Optimizing Classifier Performance via an Approximation to the Wilcoxon-Mann-Whitney Statistic

Lian Yan LIAN_YAN@CSGSYSTEMS.COM
Robert Dodier ROBERT_DODIER@CSGSYSTEMS.COM
CSG Systems, Inc., 11080 Circle Point Rd., Westminster, CO 80020

Michael C. Mozer MOZER@CS.COLORADO.EDU
Department of Computer Science, University of Colorado at Boulder, Boulder, CO 80309

Richard Wolniewicz RICHARD_WOLNIEWICZ@CSGSYSTEMS.COM
CSG Systems, Inc., 11080 Circle Point Rd., Westminster, CO 80020

Abstract

When the goal is to achieve the best correct classification rate, cross entropy and mean squared error are typical cost functions used to optimize classifier performance. However, for many real-world classification problems, the ROC curve is a more meaningful performance measure. We demonstrate that minimizing cross entropy or mean squared error does not necessarily maximize the area under the ROC curve (AUC). We then consider alternative objective functions for training a classifier to maximize the AUC directly. We propose an objective function that is an approximation to the Wilcoxon-Mann-Whitney statistic, which is equivalent to the AUC. The proposed objective function is differentiable, so gradient-based methods can be used to train the classifier. We apply the new objective function to real-world customer behavior prediction problems for a wireless service provider and a cable service provider, and achieve reliable improvements in the ROC curve.

1. Introduction

For many real-world classification problems, the typical classification accuracy (correct classification rate) is not a sufficient metric for classifier performance (Provost et al., 1998). Consider the problem of predicting churn, that is, which customers will switch from one service provider to another. In the telecommunications industry, typical monthly churn rates are in the neighborhood of 2% (Mozer et al., 2000). The trivial solution of labeling all customers as nonchurners yields a 98% correct classification rate. In general, classification accuracy can be misleading for a two-class problem when the class distribution is highly imbalanced. To elaborate, consider the simple example of a data set consisting of 10 positive and 90 negative samples. Suppose the classifier outputs are binary, and that there are 8 errors. Regardless of which samples are misclassified, the classification accuracy will be 92%. However, consider two different cases leading to 8 total errors. In the case where all 8 errors are among the negative samples, the *false positive rate* (rate of negative samples being labeled as positive) is 9% and the *false negative rate* (rate of positive samples being labeled as negative) is 0%. In the case where all 8 errors are among the positive samples, the false positive rate will be 0% but the false negative rate will be 80%. The two cases yield quite different errors yet would be indistinguishable using the measure of classification accuracy.

An alternative means of evaluating classifier performance that does not confound the false-positive and false-negative error rates is the *Receiver Operating Characteristic* or *ROC* curve (Green & Swets, 1966). The ROC curve depicts the performance of a classifier by plotting the true positive rate against the false positive rate. Assuming that a classifier produces a continuous output (e.g., class posterior probabilities), then the output must be thresholded to label each sample as positive or negative. Thus, for each setting of the decision threshold, a true-positive and a false-positive rate is obtained. By varying the decision threshold over a range from 0 to 1, the ROC curve is produced. The upper graph in Figure 4 includes three ROC curves with the true and false positive rates on the y- and x-axes, respectively. The more bowed the curve is to-

Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

ward the upper left corner, the better is the classifier's ability to discriminate between the two classes. Thus, area under the ROC curve (AUC) is a general, robust measure of classifier discrimination performance, regardless of the decision threshold, which may be unknown, changeable over time, or might vary depending on how the classifier will be used.

It seems that the AUC is not easy to compute. However, note that the ROC curve of a finite set of samples is based on a step function, and it can be shown that the AUC is exactly equal to the normalized Wilcoxon-Mann-Whitney (WMW) statistic (Mann & Whitney, 1947) in the form:

$$U = \frac{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} I(x_i, y_j)}{mn}, \tag{1}$$

where

$$I(x_i, y_j) = \begin{cases} 1 & : \quad x_i > y_j \\ 0 & : \quad \text{otherwise} \end{cases}, \tag{2}$$

is based on pairwise comparisons between a sample x_i, $i = 0, \ldots, m-1$, of random variable X and the sample y_j, $j = 0, \ldots, n-1$, of random variable Y. If we identify $\{x_0, x_1, \ldots, x_{m-1}\}$ as the classifier outputs for m positive examples, and $\{y_0, y_1, \ldots, y_{n-1}\}$ as the classifier outputs for n negative examples, we obtain the AUC for our classifier via Eq. 1.[1] The expression of U can be further simplified if the $\{x_0, x_1, \ldots, x_{m-1}\}$ and $\{y_0, y_1, \ldots, y_{n-1}\}$ are merged and sorted in ascending order, yielding $U = \frac{1}{mn}(\sum_{i=0}^{m-1} r_i - \frac{m(m-1)}{2})$, where $r_i \in \{0, 1, \ldots, m+n-1\}, i = 0, \ldots, m-1$, is the rank of x_i, and $\sum_{i=0}^{m-1} r_i$ is the Wilcoxon's Rank Sum statistic (Wilcoxon, 1945). The statistic U is an estimator of $P[X > Y]$, but we only emphasize the equivalence between the WMW statistic and the AUC. The independence and other distribution requirements to use U as a test statistic are irrelevant here.

[1] We assume that the classifier produces high outputs for positive classifications and low outputs for negative classifications. We derive the equivalence between the WMW statistic and the AUC by expressing the AUC as

$$\begin{aligned} &\sum_{i=1}^{m-1} \frac{m-i}{m} \cdot \frac{\sum_{j=0}^{n-1} I(x_i, y_j) - \sum_{j=0}^{n-1} I(x_{i-1}, y_j)}{n} \\ &\qquad + \frac{\sum_{j=0}^{n-1} I(x_0, y_j)}{n} \\ &= \sum_{i=0}^{m-1} \frac{m-i}{m} \frac{\sum_{j=0}^{n-1} I(x_i, y_j)}{n} \\ &\qquad - \sum_{i=0}^{m-2} \frac{m-i-1}{m} \frac{\sum_{j=0}^{n-1} I(x_i, y_j)}{n}, \end{aligned}$$

where we assume $x_0 < x_1 < \ldots < x_{m-1}$.

The ROC curve or the AUC has been utilized extensively to assess classifier performance (e.g., Hand, 2001; Provost & Fawcett, 1998; Weiss & Provost, 2001), but how to optimize the ROC curve or the AUC has been rarely addressed. If the goal is to achieve the best correct classification rate, classifiers such as neural networks are typically trained by minimizing a mean squared error (MSE) or the cross-entropy (CE) objective function, or by maximizing the likelihood function.[2] However, as we will show in Section 2, optimizing classification accuracy by minimizing the MSE or CE cannot guarantee maximization of the WMW statistic or the AUC. Thus, if the AUC is the quantity of concern, it would be better to train a classifier to directly optimize the AUC. Unfortunately, the AUC itself, even in the equivalent form of WMW statistic, is nondifferentiable and cannot be optimized by gradient-based methods. In Verrelst et al. (1998), the authors recognized the limitation of using the MSE to optimize the ROC curve, and proposed to numerically calculate the AUC with the trapezoidal integration rule, and to maximize the AUC by simulated annealing. However, the paper reported that the simulated annealing approach does not improve the results over the MSE criterion. In Mozer et al. (2001), several approaches were proposed to try to improve a specific point on the ROC curve. However, these approaches assume that the decision threshold and/or the misclassification costs are known. Here, the focus of our work is to improve the overall ROC curve, measured by the AUC. Caruana et al. (1996) recognized that MSE might overfit the model when only relative ranking rather than explicit discrimination is required. They proposed an algorithm called *Rankprop* that tried to learn the rankings iteratively. However, the convergence properties of this algorithm were questionable, and we have experimentally found that it is difficult to obtain convergence. In the information retrieval field, some related work exists on optimizing ranks, (e.g., Bartell et al., 1994; Vogt & Cottrell, 1998). Bartell et al. use a differentiable function based on the Guttman's Point Alienation statistic to measure the correlation between the target and estimated ranking. Vogt and Cottrell showed that this measure can be seen as a scaled version of $d' = (\bar{x} - \bar{y})/\sigma_y$, where $\bar{x}$ is the mean of the x_i and $\bar{y}$ and σ_y are the mean and standard deviation of the y_j. We have explored using d' and its several variations for training neural network classifiers, and have found that they did not improve the AUC in our real-world data sets (Rattenbury & Yan, 2001). Finally, one would simply weight positive and nega-

[2] For a two-class problem, maximizing likelihood is equivalent to minimizing cross entropy.

tive samples differently during training, especially for imbalanced data sets, to improve the AUC. Weiss & Provost (2001) report that the best weight (class distribution) needs to be experimentally determined for each data set. Later we will compare the best ROC curve (in terms of the AUC) obtained by trying several weights for the MSE training with the proposed new algorithm's results.

2. Classification accuracy and the AUC

The AUC and classification accuracy criteria can be dissociated from one another, i.e., the AUC is sensitive to quite different aspects of a classifier than the classification accuracy. We argue this point with a two-alternative classifier that is used to assign a "positive" or "negative" label to each sample, in the following manner. We assume that the classifier produces a continuous output, and is run on a data set consisting of m positive and n negative samples. The set of samples can be ranked by their classifier output, where the rank varies from 0, the rank of the lowest output, to $m+n-1$. Samples with identical output are assigned the average rank. Given some threshold θ, such that all samples with rank $\geq \theta$ are labeled "positive" and other samples are labeled "negative," one can determine which samples are labeled correctly and which are errors. If k is the total number of errors in the data set, the classification accuracy is $1-k/(m+n)$. In contrast, the WMW statistic depends not only on k but also on the *ranks* of the samples which are classified incorrectly. Assuming $k < n$, for each θ, there are $\sum_{x=0}^{\min\{k,m\}} \binom{m+k-2x}{k-x}\binom{n-k+2x}{x}$ possible values of the WMW statistic depending on the ranks of the k misclassified samples. To simplify, assign 1 to all samples with rank $\geq \theta$ and 0 to others. Then, from Eq. 1, the mean and variance of this set of WMW statistic values, which associate with the same classification accuracy $1-k/(m+n)$, can be obtained:

$$\mu_u = \sum_{x=0}^{\min\{k,m\}} \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \left(1 - \frac{\frac{k-x}{n}+\frac{x}{m}}{2}\right), \tag{3}$$

$$\sigma_u^2 = \sum_{x=0}^{\min\{k,m\}} \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \left(1 - \frac{\frac{k-x}{n}+\frac{x}{m}}{2} - \mu_w\right)^2. \tag{4}$$

Figure 1 (upper) shows an example of the relationship between σ_u and the positive-example proportion $m/(m+n)$ when classification accuracy is fixed. We can see that the WMW statistic becomes increasingly dissociated from classification accuracy when the class distribution becomes more imbalanced. Generally, minimizing the MSE or CE for optimizing classification accuracy cannot guarantee maximization of the WMW statistic or the AUC, especially for imbalanced data sets. Figure 1 (lower) demonstrates that the WMW statistic does not necessarily increase with MSE's decrease when the MSE is minimized during training.

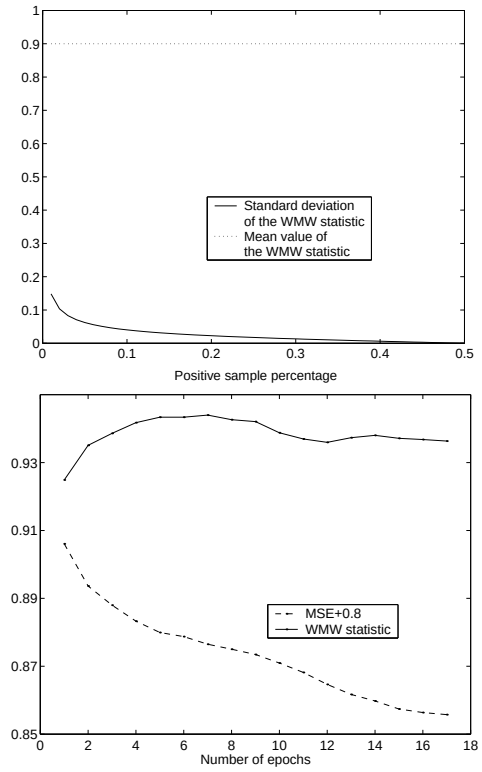

Figure 1. Upper: σ_u increases with the decrease of $m/(m+n)$, when the classification accuracy is fixed at 0.8 ($m+n=100$ and $k=10$). Lower: The WMW statistic and MSE variation during training for the UC Irvine *Credit Approval* data.

3. Direct optimization of the AUC

Here we propose an alternative objective function for training classifiers to directly optimize the ROC curve. The proposed objective function can be applied to any parametric classifier. For any such classifier, one can optimize the objective function with respect to the parameters using gradient based methods. In the results below, we apply the proposed objective to a typical multilayer perceptron (MLP) with softmax outputs, with a single hidden layer and direct connection between the input and output layers.

3.1. New objective functions

To directly optimize the AUC, we can try to maximize the WMW statistic in Eq. 1. However, the function $I(x_i, y_j)$ in Eq. 2 is nondifferentiable. A differentiable approximation to Eq. 2 is the sigmoid function

$$S(x_i, y_j) = \frac{1}{1 + e^{-\beta(x_i - y_j)}}, \tag{5}$$

where $\beta > 0$. Then, one can try to maximize

$$U_S = \frac{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} S(x_i, y_j)}{mn}. \tag{6}$$

However, as shown in Figure 2, $S(x_i, y_j)$ with a small β, e.g., $\beta = 2$, softens $I(x_i, y_j)$ too much when $0 \leq x_i \leq 1$ and $0 \leq y_j \leq 1$. A large β would make $S(x_i, y_j)$ close to $I(x_i, y_j)$, but this brings in numerical problems during optimization because of steep gradients. For the large data sets in our experiments, we have seen that the optimization could not proceed successfully with $\beta > 2$. Alternatively, we use the differentiable function

$$R_1(x_i, y_j) = \begin{cases} (-(x_i - y_j - \gamma))^p & : \quad x_i - y_j < \gamma \\ 0 & : \quad \text{otherwise} \end{cases}, \tag{7}$$

where $0 < \gamma \leq 1$ and $p > 1$, and train a classifier by *minimizing* the objective

$$U_R = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} R_1(x_i, y_j). \tag{8}$$

$R_1(x_i, y_j)$ is shown in Figure 2 with $S(x_i, y_j)$ and as well another function

$$R_2(x_i, y_j) = \begin{cases} (x_i - y_j - \gamma)^p & : \quad x_i - y_j > \gamma \\ 0 & : \quad \text{otherwise} \end{cases}, \tag{9}$$

where $0 < \gamma \leq 1$ and $p > 1$. $R_2(x_i, y_j)$ can be regarded as an approximation to $I(x_i, y_j)$, while $R_1(x_i, y_j)$ approximates $I(-x_i, -y_j)$. We have found that maximizing $\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} R_2(x_i, y_j)$ is ineffective to maximize the WMW statistic, because it makes the optimization focus on maximizing the difference between x_i and y_j rather than on moving more pairs of x_i and y_j to satisfy $x_i - y_j > \gamma$, which is the key notion in the WMW statistic using $I(x_i, y_j)$. Instead, during the process of minimizing U_R, when a positive sample has a higher output than a negative sample by a margin γ, this pair of samples will not contribute to the objective function. Essentially, the influence of the training samples is adaptively adjusted according to the pairwise comparisons during training. Note that the positive margin γ in U_R is needed for better generalization performance.

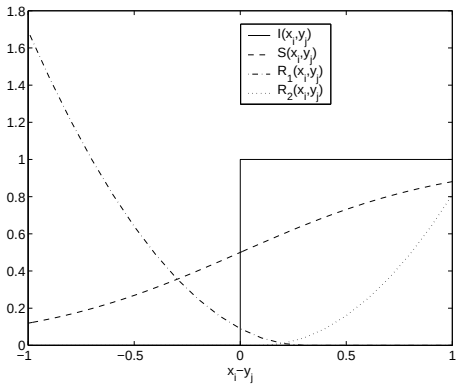

Figure 2. $S(x_i, y_j)$, $R_1(x_i, y_j)$, and $R_2(x_i, y_j)$ compare with $I(x_i, y_j)$, where $\beta = 2$ in $S(x_i, y_j)$, $\gamma = 0.3$ and $p = 2$ in $R_1(x_i, y_j)$, and $\gamma = 0.1$ and $p = 2$ in $R_2(x_i, y_j)$.

3.2. Discussion of U_R and U_S

First, we look at the training process of the different objective functions over the data set *SPECTF heart* in the UC Irvine machine learning repository, which has predefined training and test sets. Figure 3 shows the variation of the WMW statistic U in Eq. 1 as a function of the number of epochs of training, for both the training and test sets, for optimizing four objective functions, where $\gamma = 0.1$ and $p = 2$ in $R_1(x_i, y_j)$, and $\beta = 2$ in $S(x_i, y_j)$. We can see that minimizing the objective function U_R based on $R_1(x_i, y_j)$ obtains the largest WMW statistic for both training and test. Although minimizing CE achieves the same WMW statistic for the training set, it obtains the lowest WMW statistic for the test data. Overfitting is observed for both MSE and CE based training, but $R_1(x_i, y_j)$ based training does not incur obvious overfitting. Moreover, even if early stopping can be applied so that the MSE or CE training can stop when the value of the WMW statistic over the test set reaches the maximum, this maximum value is still smaller than the final WMW statistic obtained by minimizing U_R. We will see that this is true in our other experiments as well. It is also obvious in Figure 3 that maximizing U_S based on $S(x_i, y_j)$ is inferior to minimizing U_R.

The upper panel of Figure 4 presents the ROC curves for the UC Irvine *Credit Approval* data set, based on 10-fold cross validation. The figure shows that the objective function U_R generates a significantly better ROC curve than CE- and MSE-based training. Here, we chose $\gamma = 0.7$ and $p = 2$ for $R_1(x_i, y_j)$. A natural question is then how to choose γ and p. As shown in Figure 4 (lower), the WMW statistic U, i.e., the AUC, for the test data is quite insensitive to γ over a large range. In general, we can choose a value between 0.1 and 0.7 for γ. Also, we have found that $p = 2$ or 3

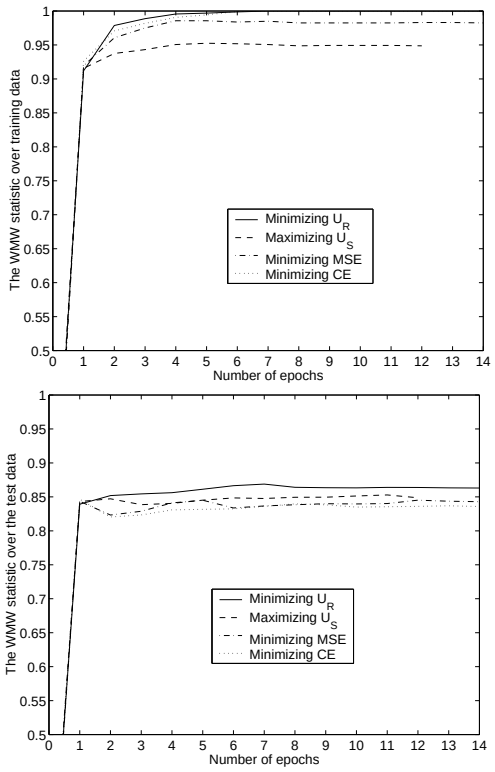

Figure 3. The WMW statistic variation over the training set (upper) and test set (lower) for the *SPECTF heart* data as a function of number of epochs of training.

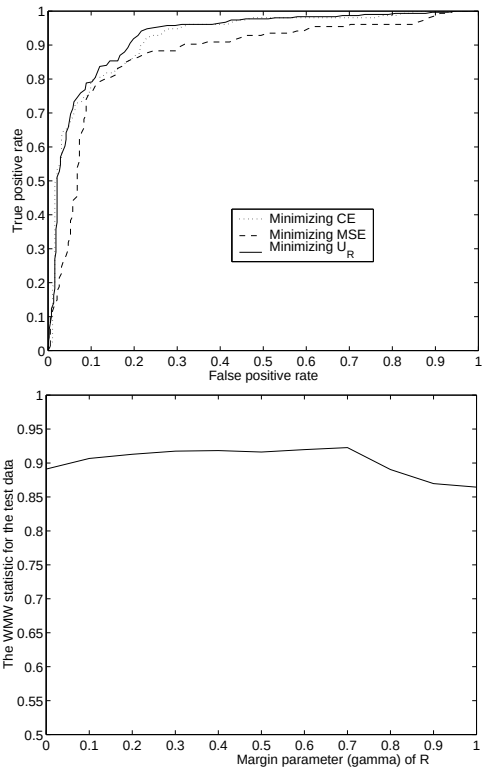

Figure 4. Upper: The ROC curves for 10-fold cross validation over *Credit Approval*. The results are insensitive to the number of hidden units, and these curves are based on 0 hidden unit. Lower: The WMW statistic over the test data at different γ's for *Credit Approval*. The test data is the union of the 10 out-of-sample subsets for the 10-fold cross validation.

achieves similar, and generally the best results.

4. Applications

4.1. Churn prediction

In the wireless telecommunications industry, it costs around five times as much to sign on a new subscriber as to retain an existing one. It is thus crucial to predict customer behavior and proactively build lasting customer relationships. We build neural network models to predict churn, using both static customer data such as demographics and application information, and time series data of historical usage, billing, and customer service. Up to one year's worth of past usage, billing, and customer service data can be used to predict which customers will likely churn within the next one or two months. For example, a model is trained over input data from June to September, with the class label (churn or nonchurn) given for October. The trained model is then used over data from July to October to make predictions for November. The training and test data extracted from different time windows make the problem even more difficult because of the data nonstationarity issue (Yan et al, 2001).

The results in this study are based on data sets from a major US wireless service provider. The number of customers in the training set is close to 70,000, and the test set from a forward-shifted time window has over 70,000 customers. Six months past usage, billing, and customer service data were available for both training and test sets. The task is to predict churn over a two month window, for which the average churn rate is roughly 6%. 134 raw features were extracted from a variety of data sources. After preprocessing of the raw data, a 55-dimensional feature vector was used to represent each customer. In Figure 5 (upper), we compare the ROC curves over the test data by different training methods. The model trained by minimizing the new objective function U_R in Eq. 8 achieves a better ROC curve than the models trained by MSE or CE. Moreover, by trying several weights for positive samples during the MSE-based training, we have found weighting all the positive samples by 2 (versus 1 for all negative samples) achieved the best ROC curve over the test set. Still, we can see in Figure 5 that this ROC curve, obtained by extensive search of weights,

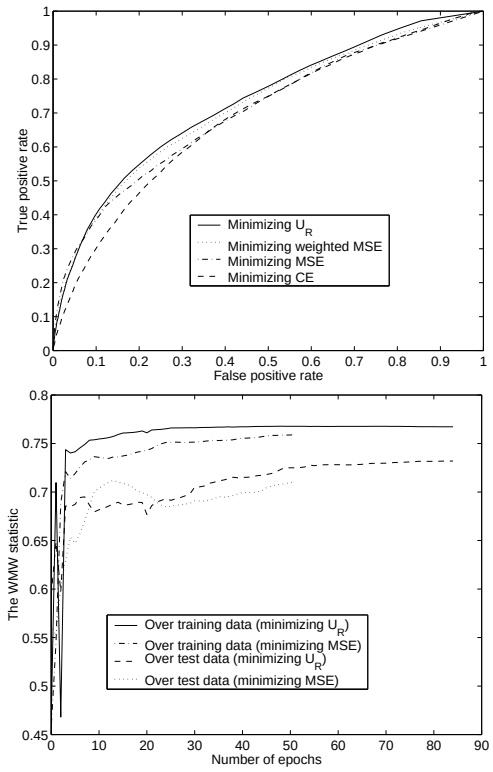

Figure 5. Upper: The ROC curves for churn prediction for a major US wireless service provider. The MLP structure has 5 hidden units. Lower: The WMW statistic variation over the training and test sets during training for the churn prediction model. Here, $\gamma = 0.3$ and $p = 2$ in $R_1(x_i, y_j)$.

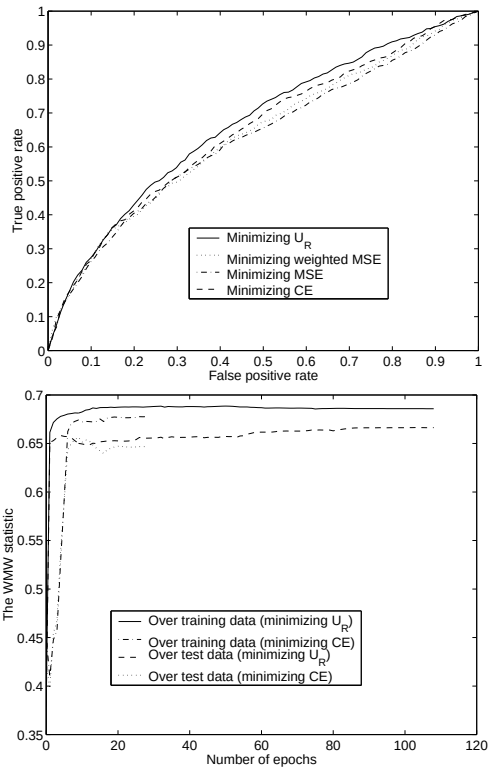

Figure 6. Upper: The ROC curves for predicting new service acceptance for a major US cable service provider. The MLP structure has 5 hidden units. Lower: The WMW statistic variation over the training and test sets during training for the acceptance prediction model. Here, $\gamma = 0.05$ and $p = 2$ in $R_1(x_i, y_j)$.

is overall worse than the one achieved by minimizing U_R. We also show the training process in Figure 5 (lower), which compares the variation of the WMW statistic U during training over both the training and test sets for the new objective function U_R and MSE. Similar to previous results, minimizing the new objective function obtains larger WMW statistic for both training and test data, and, unlike MSE-based training, it does not overfit the training data. Note again that even the largest value of the test set WMW statistic for MSE minimization (at epoch 13) is smaller than the final WMW statistic obtained by minimizing U_R.

4.2. Cross-sell acceptance prediction

In this application, we build models for an US cable service provider to predict which customers will likely accept a cross-sell offer of a particular new service. Accurate prediction of the acceptance can significantly save the service provider's campaign costs and dramatically improve the offer acceptance rate by targeting only those customers who will likely accept the new service. From the similar data sources to those of churn prediction, 135 raw data fields are extracted, and 41 transformed features are used as the neural network inputs. In this study, five months worth of past usage, billing, and customer service data are used to predict which customers will likely accept the new service within the next month. Training and test sets were extracted from different time windows, with the test data window forward-shifted by one month. For training, we have over 130,000 customers with a positive sample rate of only 0.5%. The test set consists of a similar number of customers with a positive sample rate of 0.7%. Figure 6 (upper) demonstrates again that the new objective function U_R achieves substantial improvement of the ROC curve over both the MSE and CE training. For this data set, weighting the positive samples by 10 for the MSE training obtained the best ROC curve over the test set, but it is still worse than the one achieved by minimizing R_U. Figure 6 (lower) shows the difference between the training processes of the new objective function U_R and the CE.

5. Extensions

5.1. Multi-membership classification

In some applications, samples are associated with multiple classes, e.g., people may have multiple diseases, and these are called multi-membership samples. We extend our work here to the following multi-membership classification problem. In addition to churn prediction, service providers would also like to know in advance why a subscriber wants to churn so that the customer service representative (CSR) can make appropriate offers and try to save the subscriber. Some subscribers may have multiple reasons, e.g., competition and billing problem. Also, the CSR can generally try more than one reason during the conversation with a subscriber. Thus, the conventional accuracy, just measuring if the predicted top reason is correct, is insufficient. Instead, this application actually requires ranking the reasons for each subscriber.

Assume there are L predefined reason categories, and subscriber k, $k = 0, \ldots, K-1$, is associated with m_k reasons, $0 < m_k \leq L$. An MLP structure with L outputs is trained over these K samples. Denote x_{ki}, $i = 0, \cdots, m_k - 1$, as the classifier output for subscriber k's associated reason i, and y_{kj}, $j = 0, \cdots, L - m_k - 1$, as the output for reason j which is not a stated reason of subscriber k. We then train the MLP model by minimizing the objective

$$U_m = \sum_{k=0}^{K-1} \sum_{i=0}^{m_k-1} \sum_{j=0}^{L-m_k-1} R_1(x_{ki}, y_{kj}). \qquad (10)$$

We apply this objective function to churn reason prediction for an US cable service provider. Churn reasons are predicted for potential churners in the next month based on the similar data sources mentioned in Section 4. The customer base includes more than 330,000 subscribers, but only about 3,100 subscribers with churn reasons can be used to train the model. More than 20% of the 3,100 subscribers are associated with more than one reason. There are 15 predefined reason categories, with the smallest category consisting of only 0.03% of the churners and the largest category consisting of 40.8% of the churners. Typically, reason prediction is a much more difficult problem than churn prediction because of available training set size, very noise data, and imbalanced, overlapping categories. Table 1 shows the 10-fold cross validation results for the new objective function U_m, the MSE, and just using the prior distribution, i.e., ordering the reasons for each subscriber according to the prior class distribution. P_1 is defined as the rate of the number of churners, whose top predicted reason is among the true reasons, to the total number of churners, and P_2 and P_3 are the rates of the number of churners, who have at least one of the true reasons among the top two or three predicted reasons, respectively, to the total number of churners. We can see that the new objective U_m outperforms the MSE for all the accuracy measures.

Method	P_1	P_2	P_3
Minimizing U_m	0.451	0.766	0.893
Minimizing MSE	0.439	0.734	0.882
Using the priors	0.408	0.731	0.828

Table 1. Accuracy measures of churn reason prediction for an US cable service provider. The number of hidden units of the MLP structure is chosen as 5 for both U_m- and MSE-based training.

5.2. Focusing on the lower-left part of the ROC curve

If the classifier produces high outputs for positive classifications and low outputs for negative classifications, the lower-left part of the ROC curve measures the performance over those samples with higher classifier outputs. This portion of the ROC curve is important because many applications restrict the actions, e.g., contacting customers, among the classifier outputs above a certain threshold or within a certain percentage in the top. One way to extend our algorithm to focusing on the lower-left part of the ROC curve is to map the classifier output s by the function

$$f(s) = \begin{cases} (s - \beta \cdot \mu_s)^\alpha & : \quad s > \beta \cdot \mu_s \\ 0 & : \quad \text{otherwise} \end{cases}, \qquad (11)$$

with $\mu_s = \frac{\sum_{i=0}^{m-1} x_i + \sum_{j=0}^{n-1} y_j}{m+n}$, the mean value of the classifier outputs. Here, $\alpha > 1$ but is close to 1, e.g., $\alpha = 1.1$, and $\beta \geq 1$ and is generally chosen as 1. Thus, this function maps the classifier outputs below $\beta \cdot \mu_s$ to zero. Then, when we train a classifier by minimizing the objective

$$U_f = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} R_1(f(x_i), f(y_j)), \qquad (12)$$

the optimization is focused on those samples with outputs larger than $\beta \cdot \mu_s$. Figure 7 demonstrates the improved lower-left part of the ROC curve over *Credit Approval* achieved by minimizing U_f. Two 10-fold cross validation ROC curves are shown in Figure 7, with the curve obtained by minimizing U_R replicated from Figure 4.

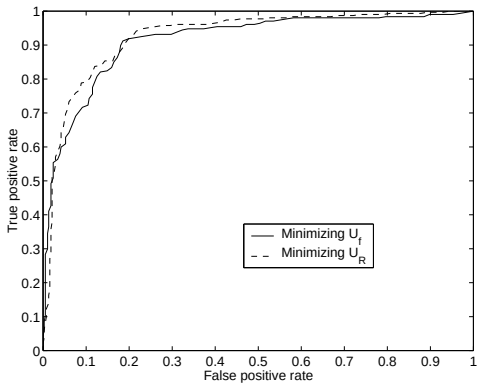

Figure 7. Minimizing U_f improves the lower-left part of the ROC curve over that obtained by minimizing U_R. Here, for U_f, $\alpha = 1.1$, $\beta = 1$ in $f(s)$, and $\gamma = 0.2$, $p = 3$ in $R_1(f(x_i), f(y_j))$.

6. Conclusion and future work

We have demonstrated that mean-squared error and cross entropy are not the most appropriate objective functions for training a classifier when the goal is to maximize the discriminative ability of the classifier across a range of decision thresholds. We proposed a new objective function that is a differentiable approximation to the WMW statistic, or to the area under the ROC curve. Results both over two UC Irvine data sets and several real-world customer behavior prediction problems demonstrate reliable improvements achieved by our new algorithm. In some cases the magnitude of the improvement may appear to be small, but the resulting economic gain is substantial in a large-scale real world application (Mozer et al., 2001).

Our work can still be extended in several directions. Hand and Till (2001) proposed a generalization of the AUC for multi-class classification problems based on the WMW statistic. The algorithm proposed here can be directly applied to this generalization, and may be advantageous for problems with imbalanced class distribution. Secondly, our work is related to Bayesian similarity (Breuel, 2001). In (Breuel, 2001), the optimization requires mn training patterns and a subsequent stage of simulated annealing. For large m and n, this process is intractable. We can optimize an objective associated with Bayesian similarity in a similar way to the algorithm here for ROC optimization. Finally, it would also be interesting to extend the algorithm to focus on any specified range of the ROC curve.

Acknowledgments

An earlier version of this work was presented at the NIPS 2002 workshop *Learning Rankings* in Whistler, Canada. The authors thank Tye Rattenbury of UC Berkeley, Cesar Guerra of Minnesota State University, Rich Caruana of Cornell University for their help and discussions, and the anonymous reviewers for valuable comments.

References

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1994). Optimizing parameters in a ranked retrieval system using multi-query relevance feedback. In *Proc. of Symposium on Document Analysis and Information Retrieval*.

Breuel, T. M. (2001). Classification by probabilistic clustering. In *Proc. of IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 1333-1336.

Caruana, R., Baluja, S., & Mitchell, T. (1996). Using the future to "sort out" the present: Rankprop and multitask learning for medical risk evaluation. *Advances in Neural Information Processing Systems*, 8.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons, New York.

Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45, pp. 171-186.

Mann, H. B. & Whitney, D. R. (1947). On a test whether one of two random variables is stochastically larger than the other. *Ann. Math. Statist.*, 18, pp. 50-60.

Mozer, M. C., Dodier, R., Colagrosso, M., Guerra-Salcedo, C., & Wolniewicz, R. (2001). Prodding the ROC curve: constrained optimization of classifier performance. *Advances in Neural Information Processing Systems*, 14.

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Trans. on Neural Networks*, 11: 690-696.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proc. of the 15th Intl. Conf. on Machine Learning*, pp. 445-453.

Rattenbury, T. & Yan, L. (2001). Using *d'* and its variations to train neural network classifiers. Unpublished manuscript, Advanced Technology Group, CSG Analytics.

Verrelst, H., Moreau, Y., Vandewalle, J., & Timmerman, D. (1998). Use of a multi-layer perceptron to predict malignancy in ovarian tumors. *Advances in Neural Information Processing Systems*, 10.

Vogt, C. C. & Cottrell, G. W. (1998). Using d' to optimize rankings. Technical report, U.C. San Diego, CSE Department. CS98-601.

Weiss, G. M. & Provost, F. (2001). The effect of class distribution on classifier learning: an empirical study. Technical report ML-TR-44, Rutgers University, Department of Computer Science.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, pp. 80-83.

Yan, L., Miller, D. J., Mozer, M. C., & Wolniewicz, R. (2001). Improving prediction of customer behavior in nonstationary environments. In *Proc. of Intl. Joint Conf. on Neural Networks*, pp. 2258-2263.