Exploiting AUC for Optimal Linear Combinations of Dichotomizers

Claudio Marrocco, Mario Molinara and Francesco Tortorella

Dipartimento di Automazione, Elettromagnetismo,
Ingegneria dell'Informazione e Matematica Industriale
Università degli Studi di Cassino
03043 Cassino (FR), Italy
{c.marrocco, m.molinara, tortorella}@unicas.it
Accepted for publication on *Pattern Recognition Letters*, Elsevier.

Abstract

The combination of classifiers is an established technique to improve the classification performance. The possible combination rules proposed up to now generally try to decrease the classification error rate, which is a performance measure not suitable in many real situations and particularly when dealing with two class problems. In this case, a good alternative is given by the Area under the Receiver Operating Characteristic curve (AUC), whose effectiveness in measuring the classification quality has been proved in many recent papers.

In this paper, we propose a method to achieve the optimal linear combination of two dichotomizers based on the maximization of the AUC of the resulting classification system. The effectiveness of the approach has been confirmed by the tests performed on standard datasets.

1. Introduction

Dichotomizers (i.e. two-class classifiers) are used in many critical applications (e.g., automated diagnosis, fraud detection, currency verification) which require highly discriminating classifiers. In order to improve the classification performance, a well established technique is to combine more classifiers so as to take advantage of the strengths of the single classifiers and avoid their weaknesses. To this aim, a huge number of possible combination rules has been proposed up to now which generally try to decrease the classification error. However, there are many real situations in which the overall error rate, usually employed as a reference performance measure in classification problems, is not a suitable metric for evaluating the quality of the classifier (Provost et al., 1998). Many applications, for example, involve highly dissymmetrical costs since the consequences coming from distinct kinds of errors are very different and for this reason it is mandatory to consider separately the errors on the single classes. Another case is given by applications such as document retrieval (Cortes and Mohri, 2003) where the dichotomizer is a search engine which

selects a prefixed number of documents from a huge database on the basis of some search criteria and prompts them to the user according to an estimated order of relevance. In this case, the actually significant performance measure is the ranking of the documents rather than the proportion of the correct classifications.

A more effective performance measure for correctly evaluating the dichotomizer in such situations is the Area under the Receiver Operating Characteristic curve (AUC). The advantages of the AUC over the accuracy for evaluating the quality of dichotomizers were described for the first time in the seminal paper by Bradley (Bradley, 1997), who pointed out the increased sensitivity in the Analysis of Variance (ANOVA) tests, the independence from the decision threshold and the invariance to prior class probabilities. More recently, (Huang and Ling, 2005) have established that AUC is an evaluation criterion for the predictive performance of dichotomizers more discriminating than accuracy while (Cortes and Mohri, 2003) have demonstrated that algorithms designed to minimize the error rate may not lead to the best possible AUC values, thus motivating the use of algorithms directly optimizing the AUC. Finally, (Rosset, 2004) has shown that the AUC may be better than empirical error for discriminating between models, even when the ultimate goal is to maximize the accuracy. On these bases several learning algorithms have been proposed; for example (Rakotomamonjy, 2004) presents an algorithm to train Support Vector Machines which directly optimizes the AUC while (Freund et al., 2003) introduces an efficient boosting algorithm for combining multiple rankings based on the maximization of a particular function that (Cortes and Mohri, 2003) show to be coincident with the AUC.

In this paper we propose a method based on AUC maximization to achieve an optimal combination between already trained dichotomizers. In particular, we consider the linear combination since it is the most frequently adopted rule; therefore, the problem faced in this paper is to find the optimal weight which maximizes the AUC of the resulting classification system. To this aim, an analysis of the dependence of the AUC on the weight has been performed and a method to find the optimal weight has been carried out. Experiments made on standard datasets have confirmed the effectiveness of the approach.

The rest of the paper is organized as follows: in the next section we present a short description of the ROC curve and of the AUC measure. Section 3 contains a description of the proposed method, while section 4 shows the obtained experimental results. Some conclusions and possible future developments are drawn in the last section.

2. The ROC curve and the Area under the ROC Curve

In binary classification problems, a sample can be assigned to one of two mutually exclusive classes that can be generically called *Positive* (*P*) class and *Negative* (*N*) class. Without loss of generality, let us assume that the dichotomizer *f* provides, for each sample q, a real value *f*(q) which is a confidence degree that the sample belongs to one of the two classes, e.g. the class *P*. The sample should be consequently assigned to the class *N* if $f(q) \to -\infty$ and to the class *P* if $f(q) \to +\infty$. A threshold *t* is usually chosen, so as to attribute the sample q to the class *N* if $f(q) \leq t$ and to the class *P* if $f(q) > t$. For a given threshold value *t*, two appropriate performance figures are given by the *True Positive Rate TPR*(*t*), i.e. the fraction of actually-positive cases correctly classified and by the *False Positive Rate FPR*(*t*), given by the fraction of actually-negative cases incorrectly classified as "positive". It is important to take into account both quantities for a particular choice of *t* since the consequences of false-negative and false-positive errors are often very different and hard to quantify. The ROC curve plots *TPR*(*t*) vs. *FPR*(*t*) by sweeping the threshold *t* into the whole range of *f*, thus providing a description of the performance of the dichotomizer at different operating points; more important, such description is independent of the prior probabilities of the two classes. A perfectly discriminating dichotomizer has an ROC curve that passes through the upper left corner (where $TPR = 1.0$ and $FPR = 0.0$), while a non discriminating classifier is represented by a 45° diagonal line from the lower left to the upper right corner. Qualitatively, the closer the curve to the upper left corner, the better the dichotomizer.

As well described in many papers devoted to the ROC analysis (Provost and Fawcett, 2001; Fawcett, 2003; Flach, 2003; Tortorella, 2005) the geometrical properties of the ROC curve can be profitably used for optimizing the performance of a dichotomizer with reference to various metrics and classification requirements. However it is often preferable to employ a single value measure which summarizes the performance of the dichotomizer, e.g. because there are some dichotomizers to be compared and there is no any clear predominance of some ROC curve above the others.

The most widely used single measure is the Area Under the ROC Curve (AUC), the value of which intuitively provides an estimate of the quality of the dichotomizer ($AUC = 0.5$ for a non discriminating dichotomizer, $AUC = 1$ for a perfectly discriminating dichotomizer). The AUC has been recently proposed as an alternative single-number measure for evaluating the predictive ability of learning algorithms. (Huang and Ling, 2005) have been shown theoretically and empirically that AUC is a better measure than accuracy and should replace it in comparing learning algorithms. Moreover, the AUC has an important statistical property: the AUC of a dichotomizer measures the probability of correct pair-wise ranking (Hanley and McNeil, 1982; Hand and Till, 2001), i.e. the probability that, given two samples n and p randomly extracted from *N* and *P*, the former will be

given a confidence degree lower than the latter. Such probability can be also estimated by means of the Wilcoxon-Mann-Whitney statistic (Mann and Whitney, 1947) which is defined as follows.

Let S be a set of samples containing n_+ samples belonging to class *P* and n_- samples belonging to class *N* and let *f* be a dichotomizer applied on S. Moreover, let be p_i a positive sample and n_j a negative sample, both coming from S, and let $x_i = f(p_i)$ and $y_j = f(n_j)$ be the outputs of the dichotomizer on such samples. The Wilcoxon-Mann-Whitney (WMW) statistic is defined as:

$$\frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(x_i, y_j)}{n_+ \cdot n_-} \tag{1}$$

where $I(x,y)$ is an indicator function defined as:

$$I(x,y) = \begin{cases} 1 & \text{if} \quad x > y \\ 0.5 & \text{if} \quad x = y \\ 0 & \text{if} \quad x < y \end{cases}$$

In this way, it is possible to evaluate the AUC of *f* directly through (1) without explicitly plotting the ROC curve and estimating the area with a numerical integration.

3. Linear Combination of Dichotomizers via AUC

Let S be a set of samples defined as above. Let us consider two dichotomizers f_0 and f_1 whose outputs on positive and negative samples are:

$$\begin{aligned} x_i^0 &= f_0(p_i) & x_i^1 &= f_1(p_i) \\ y_j^0 &= f_0(n_j) & y_j^1 &= f_1(n_j) \end{aligned} \tag{2}$$

The AUC's for the two dichotomizers evaluated according to the WMW statistic are:

$$AUC_0 = \frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(x_i^0, y_j^0)}{n_+ \cdot n_-} \qquad AUC_1 = \frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(x_i^1, y_j^1)}{n_+ \cdot n_-} \tag{3}$$

Let us now consider a linear combination of f_0 and f_1. Without any loss of generality[1], the resulting classifier can be represented by:

$$f_{lc} = f_0 + \alpha \cdot f_1 \tag{4}$$

where α is the relative weight of f_1 with respect to f_0. The outputs of f_{lc} to p_i and n_j will be consequently:

[1] In general, a linear combination of two classifier is given by $\alpha_0 \cdot f_0 + \alpha_1 \cdot f_1$. However, any decision rule based on the comparison with a threshold τ is equivalent to the decision rule which compares the output of the classifier f_{lc} with the threshold τ/α_0.

$$\begin{aligned}\xi_i &= f_{lc}(\mathrm{p}_i) = x_i^0 + \alpha \cdot x_i^1 \\ \eta_j &= f_{lc}(\mathrm{n}_j) = y_j^0 + \alpha \cdot y_j^1\end{aligned} \tag{5}$$

According to the WMW statistic the AUC of f_{lc} is given by:

$$\mathrm{AUC}_{lc} = \frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(\xi_i, \eta_j)}{n_+ \cdot n_-} \tag{6}$$

and depends on the value of α. Therefore the optimal choice for the weight is the value maximizing AUC_{lc}:

$$\alpha_{opt} = \arg\max \mathrm{AUC}_{lc}(\alpha) \tag{7}$$

To this aim, let us analyze the term $I(\xi_i, \eta_j)$ and study how it depends on the values of $I(x_i^0, y_j^0)$ and $I(x_i^1, y_j^1)$; for the following analysis we consider a tie as an error and thus we group together the cases $I(x,y) = 0.5$ and $I(x,y) = 0$. With this assumption, we can distinguish three cases:

- $I(x_i^0, y_j^0) = 1$ and $I(x_i^1, y_j^1) = 1$: in this case both the dichotomizers rank correctly the two samples and $I(\xi_i, \eta_j) = 1$ whatever the value of α.
- $I(x_i^0, y_j^0) = 0$ and $I(x_i^1, y_j^1) = 0$: in this case neither dichotomizer ranks correctly the samples and thus $I(\xi_i, \eta_j) = 0$ whatever the value of α.
- $I(x_i^0, y_j^0)$ xor $I(x_i^1, y_j^1) = 1$: 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.

According to this result, the set of all the pairs on which AUC_{lc} is evaluated can be split in four subsets S_{00}, S_{11}, S_{01}, S_{10}, where S_{uv} is defined as:

$$S_{uv} = \left\{ (i, j) \middle| I(x_i^0, y_j^0) = u \text{ and } I(x_i^1, y_j^1) = v \right\} \tag{8}$$

As a consequence, the expression for AUC_{lc} can be written as:

$$\mathrm{AUC}_{lc} = \frac{1}{n_+ \cdot n_-}\left[\sum_{(i,j)\in S_{00}} I(\xi_i, \eta_j) + \sum_{(i,j)\in S_{11}} I(\xi_i, \eta_j) + \sum_{(i,j)\in S_{10}\cup S_{01}} I(\xi_i, \eta_j) \right] = \frac{1}{n_+ \cdot n_-}\left[0 + card(S_{11}) + v(\alpha)\right] \tag{9}$$

In other words, while the pairs on which both dichotomizers are wrong do not contribute to AUC_{lc} and the pairs correctly ranked by both the dichotomizers give a contribution independent of the value of α, the dependence of AUC_{lc} on α is limited to the set of pairs *on which the dichotomizers disagree*. Therefore, the larger the set $S_{10} \cup S_{01}$ (i.e., the higher the disagreement between f_0 and f_1), the higher the value of AUC_{lc} which, in principle, can be obtained. Taking into account eq. (9), eq. (7) can be restated as:

$$\alpha_{opt} = \arg\max v(\alpha) \tag{10}$$

In order to find the value of α_{opt} let us make explicit the dependence of $I(\xi_i,\eta_j)$ on α. To this aim, recall that the indicator function is not null only if $\xi_i > \eta_j$, i.e. if:

$$\Delta_{ij}^0 + \alpha \cdot \Delta_{ij}^1 > 0 \tag{11}$$

where $\Delta_{ij}^0 = x_i^0 - y_j^0$ and $\Delta_{ij}^1 = x_i^1 - y_j^1$. The condition (11) leads to different constraint on α depending on which of the two sets S_{01}, S_{10} we consider; in particular we obtain:

$$\alpha < -\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \text{ if } (i,j) \in S_{10} \qquad \alpha > -\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \text{ if } (i,j) \in S_{01} \tag{12}$$

If such conditions were verified for each pair $(i,j) \in S_{10} \cup S_{01}$, we would obtain the max value allowable for $\nu(\alpha)$, i.e. $card(S_{10} \cup S_{01})$. In this case, there would exist an α_{opt} such that

$$\max_{(i,j)\in S_{01}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right) \leq \alpha_{opt} \leq \min_{(i,j)\in S_{10}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right) \tag{13}$$

and the resulting AUC would be:

$$\mathrm{AUC}_{\mathrm{lc}} = \frac{card(S_{11}) + card(S_{10}) + card(S_{01})}{n_+ \cdot n_-} = \mathrm{AUC}_0 + \mathrm{AUC}_1 - \frac{card(S_{11})}{n_+ \cdot n_-} \tag{14}$$

where $\mathrm{AUC}_0 = \dfrac{card(S_{11}) + card(S_{10})}{n_+ \cdot n_-}$ and $\mathrm{AUC}_1 = \dfrac{card(S_{11}) + card(S_{01})}{n_+ \cdot n_-}$.

However the condition $\max_{(i,j)\in S_{01}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right) \leq \min_{(i,j)\in S_{10}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ is verified only when the two dichotomizers exhibit a high degree of complementarity. In particular, the term $\min_{(i,j)\in S_{10}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ becomes high when, for each pair $(i,j) \in S_{10}$, the dichotomizer f_0 correctly ranks the corresponding pair (p_i,n_j) producing a high difference $\left| x_i^0 - y_j^0 \right|$ between the outputs, while f_1, even though incorrectly ranking (p_i,n_j), provides a low difference $\left| x_i^1 - y_j^1 \right|$. This means that the errors made by f_1 can be recovered thanks to the good performance of f_0 on the same pairs. Conversely, a low value for the term $\max_{(i,j)\in S_{01}} \left(-\frac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ is obtained when, for each pair $(i,j) \in S_{01}$, the dichotomizer f_1 correctly ranks the corresponding pair (p_i,n_j) with a high difference $\left| x_i^1 - y_j^1 \right|$ between the outputs, while f_0 incorrectly ranks (p_i,n_j) but with a low difference $\left| x_i^0 - y_j^0 \right|$. In this case f_1 helps in recovering the erroneous rankings produced by f_0. When eq. (13) is verified, the value of α_{opt} allows eliminating all the errors made by both the dichotomizers, except for those on which f_0 and f_1 agree.

Unfortunately, such condition is only rarely verified since the distributions of the ratio $-\frac{\Delta_{ij}^0}{\Delta_{ij}^1}$ evaluated on the two sets S_{10} and S_{01} are usually not separated (see fig. 1).

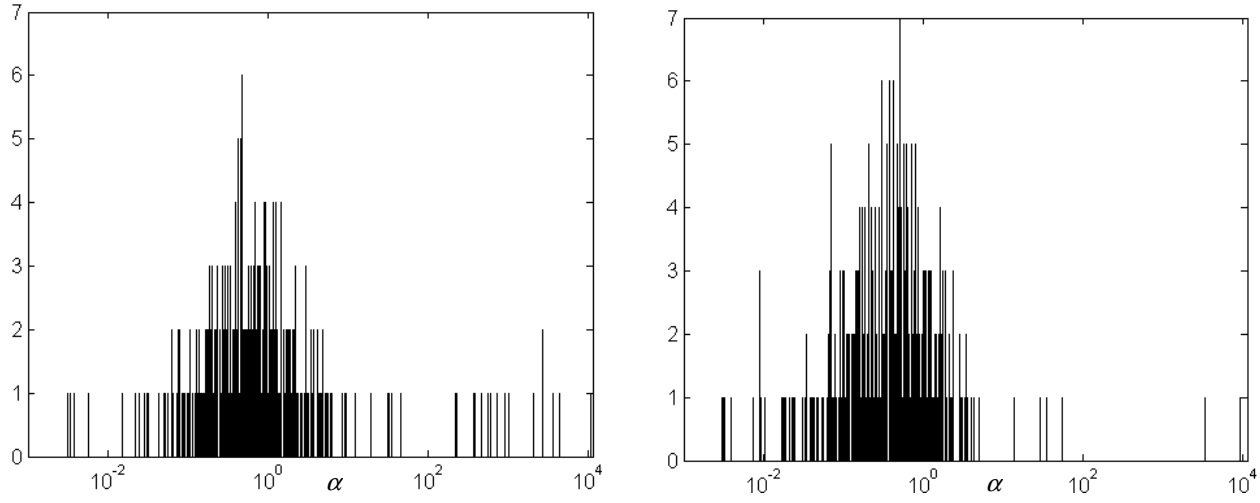


Fig. 1. Example of the distributions of the ratio $-\frac{\Delta_{ij}^0}{\Delta_{ij}^1}$ evaluated on S_{10} **(a)** and S_{01} **(b)**.

As a consequence, α_{opt} has to be found by maximizing the number of the pairs satisfying eq. (12). To this aim, if we consider the cumulative functions

$$F_{10}(\alpha) = card\left((i,j) \in S_{10} \left| -\frac{\Delta_{ij}^0}{\Delta_{ij}^1} > \alpha \right. \right) \qquad F_{01}(\alpha) = card\left((i,j) \in S_{01} \left| -\frac{\Delta_{ij}^0}{\Delta_{ij}^1} < \alpha \right. \right)$$

the function to be maximized can be defined as:

$$\nu(\alpha) = F_{10}(\alpha) + F_{01}(\alpha)$$

and the optimal value is given by:

$$\alpha_{opt} = \arg\max F_{10}(\alpha) + F_{01}(\alpha) \tag{15}$$

that can be easily found by means of a linear search.

An example of real distributions of the ratio $-\frac{\Delta_{ij}^0}{\Delta_{ij}^1}$ evaluated on the two sets S_{10} and S_{01} is shown in fig. 1, while fig. 2 shows the relative function $\nu(\alpha) = F_{10}(\alpha) + F_{01}(\alpha)$.

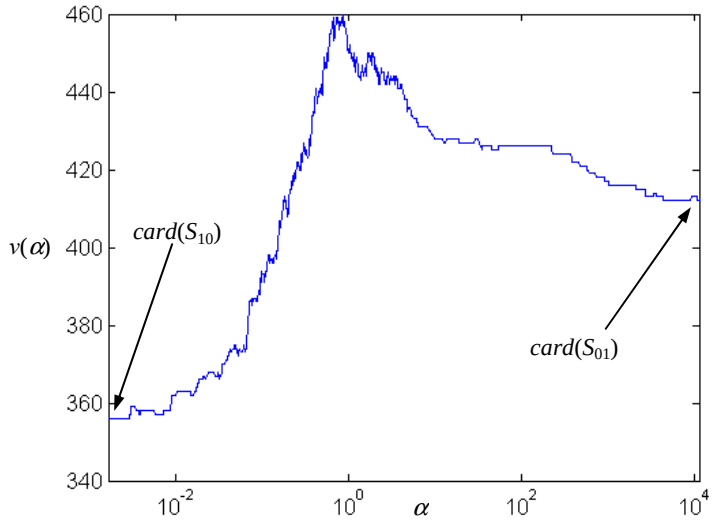


Fig. 2. The trend of $\nu(\alpha)=F_{10}(\alpha)+F_{01}(\alpha)$ obtained by the two distributions shown in fig. 1. The points in which the combination reduces to one of the dichotomizers (see text) are shown.

If we consider what happens at the bounds of the range of α, it is possible to observe that, if $\alpha \to 0$, $\nu(\alpha) \to card(S_{10})$ and the combination reduces to the only dichotomizer f_0, while when $\alpha \to \infty$, $\nu(\alpha) \to card(S_{01})$ and the combination reduces to the only dichotomizer f_1. These extreme points are shown in fig. 2: in this case, since the maximum of $\nu(\alpha)$ is higher than both the bound values, we have $\mathrm{AUC}_{lc}(\alpha_{opt}) > \mathrm{AUC}_0$ and $\mathrm{AUC}_{lc}(\alpha_{opt}) > \mathrm{AUC}_1$, i.e. the linear combination performs better than each of the two dichotomizers.

As a concluding remark, it is worth noting that the method cannot be applied when $card(S_{10}) = 0$ or $card(S_{01}) = 0$. However, in this case the combination is not profitable since it does not give better results than the single dichotomizer. In fact, if e.g. $card(S_{01}) = 0$, there are no pairs incorrectly ranked by f_0 which are correctly ranked by f_1 and thus the combination is useless since it cannot recover any error made by f_0.

4. Experimental Results

In order to evaluate the performance of the proposed method, it has been tested on five datasets publicly available at the UCI Machine Learning Repository (Blake et al., 1998). All of them have two classes and a variable number of numerical input features. The features were previously rescaled so as to have zero mean and unit standard deviation. To avoid any bias in the comparison, 10 runs of a multiple hold out procedure have been performed on all data sets. In each run, the dataset has been parted into three sets: a training set to train the classifier, a validation set to estimate the optimal weight and a test set to assess the reliability of the obtained weight. More details are given in table 1.

Table 1. Datasets used in the experiments

Datasets	# Features	# Samples	% Positive	% Negative	Training Set	Validation Set	Test Set
Pima Indian Diabetes	8	768	34.9	65.1	538	115	115
Cleveland Heart Disease	13	297	54.8	45.2	208	44	45
German Credit	24	1000	30.0	70.0	700	150	150
Contraceptive Method Choice	9	1473	57.3	42.7	1031	221	221
Breast Cancer Wisconsin	9	698	34.3	66.7	489	104	105

The employed dichotomizers are Support Vector Machines (SVM) and Multi Layer Perceptrons (MLP). The SV-based classifiers have been implemented by means of SVM^{light} tool (Joachims, 1999) while for the MLPs we have used the NODElib library (Flake and Pearlmuter, 2000). Three different kernels have been used for the SVMs while for the MLPs we have considered three classifiers with different numbers of units in the hidden layer, all trained for 10000 epochs using the back propagation algorithm with a learning rate of 0.01. The characteristics of the six different dichotomizers are reported in table 2.

Table 2. Acronyms of classifiers used in the experiments

Type of Classifier	Type of Kernel or Number of Hidden Nodes	Acronym
SVM	Linear	SL
SVM	Polynomial of degree 2	SP2
SVM	RBF with σ=1	SR1
MLP	2	M2
MLP	4	M4
MLP	6	M6

In the performed experiments we have considered all the 15 combinations which can be accomplished with the classifiers described in table 2; for each combination, we have evaluated the weight α_{opt} by means of the proposed method on the validation set and then the achieved AUC through the WMW statistic on the validation and the test set.

For the sake of comparison, we have also considered another method which trivially chooses the weight maximizing the AUC through an exhaustive search. In particular, this method considers the set of values for α varying in the range [0,50] with a step of 0.01; for each of them, the outputs of the two classifiers on the validation set are combined according to eq.(4) and the relative AUC is computed through the WMW statistic. Finally the value of α corresponding to the maximum AUC is picked out. The aim here is not to provide another algorithm to construct the optimal combination, but to obtain a reliable estimate of the weight maximizing the AUC on the validation set, which can be compared with the α_{opt} provided by the proposed method[2].

[2] It is worth noting that the involved computational complexity is very high: $O(N_p \cdot n_+ \cdot n_-)$ where N_p is the number of points considered for α.

In this way, for each data set, the hold out procedure provides 10 AUC values for each method. This allows us to employ the Wilcoxon rank-sum test (Walpole et al., 1998), so as to verify if the differences in the means of the two populations are statistically significant. All the results were provided with a significance level equal to 0.05.

Let us firstly analyze the results obtained on the validation set which are reported in table 3-7. Each entry of the tables contains the mean and the standard deviation of the AUC values obtained in the 10 runs of the hold out procedure. An underlined and bold value means that such value is significantly better than the other one. If the compared methods have undistinguishable means the values are in normal style.

Table 3. Results on the validation set for Pima dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.838±0.029	0.838±0.029
SL-M4	0.841±0.028	0.841±0.028
SL-M6	0.838±0.027	0.838±0.027
SP2-M2	0.833±0.029	0.832±0.030
SP2-M4	0.831±0.034	0.832±0.034
SP2-M6	0.832±0.032	0.832±0.032
SR1-M2	0.809±0.031	0.809±0.031
SR1-M4	0.803±0.037	0.803±0.037
SR1-M6	0.795±0.028	0.795±0.028
SL-SP2	0.842±0.029	**0.843±0.029**
SL-SR1	0.846±0.027	**0.847±0.027**
SP2-SR1	0.833±0.041	**0.833±0.041**
M2-M4	0.815±0.031	0.815±0.030
M2-M6	0.807±0.033	0.806±0.032
M4-M6	0.798±0.034	**0.798±0.034**

Table 4. Results on the validation set for German dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.803±0.040	0.803±0.040
SL-M4	0.801±0.038	0.802±0.037
SL-M6	0.808±0.043	0.808±0.043
SP2-M2	0.791±0.038	0.791±0.038
SP2-M4	0.782±0.034	**0.782±0.034**
SP2-M6	0.789±0.048	0.789±0.048
SR1-M2	0.764±0.029	0.764±0.029
SR1-M4	**0.740±0.017**	0.738±0.019
SR1-M6	0.742±0.050	0.741±0.050
SL-SP2	0.809±0.040	**0.810±0.040**
SL-SR1	0.803±0.038	0.803±0.038
SP2-SR1	0.771±0.043	0.771±0.043
M2-M4	0.762±0.029	**0.762±0.029**
M2-M6	0.776±0.040	**0.777±0.040**
M4-M6	0.764±0.037	0.764±0.037

Table 5. Results on the validation set for Cleveland dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.921±0.022	0.922±0.022
SL-M4	0.915±0.025	**0.916±0.025**
SL-M6	0.923±0.023	0.924±0.024
SP2-M2	0.895±0.037	**0.896±0.037**
SP2-M4	0.887±0.033	**0.888±0.033**
SP2-M6	0.893±0.038	0.893±0.037
SR1-M2	0.886±0.051	**0.887±0.049**
SR1-M4	0.871±0.042	0.872±0.041
SR1-M6	**0.869±0.057**	0.867±0.058
SL-SP2	0.912±0.026	0.912±0.026
SL-SR1	0.915±0.023	**0.916±0.023**
SP2-SR1	0.874±0.037	**0.875±0.037**
M2-M4	0.895±0.047	0.895±0.047
M2-M6	0.884±0.059	0.885±0.059
M4-M6	0.883±0.045	**0.884±0.045**

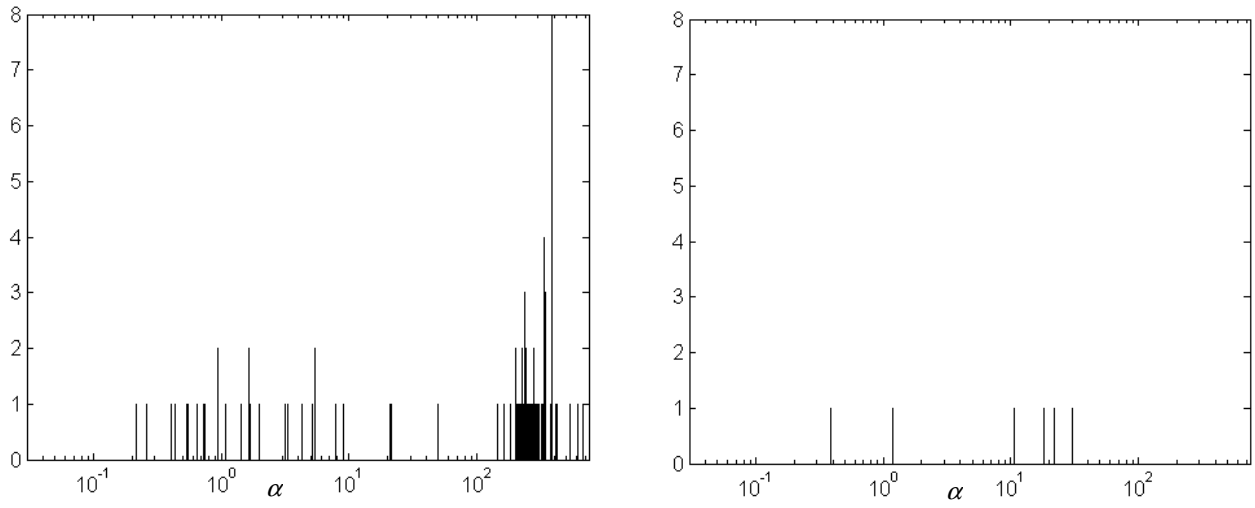
Table 6. Results on the validation set for CMC dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.757±0.037	0.757±0.037
SL-M4	0.753±0.030	0.753±0.030
SL-M6	0.746±0.038	**0.746±0.038**
SP2-M2	0.765±0.035	0.765±0.035
SP2-M4	0.763±0.032	**0.763±0.032**
SP2-M6	0.764±0.032	0.764±0.032
SR1-M2	0.756±0.034	0.755±0.034
SR1-M4	0.746±0.027	0.747±0.027
SR1-M6	0.742±0.033	**0.742±0.034**
SL-SP2	0.761±0.037	0.761±0.037
SL-SR1	0.743±0.033	**0.743±0.033**
SP2-SR1	0.757±0.036	0.757±0.036
M2-M4	0.758±0.029	0.758±0.029
M2-M6	0.758±0.035	0.758±0.035
M4-M6	0.750±0.028	0.750±0.028

Table 7. Results on the validation set for Breast dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.995±0.003	0.995±0.003
SL-M4	0.995±0.003	**0.996±0.003**
SL-M6	0.995±0.004	**0.995±0.004**
SP2-M2	0.995±0.004	**0.995±0.004**
SP2-M4	0.995±0.004	0.996±0.003
SP2-M6	0.995±0.004	0.996±0.004
SR1-M2	0.986±0.009	0.986±0.009
SR1-M4	0.990±0.005	**0.990±0.005**
SR1-M6	0.987±0.009	**0.987±0.009**
SL-SP2	0.996±0.003	0.996±0.003
SL-SR1	0.995±0.004	0.995±0.003
SP2-SR1	0.996±0.004	0.996±0.004
M2-M4	0.980±0.016	0.980±0.016
M2-M6	0.980±0.017	0.980±0.017
M4-M6	0.973±0.027	**0.976±0.027**

From these results we can see that the proposed method performs well since the AUC values obtained are quite indistinguishable from those provided by the exhaustive search and thus the α_{opt} is actually able to maximize the AUC of the resulting classifier. It is worth noting that, in the case of Breast dataset, we have frequently obtained an extreme value for α_{opt} which excludes one dichotomizer from the combination. This is due to the very good performance reached by the best single dichotomizer which leads to two possible situations: one of the sets S_{01} or S_{10} is empty (i.e. the samples erroneously classified by a classifier are not correctly classified by the other) or there is a very low number of samples in one of the two sets. In the first case, one of the two dichotomizers is useless (as said at the end of section 3) because it cannot recover any error made by the other classifier. In the second case, the distributions of the ratio $-\frac{\Delta_{ij}^0}{\Delta_{ij}^1}$ evaluated on the two sets S_{10} and S_{01} can be very imbalanced as shown in fig. 3, where the distributions for the linear combination of an SL and an M4 are reported.



(a) (b)

Fig. 3. The distributions of the ratio $\frac{\Delta_{ij}^{0}}{\Delta_{ij}^{1}}$ evaluated on S_{10} **(a)** and S_{01} **(b)** for the linear combination of an SL with an M4.

In this case, each value of α greater than zero leads to a lower value for $F_{01}(\alpha)+F_{10}(\alpha)$ because the minimum value of α which allows some errors of f_0 to be recovered produces a higher number of errors of f_1 which can be no longer recovered. This can be clearly seen in figure 4 where is shown that in this case $F_{01}(\alpha)+F_{10}(\alpha)$ is a monotonically decreasing function whose maximum is reached for α=0, i.e. when the combination reduces to one dichotomizer.

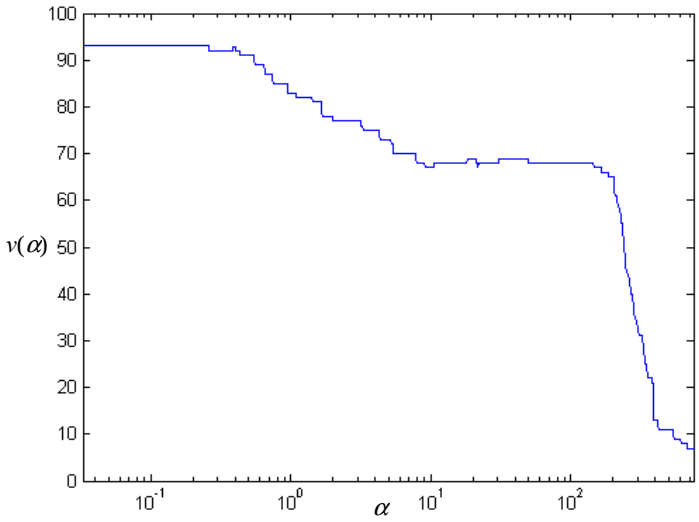


Fig. 4. The trend of $\nu(\alpha)=F_{01}(\alpha)+F_{10}(\alpha)$ for the linear combination of an SL with an M4.

Let us now analyze the results obtained on the test sets in terms of the AUC calculated combining the two dichotomizers with the weights estimated on the validation set. The results are reported in tables 8-12 and are structured in the same way as before. Even in this case the proposed method provides practically the same results as the exhaustive search, thus proving that α_{opt} is a good estimate of the optimal combination weight also on the test sets.

Table 8. Results on the test set for Pima dataset

	Proposed Method	Exhaustive Search
SL-M2	0.837±0.036	0.836±0.036
SL-M4	0.831±0.026	0.831±0.027
SL-M6	0.834±0.036	0.835±0.033
SP2-M2	0.831±0.032	0.831±0.031
SP2-M4	0.826±0.031	0.827±0.031
SP2-M6	0.831±0.031	0.831±0.031
SR1-M2	0.796±0.032	0.797±0.032
SR1-M4	0.796±0.021	0.793±0.022
SR1-M6	0.773±0.033	0.773±0.033
SL-SP2	0.835±0.031	0.835±0.031
SL-SR1	0.829±0.024	0.829±0.025
SP2-SR1	0.825±0.022	0.826±0.022
M2-M4	0.811±0.037	0.812±0.037
M2-M6	0.798±0.045	0.798±0.045
M4-M6	0.790±0.045	0.790±0.046

Table 9. Results on the test set for German dataset

	Proposed Method	Exhaustive Search
SL-M2	0.793±0.045	0.793±0.046
SL-M4	0.790±0.041	0.790±0.041
SL-M6	**0.790±0.042**	0.789±0.042
SP2-M2	0.755±0.049	0.755±0.049
SP2-M4	0.720±0.038	0.719±0.042
SP2-M6	0.705±0.049	0.711±0.043
SR1-M2	0.764±0.029	0.764±0.029
SR1-M4	0.740±0.017	0.738±0.019
SR1-M6	0.742±0.050	0.741±0.050
SL-SP2	0.796±0.045	0.796±0.045
SL-SR1	0.788±0.044	0.790±0.042
SP2-SR1	0.744±0.049	0.746±0.049
M2-M4	0.750±0.040	0.751±0.038
M2-M6	**0.753±0.057**	0.753±0.058
M4-M6	0.731±0.042	0.732±0.041

Table 10. Results on the test set for Cleveland dataset

	Proposed Method	Exhaustive Search
SL-M2	0.899±0.028	0.900±0.027
SL-M4	0.902±0.025	0.901±0.026
SL-M6	0.888±0.051	**0.891±0.049**
SP2-M2	0.877±0.033	0.876±0.032
SP2-M4	0.874±0.045	0.874±0.047
SP2-M6	0.846±0.052	0.847±0.053
SR1-M2	0.869±0.056	0.870±0.055
SR1-M4	0.827±0.070	0.835±0.060
SR1-M6	0.806±0.075	0.808±0.074
SL-SP2	0.901±0.027	0.902±0.028
SL-SR1	0.900±0.030	0.901±0.027
SP2-SR1	0.839±0.045	0.850±0.055
M2-M4	0.871±0.044	0.867±0.050
M2-M6	**0.856±0.065**	0.850±0.066
M4-M6	0.835±0.058	0.836±0.060

Table 11. Results on the test set for CMC dataset

	Proposed Method	Exhaustive Search
SL-M2	0.756±0.029	0.755±0.030
SL-M4	0.735±0.033	0.736±0.033
SL-M6	0.745±0.034	0.745±0.035
SP2-M2	0.752±0.024	**0.753±0.025**
SP2-M4	0.746±0.027	0.746±0.027
SP2-M6	0.750±0.035	0.751±0.036
SR1-M2	0.757±0.021	0.756±0.021
SR1-M4	0.732±0.022	0.733±0.022
SR1-M6	0.740±0.027	0.740±0.027
SL-SP2	0.758±0.032	0.758±0.031
SL-SR1	0.737±0.034	0.737±0.034
SP2-SR1	0.753±0.029	0.753±0.029
M2-M4	0.746±0.023	0.746±0.023
M2-M6	0.756±0.025	0.755±0.025
M4-M6	0.756±0.025	0.755±0.025

Table 12. Results on the test set for Breast dataset

	Proposed Method	**Exhaustive Search**
SL-M2	0.988±0.024	0.988±0.024
SL-M4	0.991±0.011	**0.992±0.009**
SL-M6	0.980±0.049	**0.980±0.049**
SP2-M2	0.988±0.023	0.990±0.023
SP2-M4	0.995±0.004	0.995±0.004
SP2-M6	0.981±0.049	0.981±0.049
SR1-M2	0.975±0.025	**0.977±0.025**
SR1-M4	0.982±0.010	0.982±0.010
SR1-M6	0.970±0.046	0.970±0.045
SL-SP2	**0.995±0.004**	0.995±0.005
SL-SR1	0.994±0.005	0.994±0.005
SP2-SR1	0.995±0.004	0.993±0.007
M2-M4	0.960±0.046	0.960±0.046
M2-M6	0.950±0.051	0.950±0.051
M4-M6	0.950±0.055	0.950±0.055

Finally, let us make some considerations about the computational complexity of the proposed method. The first step estimates the distributions of the two ratios $-\Delta_{ij}^{0}/\Delta_{ij}^{1}$: the complexity is $O(n_+ \cdot n_-)$since it depends on the number of pairs (p_i,n_j) to be considered, that are $n_+ \cdot n_-$. The second step is the location of the maximum of the sum of the two cumulative functions F_{10} and F_{01}; in the worst case (i.e., if we don't use any efficient search and any pair provides a different value for $-\Delta_{ij}^{0}/\Delta_{ij}^{1}$), this requires $n_+ \cdot n_-$ comparisons and thus also this step is $O(n_+ \cdot n_-)$. As a consequence, the overall computational complexity of the method is $O(n_+ \cdot n_-)$.

Conclusions

In this paper we have proposed a method for optimizing in terms of AUC the linear combination of two dichotomizers. The method is based on an analysis of the dependence of the AUC of the linear combiner on the weight α. Such analysis has also pointed out that there is a direct relation between the performance improvement attainable with the combination and a measure of the disagreement existing between the dichotomizers. The experiments made on standard datasets have demonstrated that the algorithm actually allows the optimal value of the weight to be found.

The proposed approach has considered the linear combination between two dichotomizers. The future researches will consider the extension of the linear combiner to more than two dichotomizers as well as other combination rules.

References

Blake, C., Keogh, E., Merz, C.J., 1998. UCI Repository of Machine Learning Databases. [www.ics.uci.edu/~mlearn/MLRepository.html]

Bradley, A.P., 1997. The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition 30, 1145-1159.

Cortes, C., Mohri, M., 2003. AUC Optimization vs. Error Rate Minimization. Advances in Neural Information Processing Systems (NIPS 2003).

Fawcett, T., 2003. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. HP Labs Tech Report HPL-2003-4.

Flach P.A., 2003. The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. Proc. Int. Conf. on Machine Learning (ICML-2003).

Flake, G.W., Pearlmuter, B.A., 2000. Differentiating Functions of the Jacobian with Respect to the Weights. In S. A. Solla, T. K. Leen, and K. Müller, eds., Advances in Neural Information Processing Systems 12, The MIT Press.

Freund, Y., Iyer, R., Schapire, R., & Singer, Y., 2003. An Efficient Boosting Algorithm for Combining Preferences. Journal of Machine Learning Research 4, 933-969.

Hand D.J., Till R.J., 2001. A Simple Generalization of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171-186.

Hanley J.A., McNeil B.J., 1982. The Meaning and the Use of the Area under a Receiver Operating Characteristic Curve. Radiology 143, 29-36.

Huang, J., Ling, C.X., 2005. Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering 17, 299-310.

Joachims, T., 1999. Making Large-Scale SVM Learning Practical. In B. Schölkopf, C.J.C. Burges, A.J. Smola, eds., Advances in Kernel Methods - Support Vector Learning, MIT Press, 169-184.

Mann, H.B., Whitney, D.R., 1947. On a Test whether One of Two Random Variable Is Stochastically Larger than the Other. Annals of Mathematical Statistics 18, 50-60.

Provost, F., Fawcett, T., 2001. Robust Classification for Imprecise Environments. Machine Learning 42, 203-231.

Provost, F., Fawcett, T., Kohavi, R., 1998. The Case against Accuracy Estimation for Comparing Induction Algorithms. Proc. Int. Conf. on Machine Learning (ICML-1998), Morgan Kaufmann, 445-453.

Rakotomamonjy, A., 2004. Optimizing Area Under ROC Curve with SVMs. Workshop on ROC Analysis in Artificial Intelligence.

Rosset, S., 2004. Model selection via the AUC. Proc. Int. Conf. on Machine Learning (ICML-2004).

Tortorella, F., 2005. A ROC-based Reject Rule for Dichotomizers. Pattern Recognition Letters 26, 167-180.

Walpole, R.E., Myers, R.H., Myers, S.L., 1998. Probability and Statistics for Engineers and Scientists. 6th ed., Prentice Hall Int., London.