An ensemble classifier learning approach to ROC optimization

Sheng Gao*, Chin-Hui Lee‡ and Joo Hwee Lim*

*Institute for Infocomm Research, Singapore 119613
‡School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
{gaosheng, joohwee}@i2r.a-star.edu.sg   chl@ece.gatech.edu

ABSTRACT

An ensemble learning framework is proposed to optimize the receiver operating characteristic (ROC) curve corresponding to a given classifier. The proposed ensemble maximal figure-of-merit (E-MFoM) learning framework meets four key requirements desirable for ROC optimization, namely: (1) each classifier in the ensemble can be learned with any specified performance metric for any given classifier design; (2) such a classifier is discriminative in nature and attempts to optimize a particular operating point on the ROC curve of the classifier; (3) an ensemble approximation to the overall behavior of the ROC curve can be established by sampling a set of operating points; and (4) ensemble decision rules can be formulated by grouping these sampled classifiers with a uniform scoring function. We evaluate the proposed framework using 3 testing databases, the Reuters and two UCI sets. Our experimental results clearly show that E-MFoM learning outperforms the state-of-the-art algorithms using Wilcoxon-Mann-Whitney rank statistics.

1. Introduction

In recent years many machine learning algorithms have been proposed and developed to handle a wide range of real-world classification and verification applications. One crucial and practical issue is how to evaluate these algorithms and select an optimal solution that meets a wide variety of operating needs for a particular task. In some cases, such requirements are not clearly specified at the designing stage. For example, authorizing a withdraw of a large amount of fund from a user bank account often demands a very low false negative rate, while taking a small amount out the account requires a low false positive rate to minimize user inconvenience. These two sets of conflicting goals need to be simultaneously satisfied in order for the authentication system to work for a large population of diverse users in real-world environments. Most conventional classifiers are designed with a pre-set performance objective that cannot be easily adjusted, and work only in limited conditions. Recently, there is a growing interest to adopt the ROC as the metric for performance evaluation and classifier design. In [1, 3, 5, 7], the properties of the ROC measures and their relationship with the commonly used metrics, such as classification error, precision, recall, $F_1$ measure, Gini index, and information gain, have been discussed and analyzed in great detail. They clearly show that the ROC metric is a good candidate to measure the overall behavior of a given classifier. One summary measure is the area under an ROC curve (AUC), which characterizes the correct rank statistics and has been shown equal to the normalized Wilcoxon-Mann-Whitney (WMW) statistic [4, 10, 14]. AUC maximization has thus been proposed for learning a classifier with optimal ROC. Since the ROC curve is considered as an outcome of a collection of all possible decision rules, some learning algorithms have been developed to select a sub-set of rules for ROC optimization. These multiple rules are then ordered based on some convex hull algorithms [6, 13]. Since the rules are selected to construct a convex ROC curve, the AUC over the training set is optimized. These methods illustrate two key features in ROC optimization, namely: (1) optimizing a preferred metric directly, e.g. AUC; and (2) learning multiple decision rules to approximate the overall ROC curve. This motivates the discriminative ensemble learning.

In this paper we propose an ensemble maximal figure-of-merit (E-MFoM) learning framework that meets four critical requirements desirable for ROC optimization, namely: (1) each classifier in the ensemble can be learned with any specified performance metric for any given classifier design; (2) such a classifier is discriminative in nature and attempts to optimize a particular operating point on the ROC curve of the classifier; (3) an ensemble approximation to the ROC curve can be established by sampling a collection of operating points; and (4) a set of ensemble decision rules can be formulated by grouping these sampled classifiers with a uniform scoring function that is capable of performing multi-category comparisons.

2. ROC optimization via AUC maximization

Most conventional classifier learning algorithms are designed to optimize a single performance measure. This corresponds to one operating point on the ROC curve. Since the desired operating point is usually not defined at the time of the design, the learned classifiers are often non-optimal during operation. Therefore it is desirable to optimize the classifier over a wide range of potential operating conditions. Direct ROC optimization is non-trivial. One solution is to define a summary metric to simulate the overall ROC measure. The AUC is one such metric, which is equal to the normalized WMW statistic [4, 10, 14], defined as:

$$U = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} I(x_i, y_j)$$

where $I(x_i, y_j)$ is 1 if $x_i > y_j$ and zero otherwise, $x_i$ and $y_j$ are the scores from the classifier for the $i$-th of the $M$ positive samples, and the $j$-th of the $N$ negative examples in the training set, respectively. A classifier can then be trained by maximizing Eq. (1). In general some smoothing methods, e.g. a sigmoid function in Eq. (2) or a differentiable function in Eq. (3), are used to approximate the discrete ranking count, $I(x_i, y_j)$ [14],

$$S(x_i, y_j) = \frac{1}{1 + e^{\alpha(x_i - y_j)}}$$

$$R_i(x_i, y_j) = \begin{cases} \left(-\left(x_i - y_j\right)\right)^\beta & : x_i - y_j < \gamma \\ 0 & : \text{Others} \\ \end{cases}$$
Similarly, the area above the ROC curve (AAC=1-AUC) can be defined as a complement of the AUC. The only difference from the AUC is \(I(x_i, y_j) = 1, \text{if } x_i \leq y_j\) for AAC.

### 3. Ensemble MFoM learning

AUC maximization is only one way to optimize the overall ROC behavior. It is sometimes desirable to encompass multiple decision rules to improve AUC flexibility. We would also like to have a discriminative learning framework that accommodates other performance metrics for any given classifiers, gives high performance, and is robust over a wide range of conditions. In the following we propose an ensemble learning framework for ROC optimization that includes the above AUC maximization formulation as a special case. In Section 3.1 we discuss smooth embedding of error counts for defining error-centric objectives. The AUC defined in Eq. (1) is such a metric. In Section 3.2 we describe MFoM discriminative classifier learning. In Section 3.3 we introduce ensemble learning by sampling the ROC curve. In Section 3.4 two ensemble decision rules are formulated by grouping these sampled classifiers. Finally in Section 3.5 the learning algorithm for E-MFoM classifier is briefly discussed.

#### 3.1 Definition of evaluation objective functions

To achieve a good performance over a wide variety of operating conditions with different performance requirements, it is often important to be able to directly optimize over the desired metrics when designing a classifier. Denote true positive, false positive, and false negative rates as \(TP, FP, FN\), respectively, and the positive and negative class as \(C_+\) and \(C_-\). It is clear that these measures and their related metrics are discrete error counts and cannot be optimized directly because they are non-differentiable functions of classifier parameters. By choosing an appropriate loss function, \(l(X; s)\), defined for a sample feature vector, \(X\), and classifier parameter set, \(s\), that approximates a 0-1 step function at the origin, and summing over all samples, \(X_n\), in a given training set, \(T\); these measures can be approximated as:

\[
TP(T; s) \approx \frac{1}{|C_+|} \sum_{i \in C_+} \left( 1 - l(X_i; s) \right) l(X_i \in C_+) \tag{4}
\]

\[
FP(T; s) \approx \frac{1}{|C_-|} \sum_{i \in C_-} l(X_i; s) l(X_i \in C_-) \tag{5}
\]

\[
FN(T; s) \approx \frac{1}{|C_+|} \sum_{i \in C_+} l(X_i; s) l(X_i \in C_-) \tag{6}
\]

where \(|C_+|\) and \(|C_-|\) are the sizes of the training samples for the positive and negative, \(l(\cdot)\) is the indicator function of a logical expression, \(E\). A sigmoid function in Eq. (2) is often used to approximate the 0-1 loss function. For binary classification, its decision rule, \(D(\cdot)\), is:

\[
D(X; s) = \begin{cases} \text{Positive, if } g(X; s) > t, \\ \text{Negative, Otherwise} \end{cases} \tag{7}
\]

where \(g(\cdot)\) is a discriminant (score) function and \(t\) is a decision threshold. It turns out that the WMW statistic defined in Eq. (1) can also be expressed as an overall empirical loss rate as:

\[
L(T; s) = \frac{1}{|C|} \sum_{i \in C} \sum_{a \in \{0,1\}} \left( 1 - l(s_i; s_n) \right) l(X_i \in C) l(X_n \in C_n) \tag{8}
\]

where \(s_i\) and \(s_n\) are the scores for the positive and negative class, and \(l(s_i; s_n) = 1 - S(s_i, s_n)\) is a loss function in case of \(t=s_m\). This brief analysis shows that the AUC measure is a special case of a family of metrics that can be optimized directly.

#### 3.2 Discriminative single point approximation

Discriminative learning has been well studied. Classifiers based on support vector machines have been shown to give very good results for text categorization [11]. Gao, et al. proposed a maximal figure-of-merit (MfoM) learning approach attempting to maximize the separation between the classifier models corresponding to the target and all competing classes [8, 9]. Different from other discriminative learning algorithms that are designed to optimize a specific metric, MfoM learning can be applied to optimize any preferred performance measures for any given classifier. This corresponds to single point approximation (SPA) that optimizes a single operating point on the corresponding ROC curve for the learned classifier. It has been shown that MfoM is robust and less sensitive to data variation, especially when available training samples are limited [9].

The key to discriminative learning lies in the definition of a separation function that quantifies the distance between the target class and all competing classes. For binary classification, a misclassification measure is derived from Eq. (7) as:

\[
d(T; s) = -g(T; s) + t. \tag{9}
\]

The overall loss function, \(L(T; s)\), a function of \(d(T; s)\), for a single ROC operating point, can be approximated by a weighted sums of \(TP(T; s)\) and \(FP(T; s)\) in Eqs. (4) and (5) to be discussed next.

#### 3.3 Ensemble learning by sampling ROC curves

Given a specific operating point, \(q\), on the ROC curve, a pair of true positive and false positive rates, \((TP_q, FP_q)\), is also determined. In general, increasing the true positive rate will also increase the false positive rate. For ensemble learning an intuitive way is to first select a set of operating points of interest, train a classifier to optimize each given operating point, and then design a decision rule over the ensemble of multiple classifiers. Let \(\Lambda_q\) be the classifier parameter set corresponding to the \(q\)-th operating point. Assume that the overall loss function, which is also an error function, referred to as ERR later, at this point can be defined as a weighted sum as:

\[
L_q(T; \Lambda_q) = \rho_q \cdot TP_q(T; \Lambda_q) + (1 - \rho_q) \cdot FN_q(T; \Lambda_q), \tag{10}
\]

where \(\rho_q\) is a weight constant \((0 \leq \rho_q \leq 1)\) determining the contribution of the false positive rate to the overall loss. For various performance requirements in operation we can choose different weighting constants. We can also sample the ROC curve on a representative set of operating points to collectively approximate the overall ROC behavior. An extreme case is that the weight can be set to zero if the false positive is not a major concern while the false reject must be avoided, like the case of withdrawing a large amount of fund from an account in user authentication. For constrained ROC optimization we sample only the operating points corresponding to the partial segments of interest. Therefore the proposed sampling method provides a flexible way to embed some prior knowledge into classifier design about the given real-world problem. For example, we can choose \(Q\) equally spacing weights from zero to one to approximate a set of \(Q\) operating points on the ROC curve. If \(Q\) is large enough we can collectively approximate the ROC Curve using multiple point approximation (MPA).
Since each choice of a weight constant corresponds to a different objective function as defined in Eq. (10), a learning framework that accommodates any evaluation measures for any classifiers is desired. Furthermore we would like to have a discriminative learning framework that not only optimizes performance at each operating point but also is robust to data variations and changing class distributions. We believe the MFoM classifier learning framework [8, 9] is ideal to meet these needs. Once we have a collection of MFoM-learned classifiers, we need an ensemble decision strategy to make final classification choices.

3.4 Ensemble decision rules

Given an ensemble of classifiers, we have a number of flexible decision strategies to be considered in the evaluation stage by taking advantage of multiple decisions together. The first is to select the classifier with the highest scores, and make classifier decision accordingly. It is called a MPA-Best decision defined as:

$$D(X; \Lambda) = \begin{cases} 
\text{Positive, if } \max_t g(X|\Lambda_t) - t > 0 \\
\text{Negative, Otherwise}
\end{cases}$$

(11)

The second is to use the averaging confidence score for decision. We called it MPA-Avg, which is defined as:

$$D(X; \Lambda) = \begin{cases} 
\text{Positive, if } \sum_q Q^t g(X|\Lambda_q) - t > 0 \\
\text{Negative, Otherwise}
\end{cases}$$

(12)

The third is to use a single classifier for decision if the weight ratio is known in advance. In some applications, the user may specify the ratio, e.g. for the evaluation of speaker recognition [12]. If the desired classifier for a specified weight ratio is not already in the ensemble, we can select the classifier with the weight ratio closest to the specified value, or build a new classifier using some interpolation methods, e.g. classifier parameter interpolation using the two nearest classifiers. Of course, some classifiers such as those based on linear discriminant functions are easier to interpolate than others.

3.5 Learning ensemble MFoM classifiers

The learning criterion, $L_g(T; \Lambda_q)$, for SPA is defined in Eq. (10). For E-MFoM we can weigh and combine them into an overall optimization objective function as:

$$L(T; \Lambda) = \sum_q a_q \cdot L_g(T; \Lambda_q),$$

(13)

where $a_q$ is the weights to measure the significance of the q-th operating point. For simplicity, here the uniform weight is used and the classifier for each operating point is trained independently using a GPD algorithm [8, 9].

4. Experimental results and analysis

We analyze the proposed E-MFoM learning approach for ROC optimization and compare it with the AUC-based method [14]. In all experiments the classifier uses a linear discriminant function. 3 datasets were used. The first is Reuters-21578\(^1\) (ModApte version) for text categorization. 2 topics, “Acq” and “Job”, are chosen. The former has enough training samples while the latter has small size. The 1613-dimension LSI-based feature is used for document representation [8, 9]. The other 2 datasets are selected from the UCI Repository [2]. One is “SPECTF” with divided training and test sets, and another is “Credit Approval”, exclusive of the samples with missing values, and was split into 10-folds, one for training and the other nine for evaluation in each run. Normalized attributes are used as features. A detailed description of the 4 classes is shown in Table 1. The AAC (AAC=100-AUC, shown as the percentage in Tables 2 and 3) is used for evaluation because the comparison is characterized relatively better using AUC than AUC.

<table>
<thead>
<tr>
<th>Class</th>
<th>Dim.</th>
<th>Train set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acq</td>
<td>1613</td>
<td>1,650</td>
<td>6,120</td>
</tr>
<tr>
<td>Job</td>
<td>1613</td>
<td>46</td>
<td>7,724</td>
</tr>
<tr>
<td>SPECTF</td>
<td>44</td>
<td>40</td>
<td>40</td>
</tr>
</tbody>
</table>

Since we do not use any prior knowledge about the datasets except for their labels, uniform sampling of the ROC curve was adopted. Currently 9 cost ratio points were uniformly sampled in [0, 1]. They were from 0.1 to 0.9 in an equal increment of 0.1. Two extreme points, 0 and 1, were excluded. Then the set of ensemble classifiers were learned using the E-MFoM algorithm. In evaluation, two ensemble decision rules, MPA-Best and MPA-Avg, discussed in Section 3.4 were used for comparison.

4.1 E-MFoM versus SPA for ROC optimization

First the MPA and SPA based E-MFoM learning methods were compared. SPA E-MFoM is similar to binary MFoM in [8]. The optimized metrics are the weighted misclassification error defined in Eq. (10), denoted by ERR, and $F_1$, denoted by $F_1$. The results with these two metrics in optimization and three different decision rules (i.e. SPA, MPA-Best and MPA-Avg), are listed in Table 2. For the “Credit” class its mean AAC values and standard deviations (shown in parentheses under the means) over 10-fold evaluations is given. For each metric, the best result among 3 decision rules is highlighted in bold font.

<table>
<thead>
<tr>
<th>Class</th>
<th>F1 (%)</th>
<th>ERR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SPA</td>
<td>MPA-Best</td>
</tr>
<tr>
<td>Acq</td>
<td>0.33</td>
<td>0.31</td>
</tr>
<tr>
<td>Job</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>SPECTF</td>
<td>18.80</td>
<td>17.60</td>
</tr>
<tr>
<td>Credit</td>
<td>10.18</td>
<td>10.35</td>
</tr>
</tbody>
</table>

In both performance measures it is clearly shown that E-MFoM outperforms binary MFoM (SPA column), where only a single operating point optimized. For different decision rules (MPA-Best and MPA-Avg columns), the decision using the best confidence score (MPA-Best) achieves the best AACs in almost all cases for the ERR metric. Nonetheless from the comparisons between the $F_1$ and the ERR columns, only little difference in performance is observed in most classes for $F_1$. In Fig. 1(a) we plot the ROC curves for the SPECTF data with the three classifiers (in ERR metric). They clearly show that MPA-Best outperforms MPA-Avg and SPA at most of the operating points.

---

\(^1\)http://www.research.att.com/~lewis/reuters21578.html
Comparing two E-MFoM decisions with the AUC algorithms

Figure 1 (a) ROC curves for E-MFoM using MPA and SPA (b) ROC curves for E-MFoM learning again

4.2 E-MFoM versus AUC-Based Optimization

Next we compare E-MFoM learning with the AUC-based ROC optimization algorithms [14] we implemented. In Table 3 the two sets of results for the AUC-based algorithms are shown in the AUC column on the left, under which the results for the Sigmoid (see Eq. (2)) and Differential (see Eq. (3)) functions are listed in the columns labeled “Sig” and “Dif”, respectively. The results for E-MFoM learning with the ERR metric using the MPA-Best decision rule are listed in the ERR column on the right. Again the means and standard deviations (shown in parentheses under the means) over 10-fold evaluation for the “Credit” class. Since all AAC values are relatively small in this case, the relative improvements, listed in the rightmost column of Table 3, with positive signs indicating AAC reduction, are calculated based on the difference between the better results of the two AUC-based algorithms and the MPA-Best results. It is noted that E-MFoM learning again gives better results than the AUC-based methods in all 4 classes.

Figure 2 Curves with MPA-Best (heavy dash), sampling 9 SPA curves to form a convex hull (solid) and WMW-Dif (light dash)–zooming into the upper left corner of the ROC curve in Fig. 1(b)

5. Conclusion

An ensemble MFoM learning framework is proposed for ROC optimization. The E-MFoM learning integrates statistical sampling technique into maximal figure-of-merit learning. It is a unified ensemble learning framework, which not only enhances both robustness and discrimination power of classifier corresponding to each operating point, but also optimizes the overall ROC curve through ensemble MFoM learning and decision rules. We believe the proposed E-MFoM learning framework meets all four critical requirements desirable for ROC optimization, namely: (1) each classifier in the ensemble can be learned with any specified performance metric for any given classifier design; (2) such a classifier is discriminative in nature and attempts to optimize a particular operating point on the ROC curve of the classifier; (3) an ensemble approximation to the ROC curve can be established by sampling a collection of operating points; and (4) a set of ensemble decision rules can be formulated by grouping these sampled classifiers with a uniform scoring function.

The experimental results show that the E-MFoM learning achieves better results than those obtained with binary MFoM optimized on a single operating point. It also outperforms state-of-the-art ROC learning algorithms that optimize AUC via the WMW rank statistics.

6. References