifiers do not maximize the pAUC. In particular, decreasing the FPR range, the gap between the pROC method and the first of the two other methods considered, in this case the DROC method, increases. It means that the smaller the range the more effective the algorithm of maximization of pAUC. In fact, DROC and SuLiu maximize a metric over all the possible range of FPR, while the pROC focus the attention only in a particular range. Moreover, while DROC and SuLiu methods can be seen as a global maximization, the pROC can be associated to a local maximization. The use of one of the three algorithms depends on the scope of the application under consideration, and mainly on the range of FPRs related to the application.

Table II

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>pROC</td>
<td>1.48</td>
</tr>
<tr>
<td>DROC</td>
<td>1.77</td>
</tr>
<tr>
<td>SuLiu</td>
<td>2.75</td>
</tr>
</tbody>
</table>

Table III

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>pROC</td>
<td>1.50</td>
</tr>
<tr>
<td>DROC</td>
<td>1.79</td>
</tr>
<tr>
<td>SuLiu</td>
<td>2.71</td>
</tr>
</tbody>
</table>

Table IV

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>pROC</td>
<td>1.36</td>
</tr>
<tr>
<td>DROC</td>
<td>1.80</td>
</tr>
<tr>
<td>SuLiu</td>
<td>2.84</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a new linear combination method for maximizing the partial AUC and compared it with algorithms already introduced in literature that maximize the AUC. The new method is based on the dependence of the pAUC on the coefficient $\alpha$ which is found by means of a linear search.

The experiments, performed on a biometric dataset, have shown good performance of the proposed method compared with methods that maximize the AUC. It has been shown that the maximization of the AUC is not related with the maximization of the pAUC, in particular maximizing the metric on all the range of the FPR is not equivalent to maximize the metric in a portion on that range.

The algorithm in its present form is designed for the combination of two dichotomizers. Our immediate future work is to extend our approach to the combination of several dichotomizers.

REFERENCES


