Optimal Linear Combination of Dichotomizers via AUC

Claudio Marrocco C.MARROCCO@UNICAS.IT
Mario Molinara M.MOLINARA@UNICAS.IT
Francesco Tortorella TORTORELLA@UNICAS.IT
Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e Matematica Industriale, Università degli Studi di Cassino,03043 Cassino (FR), Italy

Abstract

A well established technique to improve the classification performances is to combine more classifiers. However, 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, an effective instrument to analyze the dichotomizers under different class and cost distributions is the Receiver Operating Characteristic (ROC) curve. In particular, a good performance measure 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 consider the linear combination of two dichotomizers since it is the most frequently adopted combination rule and propose a method to achieve the optimal weight of the combination 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 c[1]lassifiers 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 rule has been proposed up to now which generally try to decrease the classification error. However, the applications considered frequently involve cost matrices and class distributions both strongly asymmetric and dynamic and in such cases 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).

A more effective tool for correctly quantifying the performance of the dichotomizer and for analyzing it under different class and cost distributions is given by the Receiver Operating Characteristic (ROC) curve. It provides a description of the performance of the dichotomizer at different operating points, which is independent of the prior probabilities of the two classes. ROC analysis is based in statistical decision theory and in the Pattern Recognition field, is increasingly adopted for many central issues such as the evaluation of machine learning algorithms (Bradley, 1997), the robust comparison of classifier performance under imprecise class distribution and misclassification costs (Provost & Fawcett, 2001) and the definition of a reject option for dichotomizers (Tortorella, 2005). Moreover, 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.

To this aim, 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). Moreover, the AUC is a more effective performance measure for correctly evaluating the dichotomizer: the advantages of the AUC over the accuracy for evaluating the quality of

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

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, (Ling et al., 2003) have established that AUC is an evaluation criterion for the predictive performance of dichotomizers more discriminating than accuracy while (Cortes & 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.

To this aim, several approaches have been proposed to optimize the AUC of single classifiers such as Support Vector Machines (Rakotomamonjy, 2004) or to apply optimization algorithms to particular problems (Yan et al., 2003) and to classifier combination based on boosting (Freund et al., 2003).

However, while in the last case an exponential approximation of the empirical ranking is optimized, in this paper we propose a method to directly optimize the AUC. So, we consider the linear combination of two dichotomizers since it is the most frequently adopted combination rule and propose a method based on AUC maximization to achieve the optimal weight of the combination. To this aim, an analysis of the dependence of the AUC of the linear combiner on the weighting is presented together with the results of experiments performed on standard datasets, that confirmed the effectiveness of the approach.

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

2. Linear Combination of Dichotomizers based on AUC Maximization

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. A threshold *t* is usually chosen, so as to attribute the sample q to the class *N* if $f(q) \le t$ 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". 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. Qualitatively, the closer the curve to the upper left corner, the better the dichotomizer.

An effective performance measure to correctly evaluate the dichotomizer is the AUC. Actually the AUC of a dichotomizer measures the probability of correct pair-wise ranking (Hanley & McNeil, 1982; Hand & Till, 2001), i.e. the probability that, given two samples n and p randomly extracted from *N* and *P*, the former has a confidence degree lower than the latter. This probability can be estimated by means of the Wilcoxon-Mann-Whitney statistic .

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 and n_j respectively a positive and a negative sample, both coming from S, and let $x_j = f(p_j)$ and $y_j = f(n_j)$ be the outputs of the dichotomizer on such samples. Then, 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:

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

In this way, the AUC of *f* is directly evaluated without explicitly plotting the ROC curve and estimating the area with a numerical integration.

Now, 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:

$$\mathrm{AUC}_0 = \frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(x_i^0, y_j^0)}{n_+ \cdot n_-} \qquad \mathrm{AUC}_1 = \frac{\sum_{i=1}^{n_+}\sum_{j=1}^{n_-} I(x_i^1, y_j^1)}{n_+ \cdot n_-} \tag{3}$$

Let us now consider a linear combination of f_0 and f_1. Without any loss of generality[2], 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:

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

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 analysis following 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.
- $I(x_i^0, y_j^0)$ 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:

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

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:

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

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:

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

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

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

and the resulting AUC would be:

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

[2] In general, a linear combination of two classifier is given by $\alpha_0 \cdot \mathrm{f}_0 + \alpha_1 \cdot \mathrm{f}_1$. 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.

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 $\max_{(i,j)\in S_{01}} \left(-\dfrac{\Delta_{ij}^0}{\Delta_{ij}^1} \right) \leq \min_{(i,j)\in S_{10}} \left(-\dfrac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ is verified only when the two dichotomizers exhibit together a high degree of complementarity. In particular, the term $\min_{(i,j)\in S_{10}} \left(-\dfrac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ 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 $\max_{(i,j)\in S_{01}} \left(-\dfrac{\Delta_{ij}^0}{\Delta_{ij}^1} \right)$ 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 the recovery of all the errors made by both the dichotomizers, excepting those on which f_0 and f_1 agree.

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

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

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

the function to be maximized can be defined as:

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

and the optimal value is given by:

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

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

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.

3. Experimental Results

In these experiments we want to show how the proposed method can be used to identify the optimal weight without evaluating the real output of the combined classifier.

To this aim, three datasets publicly available at the UCI Machine Learning Repository (Blake et al., 1998) have been used; 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 were performed on all data sets. In each run, the dataset has been divided into two sets: a Training Set used to train the classifier and a Validation set used to evaluate the effectiveness of the proposed method, i.e. to optimize the weight of the combination. More details are given in table 1.

Table 1. Datasets used in the experiments

Datasets	# Features	# Samples	Training Set	Validation Set
Pima	8	768	538	230
German	24	1000	700	300
Breast	9	698	489	209

The dichotomizers employed are Support Vector Machine (SVM) and Multi Layer Perceptron (MLP). The former has been implemented by means of SVM^{light} tool (Joachims, 1999) while for the latter we used the NODElib library (Flake & Pearlmuter, 2000). Three different kernels have been used for the SVM while for the MLP we varied the number of the units in the hidden layer so that to obtain two different dichotomizers. Their characteristics are described in table 2. The training of the MLP has been performed on 10000 epochs using the back propagation algorithm with a learning rate of 0.01.

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 variance 1	SR1
MLP	2	M2
MLP	4	M4

In the performed experiments we have considered all the 10 combinations which can be accomplished with the classifiers described in table 2. For each combination, we have compared the AUC achieved by using the value of α_{opt} obtained by means of the proposed method with the highest AUC that could be obtained through an exhaustive search on α. In particular, for the latter case, we have computed the values of AUC obtained with α varying in the range [0,50] with a step of 0.01 and chosen the maximum.

The results obtained for the first two datasets are shown in figs. 1-2. Each graph reports, for each combination, the mean AUCs obtained with the two methods together with the mean AUC of the best dichotomizer. In each graph we have also reported the error bars for each considered method.

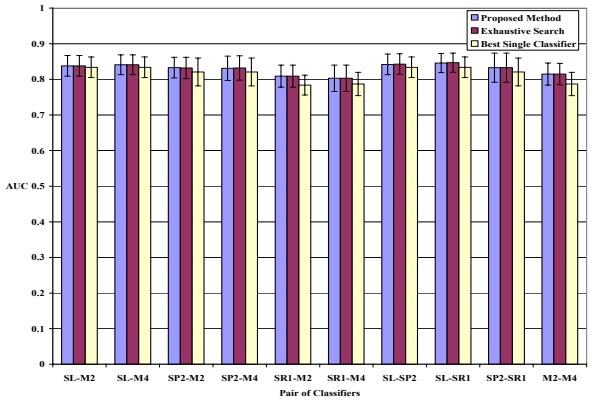

Figure1. Results on Pima Indian Diabetes dataset in terms of AUC.

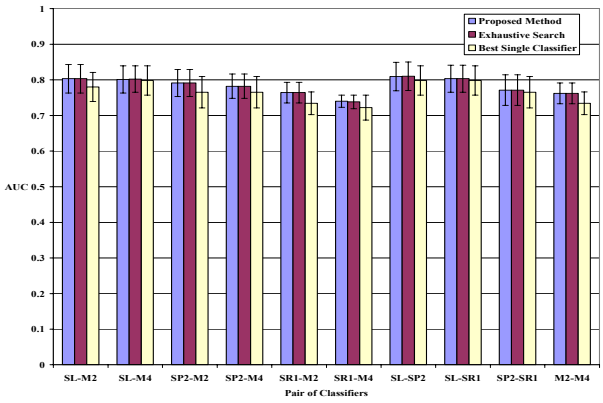

Figure2. Results on German Credit dataset in terms of AUC.

From these results we can see that the proposed method actually determines the optimal weight. In fact, the performance of the combined classifier are quite the same in all the examined cases for all the datasets; moreover, our method provides in some cases a weight considerably better than the exhaustive search.

The results obtained with the Breast dataset (fig. 3) should be more thoroughly analyzed since, in this case, the combination can improve the performance of the best dichotomizer only in few cases. This is due to the very good performance reached by the best dichotomizer (fig. 3 shows that the AUC of the best dichotomizer is at least 0.97 on this dataset) 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.

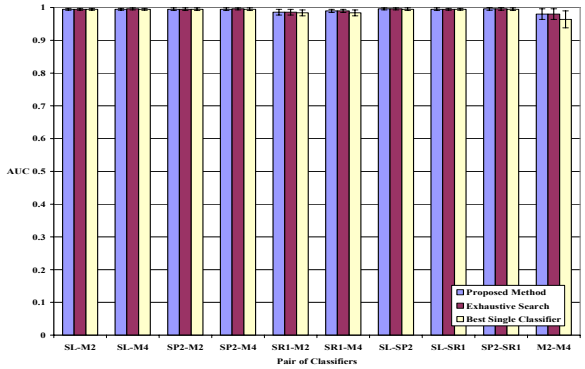

Figure 3. Results on Breast Cancer Wisconsin dataset in terms of AUC.

In the former case, one of the two dichotomizers is useless (as said at the end of the previous section) because it cannot recover any error made by the other classifier. In the second situation, 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. 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.

4. 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.

Acknowledgements

This work has been partially supported by MIUR (Italian Ministry of University and Research) under PRIN 2003 project *A system for computer aided analysis and remote access of mammographic images for early diagnosis of breast cancer*.

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*.

Flake, G. W., & Pearlmuter, B. A., (2000). Differentiating functions of the jacobian with respect to the weights. In S. A. Solla, T. K. Leen, & K. Müller (Ed.), *Advances in Neural Information Processing Systems, vol. 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, vol. 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.

Joachims, T., (1999). Making large-scale SVM learning practical. In B. Schölkopf, C.J.C. Burges, A.J. Smola, (Ed.), *Advances in Kernel Methods - Support Vector Learning*. The MIT Press.

Ling C. X., Huang J., & Zhang H., (2003). AUC: a statistically consistent and more discriminating measure than accuracy. *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 519-526).

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. *Proceedings of the International Conference on Machine Learning* (pp. 445-453). Morgan Kaufmann.

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

Tortorella, F., (2005). A ROC-based reject rule for dichotomizers. *Pattern Recognition Letters, 26*. 167-180.

Yan, L., Dodier, R., Mozer, M.C., & Wolniewicz, R., (2003). Optimizing classifier performance via the Wilcoxon-Mann-Whitney statistic. *Proceedings of the International Conference on Machine Learning* (pp. 848-855).