International Journal of Fuzzy Logic and Intelligent Systems, vol. 5, no. 3, September 2005 pp. 263-267

Optimization of Classifier Performance at Local Operating Range: A Case Study in Fraud Detection

Lae-Jeong Park and Jung-Ho Moon

Department of Electronics Engineering, Kangnung National University
123 Chibyun-Dong, Kangnung 210-702, Korea

Abstract

Building classifiers for financial real-world classification problems is often plagued by severely overlapping and highly skewed class distribution. New performance measures such as receiver operating characteristic (ROC) curve and area under ROC curve (AUC) have been recently introduced in evaluating and building classifiers for those kind of problems. They are, however, ineffective to evaluation of classifier's discrimination performance in a particular class of the classification problems that interests lie in only a local operating range of the classifier. In this paper, a new method is proposed that enables us to directly improve classifier's discrimination performance at a desired local operating range by defining and optimizing a partial area under ROC curve or domain-specific curve, which is difficult to achieve with conventional classification accuracy based learning methods. The effectiveness of the proposed approach is demonstrated in terms of fraud detection capability in a real-world fraud detection problem compared with the MSE-based approach.

Key Words : Classifier, ROC, AUC, Learning

1. Introduction

In general, design and learning of classifiers for financial real-world two-class classification problems are often plagued by severely overlapping class distribution since it is likely that samples in one class are similar or even identical to those in the other class due to the nature of the problems. Examples are database marketing, churn prediction, and fraud detection.

Classification accuracy or minimization of misclassifications has been widely used to evaluate classifier's ability to discriminate between members of the two classes. For a neural network (NN) classifier, the mean squared error (MSE) between the actual output and the desired target is defined and minimized to reduce the number of misclassifications. The MSE minimization may be, however, unsuitable to evaluation of a NN classifier in the severely overlapping class problem because the MSE may not represent true positive rate (TPR) and false positive rate (FPR) of the classifier explicitly. A trade-off between TPR and FPR is often unavoidable and thus should be carefully determined in the severely overlapping class distribution problems. Furthermore, NN classifier's performance varies with time because change of class ratio and/or class distribution is not rare and even occurs frequently and unpredictably in financial classification problems. It is, sometimes, desirable to evaluate and optimize NN classifier's discrimination performance across all or partial operating ranges that depend on changes in class ratio and/or distribution, which is difficult to achieve through the MSE minimization.

Recently, receiver operating characteristic (ROC) curve[1] has been widely used to evaluate classifier's performance in the skewed and overlapping data sets since its introduction by [1] in the machine learning and data mining communities. The ROC curve of a NN classifier is obtained by adjusting a threshold on the continuous single output. The ROC curve makes it possible to visualize a trade-off between classifier's discrimination capability between the two classes, which is undistinguishable in the MSE measure, as well as a relationship between TPR and FPR. Given a class ratio and misclassification costs, the optimal decision threshold of a NN classifier (corresponding to a point on the ROC space) can be determined precisely, effectively based on its ROC curve [1].

The area under a ROC curve (AUC) has been proposed as a single classifier's performance measure to deal with a problem with specifying classifier's performance in terms of a single operating point on its ROC curve [2]. AUC may be appropriate if class ratio and costs are unknown and a single classifier should be selected to handle every possible operating points. In other words, AUC reflects an average classifier's performance in the entire operating points or the entire performance space. The larger AUC is, the better classifier's discrimination ability is. Recently, much attention have been paid to design and train classifiers by maximizing the AUC in financial and medical applications [3-7].

On the other hand, in some classification applications, it is often more desirable or important to measure classifier's discrimination performance at a local operating range than in the entire operating range unlike the AUC. For example, in medical diagnosis, TPRs of less than 0.7-0.8 would be probably unacceptable, because patients with disease should be detected

Manuscript received Feb. 19, 2005; revised Jun. 28, 2005.

[1] The TPR is plotted on the Y axis and the FPR is plotted on the X axis.

even if it turns out that it was a false detection. In credit card fraud detection, fraud detection system should not operate in a range of high FPRs due to operational constraints such as limitation of the number of transactions that can be investigated daily or weekly. In order to produce meaningful high-quality classifiers in those application domains, a method for optimizing a ROC curve at a desired operating range is required, but to my knowledge, how to optimize a partial area under a ROC curve has been rarely addressed in machine learning and data mining communities.

In this paper, a new learning approach is proposed that makes it possible to optimize classifier's discrimination performance at a desired local operating range by utilizing a cost function that is related with a partial area under a ROC or domain-specific curve, which is difficult to optimize with common MSE-based learning methods. The discrimination performance of the proposed approach is examined and compared with the MSE-based approach in a credit card fraud detection that is a representative one of severely overlapping real-world classification problems.

II. Optimization of Classifier Performance based on Partial Area Under Curve

2.1 ROC curve, AUC, and Partial AUC

The ROC curve of a NN classifier is obtained by varying a threshold θ on the continuous output of the classifier, ranging from [0,1]. An example of histograms of classifier outputs of positive and negative samples is illustrated in Fig. 1. The more overlapping the two classes are, the more likely it is that the two histograms overlap each other. Samples whose their outputs are greater than θ are classified as positives, and otherwise they are classified as negatives. Given a value of θ, a point of (FPR,TPR) on the ROC space is determined by

$$FPR = \frac{negatives\ incorrectly\ classified}{total\ negatives},$$
$$TPR = \frac{positives\ correctly\ classified}{total\ positives}. \quad (1)$$

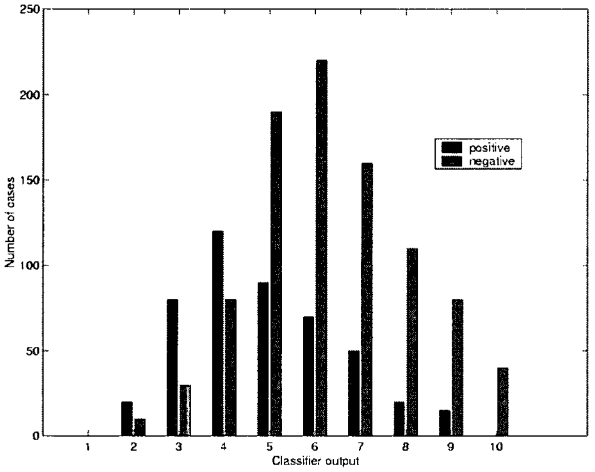


Fig. 1 Two histograms of classifier's outputs of samples of the two classes.

Fig. 2 shows an example of ROC curves of three distinct NN classifiers. Classifier *B* is completely outperformed by classifier *A*, which implies that the two histograms of outputs of classifier *B* is more overlapping than those of classifier *A*. The less likely it is that the two histograms of classifier outputs overlap each other, the better the classifier discrimination performance is, and the more bowed the ROC curve is toward the left corner of (1,1). The discrimination performance that an optimal classifier can achieve maximally depends mainly on the degree to which two classes are overlapped on the feature space.

As mentioned before, the AUC can be a single measure to quantify graphical representation of the difference of classifier performances across all decision thresholds, and thus used to compare classifier performances readily. As shown in Fig. 2, the AUC of classifier *A* is definitely larger than that of classifier *B*.

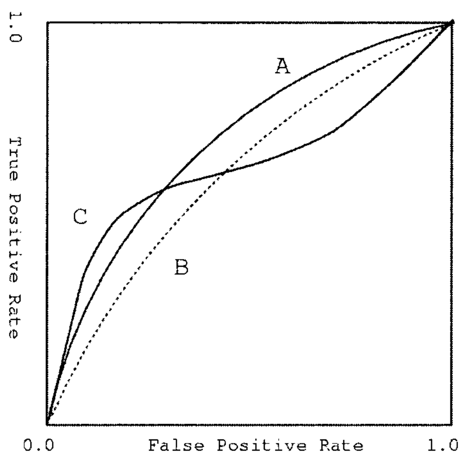


Fig. 2 ROC curves of three classifiers.

Especially, the AUC is an effective performance measure when the operating range of a classifier is completely unknown or changeable with time. The AUC is known to be exactly equal to the normalized Wilcoxon-Mann-Whitney statistic [8] and is given by

$$\frac{\sum_{i=0}^{m-1}\sum_{j=0}^{n-1} I(x_i, y_j)}{mn}, \quad (2)$$

where $I(x_i, y_j) = \begin{cases} 1 & x_i > y_j \\ 0 & otherwise \end{cases}$ is based on pairwise comparison between sample $x_i, i = 0, ..., n-1$ and sample $y_j, j = 0, ..., m-1$. The AUC has been widely adopted to assess classifier's discrimination capability informatively compared with the error rate or MSE and hence several methods to optimize the AUC directly have been proposed recently. Verrelst *et. al.* adopted simulated annealing to maximize the AUC to produce a MLP classifier in ovarian tumor malignancy prediction problem [3]. Yan *et. al.* proposed a gradient-based training algorithm for directly maximizing the AUC by using a differentiable objective function that is approximation to the Wilcoxon-Mann-Whitney statistic, which is equivalent to the AUC [4].

As mentioned before, in some cases it is important and meaningful to focus on the area under only a portion of ROC curve, for example, within a specific range of FPRs and/or TPRs because no interest lie in the entire range of FPRs and/or TPRs. For example, in a diagnostic test, much attention are generally paid to the portion of the ROC curve where TPRs are greater than a predetermined threshold. A classifier with a higher AUC value may, however, have lower discrimination performance than classifiers with lower AUC values. As shown in Fig. 2, classifier *C* is slightly better than classifier *A* at a low FPR range whereas it is much worse than classifier *A* at a high FPR range. In these applications, it is required to train classifiers by optimizing local discrimination performance at a desired local operating range that may be indistinguishable by the AUC performance measure. To my knowledge, there have been few attentions on importance of classifier's discrimination performance at a local operating range in machine learning and data mining communities and therefore, there have been few researches on methods designed to train NN classifiers by directly optimizing a partial area under the ROC curve or similar curve (PAUC) value. A method of maximizing the PAUC value is described in detail in the next subsection.

2.2 PAUC Optimization

Unfortunately, the backpropagation (BP) algorithm that minimizes the MSE at the output of a NN classifier is not suited for optimizing the classifier's performance in terms of the PAUC. It is because the MSE minimization at the output does not correspond directly to the PAUC-based optimization.

```
Input :
  [$\alpha$, $\beta$] : a range of FPRs
  $N_\theta$ : the number of thresholds between [0,1]
STEP 1:
  $N_\theta$ pairs of (TPR, FPR) are calculated by changing
    the threshold on the NN classifier's output
STEP 2:
  A ROC curve is constructed by using $N_\theta$ pairs of
    (TPR, FPR)
STEP 3:
  If the number of (TPR, FPR) falling between [$\alpha$,$\beta$] is
    larger than a predetermined value $\kappa$, a partial area
    under the curve between [$\alpha$,$\beta$] is calculated with the
    trapezoidal integration rule
  Otherwise, the penalty is returned
```

Fig. 3 A procedure for numerical PAUC calculation.

A nonstandard training method is needed because it is difficult to compute the partial derivatives of the PAUC with respect to the weights. A NN classifier is therefore trained by one of effective search algorithms or Evolutionary Programming (EP) rather than gradient-descent algorithms so that the weights and its threshold θ are chosen in terms of the PAUC optimization. An objective function for the PAUC optimization is given in the form of

$$\frac{\int_{\alpha}^{\beta} C(x)dx}{\int_{0}^{1} C(x)dx}. \tag{3}$$

where [α,β] specifies a desired operating range, $C(x)$ is a curve or function of x, and x is a false positive rate or problem-dependent independent variable. Fig. 3 shows a procedure for the numerical PAUC calculation of a NN classifier with a single output. NN Classifiers that the number of (TPR, FPR) points falling into [α, β] is smaller than κ, is excluded or penalized during the evolutionary search in order not to produce NN classifiers that may generate an abrupt transition of histograms of outputs, which is related with poor discrimination performance.

III. Experimental Results

3.1 Fraud Detection

Credit card fraud detection is one of real-world two-class classification problems where class distribution is severely overlapping as well as the class ratio is highly skewed. Fraudulent transactions are overwhelmed by legitimate ones in number. The class ratio of fraudulent to legitimate transactions ranges from 10^{-3} to 10^{-4} [9,10]. In addition, fraudulent transactions resemble legitimate ones so that it is impossible to detect fraudulent transactions without false detection of legitimate ones. Hence, in practice, a classifier for credit card fraud detection should be operated on a low range of FPRs in order to prevent the number of legitimate transactions incorrectly detected from being larger than the maximum number of suspicious transactions investigated daily or weekly. In other words, a constraint that the total number of transactions investigated on a regular basis is imposed on credit card fraud detection. Therefore, it is crucial to optimize classifier's discrimination performance at a specific local range of false positive rates associated 'with the chosen operating range, which is determined by an operator of credit card fraud detection system. Note that learning of classifiers is performed in an operating range, not a single operating point because an operating point changes with time, from a month to another, depending on the changing class distribution.

3.2 Data Sets and Classifier

Two data sets of credit card transactions labeled as legitimate or fraudulent were provided by a credit card company in Korea. One set consisting of about 51,260 transactions collected selectively during one year is used as a training data set. Another data set of about 7 millions transactions that were collected during three consecutive months is used as an evaluation data set of the NN classifier.

In this experiment, a classification approach based on common features identifying fraudulent transactions is implemented rather than a user profiling-based approach. A multi-layered perceptron (MLP) with one hidden layer and sigmoidal function is used as a NN classifier. The ten features are extracted from each transaction and are then used for the inputs of the NN classifier.

3.3 Application-specific PAUC Maximization

For credit card fraud detection, it is practically desirable that an operating range of a NN classifier is represented by a rejection rate (true positives plus false positives) since the rate is a major monitoring variable that is carefully controlled due to the constraint imposed on the capacity of fraud investigation. The number of fraudulent transactions may fluctuate with time compared with the total number of transactions. It is therefore preferred that the Y-axis on the ROC curve or TPR is replaced by the number of fraudulent transactions correctly detected. To meet the specifications, instead of a ROC curve, a domain-specific curve $C(x)$ is introduced that represents the number of correctly detected fraudulent transactions with respect to a rejection rate, x.

The EP is used to train a NN classifier in order to maximize Eq. (3) where [α, β] is [0.02, 0.04]. With the domain-specific curve, the objective is to maximize an average of the correctly detected frauds in a chosen range of rejection rates by maximizing the partial area under the curve. A penalty term is added to a cost function of EP so as to penalize classifiers whose decision boundary lie in the regions on the feature space where sample distribution is very dense. The value of κ is determined by trial-and-error and is set to 20. The parameter values of EP with Cauchy mutation for all experiments are as follows. The number of generations and the population size are chosen as 1000 and 30, respectively, for a reasonable convergence speed. The tournament size is set to 15.

3.4 Experimental Results and Discussions

Two sets of experiments were performed to evaluate and compare PAUC-based classifiers with the MSE-based ones. For each learning criterion, 30 NN classifiers were trained with 10, 15, 20, 25, 30 hidden nodes. The operating range was chosen as [0.001,0.004] to reflect reality.

Table 1 shows the averaged MSE and PAUC values of thirty classifiers trained by the PAUC criterion and thirty ones by the MSE criterion. As expected, the average MSE value of classifiers trained by the MSE criterion is significantly smaller than that of classifiers trained by the PAUC criterion whereas the average PAUC value of classifiers trained by the PAUC criterion is significantly larger than that of classifiers trained by the MSE criterion. An explanation could be that, by maximization of Eq. (3) the decision boundary of a NN classifier is formed so as to increase $C(x)$ in the operating range of [α, β] by the sacrifice of large MSE contribution of the samples that lie in regions on the feature space corresponding to the rejection rates of [β, 1.0]. The operating range of high rejection rates greater than β is of no consequence in credit card fraud detection.

Interestingly, classifiers trained by the PAUC criterion (MSE criterion) have a large standard deviation in MSE values (PAUC values) compared with the standard deviation in PAUC values (MSE values), which implies that the MSE minimization does not correspond to the PAUC maximization one-to-one. The similar results between AUC and MSE performance indices have been observed in [6,7]. It has been revealed analytically in [6] that algorithms designed to minimize the error rate may not lead to the best AUC possible values. In [7], by comparison of error surfaces in weight space of MSE, AUC, and partial AUC performance measures, they showed that MSE minimization tended to maximize AUC, but not partial AUC defined at a range of high true positive rates.

Table 1. Averages and standard deviations of MSE and PAUC values of sixty NN classifiers, a half trained by the PAUC criterion, the others by the MSE criterion.

Learning criterion	(MSE, PAUC)
MSE minimization	$(3.07 \pm 0.03, 2.91 \pm 0.10)$
PAUC maximization	$(4.05 \pm 0.85, 3.05 \pm 0.02)$

Fig. 4 shows a difference between the two averaged $C(x)$ curves of the classifiers trained by the PAUC maximization and ones trained by the MSE minimization on the evaluation set. It should be noted that even though training of NN classifiers with the PAUC criterion generates classifiers with large MSE values, it achieves the goal of increasing the partial area under curve $C(x)$ for the prescribed interval of rejection rates of [0.001,0.004], compared with the MSE-based classifiers. To compensate the increase in the range of [0.001,0.004], the $C(x)$ curve of the PAUC-based classifier increases more slowly at a medium range of rejection rates than that of MSE-based classifiers, resulting in negative values of the curve in a range of rejection rates above 0.4, as

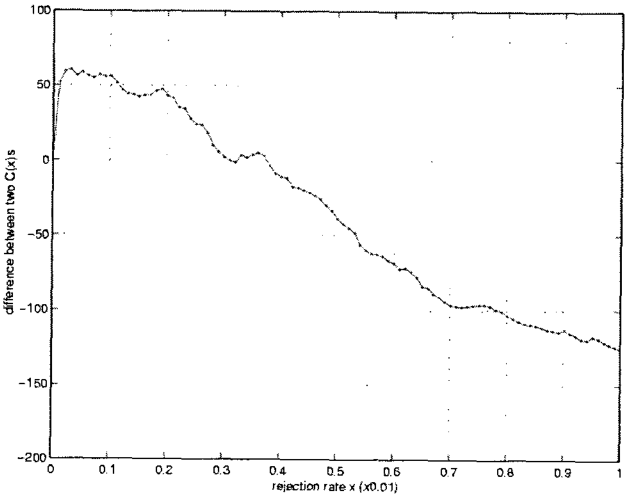


Fig. 4 Difference between the two average $C(x)$ curves of NN classifiers that were trained by in terms of the PAUC and MSE criteria.

shown in Fig. 4. The performance improvement on the specified operating range is, whether significant or minor, consistent with classifiers with the number of hidden nodes between 10 and 30.

Ⅳ. Conclusions

In this paper, a classifier learning approach has been proposed that enables to optimize neural network classifier's discrimination performance at a desired local operating range by maximizing a partial area under a ROC or domain-specific curve, which is difficult to achieve with conventional classification accuracy or MSE-based learning methods. An evolution-based search algorithm has been used to optimize the partial area under curve since the local performance measure itself is not expressed by a differentiable function with respect to the classifier's output.

The effectiveness of the proposed approach has been demonstrated and compared with the MSE-based approach in terms of local discrimination performance in a credit card fraud detection. The experimental results have shown that the proposed learning method makes it possible to detect more fraudulent transactions at a local range of very low false positive rates by maximizing the partial area under a domain-specific curve, compared with the MSE minimization. We think that the proposed approach can be also successfully applied to other two-class classification problems that interests lie not in the entire operating range but in a local operating range like fraud detection.

References

[1] F. Provost and T. Fawcett, "Analysis and visualization of classifier performance comparison under imprecise class and cost distributions," *Proc. Int'l Conf. Knowledge Discovery and Data Mining*, pp. 43-48, 1997.

[2] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol 30, pp. 1145-1159, 1997.

[3] H. Verrelst, Y. Moreau, J. Vanderwalle, and D. Timmerman, "Use of a multi-layer perceptron to predict malignancy in ovarian tumors," *Advances in Neural Information Processing Systems*, vol. 10, 1998.

[4] L. Yan, R, Dodier, M. Mozer, and R. Wolniewicz, "Optimizing classifier performance via the Wilcoxon-Mann-Whitney statistics," *Proc. Int'l Conf. Machine Learning*, pp. 848-855, 2003.

[5] B. Sahiner, H.-P. Chan, N. Petrick, S. S. Gopal, and M. M. Goodsitt, "Neural network design for optimization of the partial area under the receiver operating characteristic curve," *Proc. of IEEE Int. Conf. on Neural Networks*, 1997.

[6] C. Cortes and M. Mohri, "AUC optimization vs. error rate minimization," *Advances in Neural Information Processing Systems*, vol. 15, 2003.

[7] M. K, Markey, J. Y. Lo, R. Vargas-Woracek, G. D. Tourassi, and C. E. Floyd Jr., "Perceptron error surface analysis: a case study in breast cancer diagnosis," *Computers in Biology and Medicine*, vol. 32, pp. 99-109, 2002.

[8] H. B. Mann and D. R. Whitney, "On a test whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, pp. 50-60, 1947.

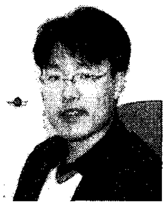
[9] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, pp. 235-255, 2002.

[10] P. K. Chan, W. Fan, A. L. Prodromidis, and S. J. Stolfo, "Distributed data mining in credit card fraud detection," *IEEE Intelligent Systems*, vol. 14, 57-74, 1999.

Lae-Jeong Park

Lae-Jeong Park was born in Pusan, Korea, on October 28, 1968. He received the B.S. degree in Electrical Engineering from Seoul National University, Seoul, Korea, in 1991, and the M.S. and Ph.D. degrees in Electrical Engineering from the Korea Advanced Institute of Science and Technology, Taejon, Korea, in 1993 and 1997, respectively. From 1997 to 1999, he was with the Information Technology Lab. at LG Corporate Institute of Technology, Seoul, Korea. He is currently an assistant professor in the Department of Electronics Engineering at Kangnung National University, Kangnung, Korea. His current research interests are machine learning, pattern recognition, and evolutionary and neural computation.

Phone : +82-33-640-2389
Fax : +82-33-646-0740
E-mail : ljpark@kangnung.ac.kr



Jung-Ho Moon

Jung-Ho Moon received the B.S. degree in control and instrumentation engineering from Seoul National University in 1991, the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology in 1993 and 1998, respectively. In 2002, he joined the faculty of the Department of Electronics Engineering, Kangnung National University, Korea, where he is an Assistant Professor. From 1998 to 2000, he worked for Samsung Electronics Co., Ltd as a Senior Research Engineer, and from 2001 to 2002, he worked for Humax Co., Ltd as a Senior H/W Engineer. His research interests include digital control, disk drive servo systems, and embedded systems.

Phone : +82-33-640-2427
Fax : +82-33-646-0740
E-mail : itsmoon@kangnung.ac.kr