

# The Use of Decision Threshold Adjustment in Classification for Cancer Prediction

James J. Chen<sup>1</sup>, Chen-An Tsai<sup>2</sup>, Hojin Moon<sup>1</sup>, Hongshik Ahn<sup>3</sup>, John J. Young<sup>1</sup>,  
and Chun-houh Chen<sup>2</sup>

<sup>1</sup> Division of Biometry and Risk Assessment  
National Center for Toxicological Research  
Food and Drug Administration  
Jefferson, Arkansas 72079

<sup>2</sup> Institute of Statistical Science  
Academia Sinica  
Taipei, 11529  
Taiwan

<sup>3</sup> Department of Applied Mathematics and Statistics  
Stony Brook University  
Stony Brook, NY, 11794

Send correspondence to:

Dr. James J. Chen  
HFT-20  
Jefferson, AR 72079  
Tel:(870)-543-7007  
Fax:(870)-543-7662  
E-mail [jchen@nctr.fda.gov](mailto:jchen@nctr.fda.gov)

Abbreviated title: Decision Thresholds in Classification

## Summary

Standard classification algorithms are generally designed to maximize the number of correct predictions (concordance). The criterion of maximizing the concordance may not be appropriate in certain applications. In practice, some applications may emphasize high sensitivity (e.g., clinical diagnostic tests) and others may emphasize high specificity (e.g., epidemiology screening studies). This paper considers effects of the decision threshold on sensitivity, specificity, and concordance for four classification methods: logistic regression, classification tree, Fisher's linear discriminant analysis, and a weighted k-nearest neighbor. We investigated the use of decision threshold adjustment to improve performance of either sensitivity or specificity of a classifier under specific conditions. We conducted a Monte Carlo simulation showing that as the decision threshold increases, the sensitivity decreases and the specificity increases; but, the concordance values in an interval around the maximum concordance are similar. For specified sensitivity and specificity levels, an optimal decision threshold might be determined in an interval around the maximum concordance that meets the specified requirement. Three example data sets were analyzed for illustrations.

KEY WORDS: concordance, cross validation, weighted k-NN, receiver operating characteristic curve, sensitivity, specificity.

## INTRODUCTION

Classification/prediction (machine learning) has been a widely used data mining technique in many areas of research and applications. Class prediction has been used to predict the activity or toxicological property of untested chemicals, for instance, to predict rodent carcinogenicity [1], *Salmonella* mutagenicity [2,3], or estrogen receptor binding activity [4] of chemicals using structure-activity relationship models. Recently, class prediction models have been developed to classify tumor and normal colon tissues based on gene expression profiles [5], to identify marker genes for distinguishing between acute lymphoblastic leukemias (ALL) and acute myeloid leukemias (AML) [6] based on gene expression data, and to diagnose ovarian and prostate cancers based on proteomic SELDI-TOF MS (Surface Enhanced Laser Desorption-Ionization Time-Of-Flight Mass Spectrometry) data [7].

Development of a class prediction algorithm generally consists of three components: 1) selection of predictors, 2) selection of a classification algorithm to develop the prediction rule, and 3) performance assessment. The first two components build a prediction model, and the third component assesses the performance of the prediction model. Sensitivity and specificity are two primary criteria used in the evaluation of the performance of a classification algorithm. The sensitivity is the proportion of correct positive classifications out of the number of true positives. The specificity is the proportion of correct negative classifications out of the number of true negatives. The concordance is the total number of correct classifications out of the total number of samples.

A classification model is developed based on a training data set. Sensitivity and specificity of a prediction algorithm (a classifier) can depend on the makeup of the numbers of positives to the number of negatives in the training samples. When the class sizes are not equal, depending on the classification methods, the derived classifier may favor the larger class. In general, the majority class of positive will have a high sensitivity and the minority class will have a low specificity, and vice versa. This problem has been addressed by learning from imbalanced data set. Applications include detection of fraudulent telephone calls, detection of oil spills in satellite images, clinical diagnostic test of rare diseases, where the positive data are rare as compared to the negative data [8-10]. In these applications, the main interest is toward correct classification of positive samples (high sensitivity in predicting the minority class samples). For example, if the ratio of positive-to-negative is on the order of 1 to 100, then a procedure will have 99% concordance by simply predicting all to be negative (99% specificity and 0% sensitivity). This procedure is obviously not useful for these applications. Other

applications such as epidemiology screening studies may emphasize high specificity. The challenge is to develop a prediction model that can provide an acceptable sensitivity (or specificity) from the available data set.

Most of the current standard classification algorithms are designed to minimize zero-one loss, in other words, to minimize the number of incorrect predictions or to maximize the concordance. Maximizing concordance criterion is based on an assumption of an equal cost of misclassifications. This criterion may not be appropriate when the class sizes are imbalanced or misclassification costs are unequal. Two approaches have been proposed to account for imbalanced class sizes or differential misclassification costs: 1) sampling techniques, and 2) adjusting decision threshold. The sampling technique is a commonly used practice in dealing with imbalanced data set by balancing the data set by either under-sampling the majority class or over-sampling the minority class [8,10-11]. Chen et al. [12] proposed using a bagging method by applying resampling techniques repeatedly to build multiple base classifiers and synthesizing their predictions to make the overall prediction by majority voting. Adjusting decision threshold approach to account for differential misclassification costs and/or prior probabilities has been proposed and discussed by several researchers via ROC (receiver operating characteristic) analysis [13-15]. For example, Provost and Fawcett [15] proposed a ROC convex hull method by combining ROC analysis with decision analysis for comparing the performance of a set of classifiers and identifying the optimal classifier or a subset of potentially optimal classifiers.

The purpose of this paper is to study the effects of changes of the decision threshold on sensitivity, specificity, and concordance for the four classification methods: logistic regression, Fisher's linear discriminant analysis, classification tree, and a weighted k-nearest neighbor. Monte Carlo simulations were conducted to examine the relative behaviors of the sensitivity, specificity, and concordance. We also adapted the ROC-type analysis to estimate the optimal decision threshold for specified misclassification costs and/or prior probability of class distribution. The primary focus is the use of the decision threshold adjustment to improve the performance of sensitivity or specificity of a classifier under specific conditions or study objectives. To our knowledge, there previously has not been a systematic analysis of sensitivity and specificity. Under the model of an equal misclassification cost and with specified desirable sensitivity and specificity levels, we estimate a range of decision thresholds that meets the specified sensitivity and specificity levels, and then determine the most appropriate decision threshold that corresponds to the maximum concordance.

## MATERIALS AND METHODS

### Classification Algorithms

Let  $\mathcal{D} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n\}$  be the set consisting of  $n$  labelled samples. Each sample consists of two parts,  $\mathbf{t}_i = (\mathbf{x}_i, y_i)$ , where  $\mathbf{x}_i$  is a vector of predictors from  $m$ -dimensional space, and  $y_i$  is a categorical variable for a set of possible labels  $\mathcal{Y}$ . In the binary classification  $\mathcal{Y}$  consists 0 (class 0 or negative samples) and 1 (class 1 or positive samples). Let  $n_0$  denote the number of negative samples and  $n_1$  denote the number of positive samples. A future unlabelled sample  $\mathbf{x}$  is classified by applying the prediction rule (a classifier) built on  $\mathcal{D}$  to predict the unknown  $y$  as either 0 or 1.

We consider the four well known classification methods: logistic regression (LR), classification tree (CTree), and Fisher’s linear discriminant analysis (FLDA), and a weighted  $k$ -nearest-neighbor classifier ( $k$ -NN), a modified  $k$ -NN. Each procedure is briefly described below assuming an equal misclassification cost and equal prior probability of the class distribution.

The functional form of the logistic regression model [16] is

$$P(y = 1|\mathbf{x}_i) = \frac{\exp(\sum_j \beta_j x_{ij})}{1 + \exp(\sum_j \beta_j x_{ij})},$$

where  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{im})$  is the predictor variable. For the given value of an predictor  $\mathbf{x}$ , the predictive output value, denoted by  $\hat{y}$ , represents the probability that the sample  $x$  is from class 1. The default decision threshold uses 0.5 to predict class membership. The decision rule assigns  $\mathbf{x}$  to class 1 if  $\hat{y} \geq 0.5$  and to class 0 if  $\hat{y} < 0.5$ . This rule implies an equal prior probability of class membership for  $\mathbf{x}$ . The decision threshold can be adjusted, for example, to  $n_1/(n_0 + n_1)$  to reflect differential class sizes or prior probabilities.

The CTree performs binary recursive partitioning [17]. The algorithm recursively partitions parent nodes into two child nodes by splitting the corresponding covariate space into regions selected on the basis of maximum reduction in node impurity measured by entropy or information,  $-\sum_j p(c|t) \log[p(c|t)]$ , or measured by the Gini index of diversity [17],  $1 - \sum_c p(c|t)^2$ , where  $p(c|t)$  is the probability that a sample is in class  $c$  (0 or 1) given that it falls into a node  $t$ . The partitioning algorithm is recursive until a terminal node is reached for which no split improves the within-node homogeneity or the node size is too small. To avoid over-fitting data, the cross-validation approach with minimal cost-complexity

pruning method is used. The CTree assigns each terminal node to the class  $c = 1$  if the terminal node  $p(c|t)$  is greater than the threshold. The threshold of 0.5 is the default.

Let  $\mu_c$  denote the mean of  $\mathbf{x}$  for the class  $c$  ( $c = 0, 1$ ), and  $\Sigma$  denote the covariance matrix. The Fisher’s linear discriminant analysis (FLDA) [16] assigns  $x$  to class 1 if

$$[x - (\mu_0 + \mu_1)/2]^T \Sigma^{-1} (\mu_1 - \mu_0) > \log(n_0/n_1);$$

otherwise, assigns  $x$  to class 0. The FLDA produces a binary output; the decision rule assigns  $x$  to either class 0 or class 1 according to the relative class size (in log scale),  $\log(n_0/n_1)$ . Unlike LR or CTree,  $\log(n_0/n_1)$  is not the probability of a class membership. For the purpose of evaluations across classification methods,  $\log(n_0/n_1)$  is re-scaled to the corresponding decision threshold  $n_0/(n_0 + n_1)$ . For example, if the negative-to-positive ratio is 2 to 1, then the default FLDA cutoff is  $\log 2$ , the corresponding decision threshold is  $2/3$ .

The  $k$ -NN classifiers [16] can be based on either a distance or a similarity metric, where  $k$  is an odd number. Given a future sample  $\mathbf{x}$ ,  $k$ -NN method finds the  $k$  nearest neighbors to  $x$ , and then classifies using majority vote among the  $k$  neighbors. The choice of  $k$  will influence the performance of a  $k$ -NN classifier. The  $k$  may be determined by cross validation. Let  $l$  denote the number of class 1 samples in the  $k$  neighbors. The  $k$ -NN method assigns  $x$  to either class 0 or class 1 by a majority voting; that is, it assigns  $\mathbf{x}$  to class 1 if  $l/k \geq 0.5$  and to class 0 if  $l/k < 0.5$ . The 0.5 can be regarded as a default decision threshold. The decision threshold can be adjusted to  $i/k$  ( $i = 1, 2, \dots, k$ ). This extension would require a large  $k$ . Alternately, a distance weighting  $k$ -NN classifier was used, a weighted  $k$ -NN algorithm. Let  $N_k(\mathbf{x})$  denote the  $k$  nearest neighbors of a future sample  $\mathbf{x}$ . The similarity between  $\mathbf{x}$  and a sample  $\mathbf{x}_l$  in the nearest neighbor ( $\mathbf{x}_l \in N_k(\mathbf{x})$ ) is denoted by  $\text{sim}(\mathbf{x}, \mathbf{x}_l)$ ,  $l = 1, \dots, k$ . The normalized similarity between  $\mathbf{x}$  and  $\mathbf{x}_l$  is then

$$w_l = \frac{\text{sim}(\mathbf{x}, \mathbf{x}_l)}{\sum_{\mathbf{x}_l \in N_k(\mathbf{x})} \text{sim}(\mathbf{x}, \mathbf{x}_l)}.$$

The probability (measure) that  $\mathbf{x}$  is in class  $j$  can be expressed as

$$P(y = j|\mathbf{x}) = \sum_{\mathbf{x}_l \in N_k(\mathbf{x})} \delta_l w_l,$$

where  $\delta_l$  is an indicator function with  $\delta_l = 1$  if  $y = j$ , and  $\delta_l = 0$  otherwise. There are various kernel functions for similarity measure. We apply the Gaussian kernel function to define the similarity, based

on our preliminary empirical studies. The Gaussian kernel function is

$$\text{sim}(\mathbf{x}, \mathbf{x}_l) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{d^2(\mathbf{x}, \mathbf{x}_l)}{2\sigma^2} \right],$$

where  $d(\mathbf{x}, \mathbf{x}_l)$  is the Euclidean distance between  $x$  and  $\mathbf{x}_l$ . The weighted  $k$ -NN classifier assigns  $\mathbf{x}$  to class 1 if  $P(y = 1|\mathbf{x}) > \tau$ , otherwise to class 0, where  $\tau$  is the decision threshold.

### Decision Threshold Adjustment

For a given classification method (e.g., logistic regression), the sensitivity, specificity, and concordance depend on the chosen threshold  $\tau$ . For a given decision threshold, the performance of a classifier can be summarized by a  $2 \times 2$  confusion matrix (Table 1). Let  $TP(\tau)$  and  $TN(\tau)$  be the numbers of correct predictions for the positive and negative samples, respectively. The fraction  $SN(\tau) = TP(\tau)/n_1$  is the sensitivity,  $SP(\tau) = TN(\tau)/n_0$  is the specificity, and  $(TN(\tau) + TP(\tau))/n$  is the concordance. For instance, change of decision threshold from 0.5 to 0.1 will generally result in increasing the sensitivity (increasing TP) and decreasing the specificity (decreasing TN). When the class sample sizes are equal, a classifier using the default threshold should have unbiased estimates of the sensitivity, specificity, and concordance. But, when the class sizes are different, a classifier using the default threshold may lead to an unacceptably low sensitivity (or specificity, depending on the objective of the study). Simulation results on the effect of the ratio between two class sample sizes on the sensitivity, specificity, and concordance are shown in the next section.

The Receiver Operating Characteristic (ROC) analysis has been developed to determine an optimal decision threshold for relative costs of false positive and false negative errors [13-15,18]. An ROC curve is the plot of sensitivity ( $SN$ ) versus false positive rate ( $FPR$ ) (or 1-specificity); each point on the curve corresponds to a different threshold  $\tau$  that separates the negative samples from the positive samples. In the remaining section, the ROC analysis was applied to improve performance of sensitivity or specificity of a classification algorithm. However, unlike the conventional use of the ROC analysis for determining the optimal decision threshold with respect to the total expected cost, the relationship between  $\tau$  and  $SN$  and between  $\tau$  and  $SP$  was applied to improve performance of either sensitivity or specificity.

Denote the prior probability of negative and positive as  $\pi_0$  and  $\pi_1$ , respectively. Let FP\$ and FN\$ denote, respectively, the cost for making false positive and false negative errors, and let  $P(FP)$  and  $P(FN)$  denote the corresponding probabilities of making false positive and false negative errors,

respectively. The expected cost for a false positive error is  $C_{FP} = P(FP) \cdot FP\$ = \pi_0 \cdot (1 - SP) \cdot FP\$$  and of making a false negative error is  $C_{FN} = P(FN) \cdot FN\$ = \pi_1 \cdot (1 - SN) \cdot FN\$$ . The cost function  $C_{FP}$  ( $C_{FN}$ ) is a non-decreasing (non-increasing) function of  $\tau$ . The total expected cost is the sum of the false positive cost and false negative cost,

$$C_{Total} = \pi_0 \cdot (1 - SP) \cdot FP\$ + \pi_1 \cdot (1 - SN) \cdot FN\$ = \pi_0 \cdot FPR \cdot FP\$ + \pi_1 \cdot (1 - SN) \cdot FN\$.$$

Note that the sensitivity ( $SN$ ) is a function of false positive rate ( $FPR$ ) by the curve of ROC. Thus, the total expected cost is equivalent to

$$C_{Total} = \pi_0 \cdot FPR \cdot FP\$ + \pi_1 \cdot [1 - ROC(FPR)] \cdot FN\$.$$

The optimal cutoff for minimal cost can be obtained by taking the derivative with respect to  $FPR$  and setting it to zero:

$$\lambda' \equiv \frac{dROC(FPR)}{dFPR} = (\pi_0/\pi_1) \cdot (FP\$/FN\$).$$

The  $\lambda'$  is the slope of the tangent to the ROC curve ( $1 - SP, SN$ ) at the optimal point. Traditionally, the optimal cutoff point is obtained by a graphic method which moves a line with the above slope that intersects (is tangent to) the ROC curve [19]. Alternatively, we suggest the optimal cutoff be computed empirically by directly evaluating all  $C_{Total}$ 's; the optimal cutoff corresponds to the ( $1 - SP$ ) for which  $C_{Total}$  is the minimum.

When the misclassification costs are equal, say,  $FP\$ = FN\$ = C$ , the total cost becomes

$$C_{Total} = C \cdot [\pi_0 \cdot (1 - SP) + \pi_1 \cdot (1 - SN)].$$

The minimization of the total cost is equivalent to minimization of the predictive error rate

$$P_{Err} = 1 - (SN \cdot \pi_1 + SP \cdot \pi_0).$$

Note that the sensitivity, specificity, and concordance measure accurate performance of the prediction rule for the current sample data, regardless of the class distribution,  $\pi_0$  and  $\pi_1$ . But, the total cost function (or predictive error rate) takes the class distribution into consideration. When the class sample proportions represent class probabilities, i.e.,  $\pi_0 = n_0/n$  and  $\pi_1 = n_1/n$ , the optimal cutoff has the tangent slope of  $\lambda' = n_0/n_1$ . Note that this ratio is the same as the cutoff of FLDA (in

log scale). The corresponding decision threshold is  $n_0/n$ . When  $n_0 > n_1$ , the  $n_0/n > 0.5$ . This proportion will impose more weights on the majority class (class 0). That is, under the model of an equal misclassification cost, setting the decision threshold at the proportion of the class 0 samples,  $n_0/n$ , will have the minimum predictive error. On the other hand, the effect of unequal class sizes might be alleviated by imposing more weight on the minority class. That is, setting the decision threshold at the proportion of the class 1 samples,  $n_1/n$ , should have better balance in the sensitivity and specificity. We will investigate the performance of these two sample proportions, denoted by  $p_0 = n_0/n$  and  $p_1 = n_1/n$ , for the decision threshold for the four classification methods.

In many practical applications, either the misclassification costs or the prior probabilities of the class distribution are not known; it is not feasible to estimate an optimal decision threshold. However, it may be possible to find a range of the decision thresholds such that the corresponding classifiers have at least the specified desirable sensitivity and specificity levels. For specified *SN* and *SP* levels, the range of  $\tau$  can be obtained via the monotonic relationship between  $\tau$  and *SN* and between  $\tau$  and *SP*. As the threshold  $\tau$  increases, *SN* decreases and *SP* increases. Let  $\tau_l$  be the largest decision threshold such that the corresponding *SN* meets the specified sensitivity level and  $\tau_u$  be the smallest decision threshold such that the corresponding *SP* meets the desired specificity. The classification models corresponding to the interval  $(\tau_l, \tau_u)$ , if  $\tau_l < \tau_u$ , will have the desired sensitivity and specificity. The classifier with the highest concordance value was chosen. The interval  $(\tau_l, \tau_u)$  is empty if  $\tau_l > \tau_u$ . This classifier is generally sub-optimal with respect to the concordance.

## RESULTS

### Simulation Experiments

A simulation study was conducted to examine the effect of the decision threshold on sensitivity, specificity, and concordance for the negative-to-positive ratio of 56:56 (an equal class size) and 112:56 (unequal class sizes). Since the LR and FLDA methods generally require the number of samples much larger than the number of predictors [20], we used 20 predictors with class sample sizes 56 and 112. For each simulated data set, ten 10-fold cross validation with 10 different partitions were performed, and the sensitivity, specificity, and concordance were calculated using the LR, CTree, FLDA, and weighted  $k$ -NN classification algorithms. The entire process were repeated 100 times to obtain different

simulated sample data. The mean and standard deviation of 1,000 (10 x 100) sensitivities, specificities, and concordances were calculated.

The first simulation considered the model (M0), in which two classes are from the same population. All 20 predictors were randomly generated from  $N(0, .2^2)$ , and samples were arbitrarily assigned to either class 0 or class 1. Because of no underlying difference between two classes, the prediction accuracy is expected to be 0.5. Figure 1 shows the plots of the sensitivity, specificity, and concordance for the (equal) class size 56:56 (upper panel) and the (unequal) class sizes 112:56 (lower panel). In both equal and unequal class sizes, the sensitivity (SN) decreases and specificity (SP) increases as the threshold increases. The concordances (CC) are almost constant at about 0.5 in the equal class size. The concordance generally increases with the decision threshold, and reaches its maximum at about 2/3, the proportion of majority class, in the unequal class sizes. LR and FLDA have low concordance values for small  $\tau$ , e.g., both CC's are about 35% when  $\tau = 0.1$ . For CTree and weighted  $k$ -NN, the concordances are close to 50% for  $\tau = 0.1$  and increase gradually. LR, CTree, and Weight  $k$ -NN have the 50% concordance at about  $\tau = 1/3$ ; while FLDA has the concordance of 50% at  $\tau = 1/2$ .

In the second and third simulations, all 20 predictors in the class 0 samples and the first 8 predictors (random noises) in the class 1 samples were generated from  $N(0, .2^2)$ . The remaining 12 predictors were generated from  $N(.1, .2^2)$  (M1) in the second simulation and from  $N(.2, .2^2)$  (M2) in the third simulation.

Figure 2 shows the plots of the sensitivity, specificity, and concordance of M1 (upper panel) and M2 (lower panel) for the equal class size. When the class sizes are equal, the two sample proportions are 0.5. All four methods reach their maximums at about  $\tau = 0.5$ , as expected. In both M1 and M2, the concordance values at  $\tau$  between 0.3 and 0.7 are less than 1% different from their respective maximums. FLDA and LR clearly outperform CTree and Weighted  $k$ -NN; CTree is the poorest in this simulation. In M1, the maxima are 75%, 60%, 75%, and 63% for LR, CTree, FLDA, and weighted  $k$ -NN, respectively. In M2, FLDA appears to be slightly better than LR. The maxima are 90%, 76%, 93%, and 87% for the four methods, respectively. For example, LR has SN = 74.5%, SP = 74.7%, and CC = 74.6% for M1 and SN = 89.7%, SP = 89.7%, and CC = 89.7% for M2.

Figure 3 shows the plots for the unequal class sizes. The general patterns of the sensitivity, specificity, and concordance are similar to those shown in Figures 1 and 2. FLDA has the best concordances in M1 and M2 and CTree is the poorest. LR and FLDA reach their respective maximums

at the decision threshold  $\tau = p_0 = 0.67$ . However, FLDA has the best balance between the sensitivity and specificity at about  $\tau = 0.5$ , while LR has the best balance at about the decision threshold  $\tau = p_1 = 0.33$ . For example, in M1 FLDA has SN = 65.4%, SP = 86.3%, and CC = 79.3% for  $\tau = 0.67$ ; it has SN = 76.0%, SP = 78.5%, and CC = 77.7% for  $\tau = 0.50$ . LR has SN = 55.1%, SP = 90.3%, and CC = 78.6% for  $\tau = 0.67$ ; it has SN = 74.0%, SP = 79.3%, and CC = 77.6% for  $\tau = 0.33$ . Note that the differences in the concordances between the optimal threshold and best balance threshold is about 1%. CTree and Weighted  $k$ -NN have the maximum concordances at about  $\tau = 0.9$ , where both have low SN and high SP. CTree and Weighted  $k$ -NN have the best balances between the sensitivity and specificity at about  $\tau = 0.1$ . For example, in M1, weighted  $k$ -NN has SN = 33.9%, SP = 87.9%, and CC = 69.9% at  $\tau = 0.9$  and has SN = 63.3%, SP = 66.4%, and CC = 65.9% at  $\tau = 0.1$ . In M2, weighted  $k$ -NN has SN = 70.1%, SP = 96.2%, and CC = 87.5% at  $\tau = 0.9$  and has SN = 87.2%, SP = 88.5%, and CC = 88.1% at  $\tau = 0.1$ . Figure 3 shows that the decision threshold has less impact on the sensitivity, specificity, and concordance when the separation between the two class means is large. It can be seen that the ranges of the sensitivity, specificity, and concordance in M2 are much smaller than the ranges in M1.

In summary, setting the decision threshold at  $p_1$  from the default of 0.5 or  $p_0$  appears to improve the balance between the sensitivity and specificity except for FLDA. Perhaps the outputs of the logistic regression, CTree, and weighted  $k$ -NN represent the probability of class membership; the decision boundary for FLDA is based on the mean under the equal variance model. Also, FLDA outperforms the other three methods this is because the data were generated from the normal models.

## Examples

We considered three data sets. The first two data sets are applications of SAR (structure-activity relationship) models to predict toxicologic effects of chemicals. The first data set is to predict animal liver carcinogenicity, and the second data set is to predict estrogen receptor binding activity. (Both data sets are available from the authors on request.) Both data sets consist of more than 200 predictors. The third data set is the public colon tumor data set [5] with 2000 predictors. We selected 32 highest ranked predictors based on t-statistic. Weighted  $k$ -NN method used  $k = 5$  for all data sets. This number was based on our empirical investigations; this value showed the most consistent results for all data sets.

## NCTR Liver Tumor Data Set

The NCTR liver tumor database was derived from Gold’s Carcinogenic Potency Database [21] of rodent bioassays. Rodent bioassays are conducted to assess carcinogenic effects of a chemical on humans. Because each study costs several million dollars and takes several years to complete, rodent bioassays are conducted on only a small fraction of the thousands of chemicals in use or in the environment. SAR models have been developed to predict potential genetic toxicants [3]. The SAR model can be applied to identify hazardous chemicals at low cost and to reduce the number of laboratory animal experiments. This example applies the SAR model to the prediction of animal liver carcinogenicity. The NCTR liver tumor data set consists of 282 liver carcinogens and 714 non-liver carcinogens. The SAR model was based on 282 descriptors (predictors) mostly generated by Cerius2 (Accelrys, Inc., San Diego, CA). The ratio of negatives to positives is about 2.5:1.

This data set has been analyzed by Young et al. [22] using the following four classification methods: multivariate adaptive regression spline, rough sets, support vector machines (SVM), and partial least square discriminant function. The reported concordance ranged from 54 to 71%, the sensitivity from 12 to 57% and specificity from 68 to 91%. Note that the concordances for this database are in the range between 24% and 79% reported in the Predictive Toxicology Challenge [1]. The SVM classifier had the best concordance of  $CC = 71\%$  with  $SN = 26\%$  and  $SP = 91\%$ . However, a ‘naive’ procedure which classifies all chemicals to be negative will have  $CC = 71.6\%$ ; this data set is used for illustrative purposes.

For the 32 selected predictors, 10-fold cross-validation with 100 different partitions was performed using decision thresholds of  $\{0.1, 0.2, \dots, 0.9\}$ , and the two sample proportions  $p_0$  threshold and  $p_1$  threshold. Table 2 shows the means and the standard deviations of sensitivity, specificity and concordance for the four classification methods. The results given in Table 2 are consistent with the simulation results: sensitivity (SN) decreases and specificity (SP) increases as the threshold increases. Both logistic regression and FLDA appear to be sensitive to the changes of decision threshold. The SN’s are greater than 90% at  $\tau = 0.1$  and the SP’s are 99% at  $\tau = 0.9$ . All four methods show that the concordances increases as  $\tau$  increases from 0.1 to 0.5. Logistic regression reaches the maximum concordance at  $\tau = 0.5$  and FLDA reaches the maximum at  $\tau = 0.7$ ; both concordances then slowly decrease. CTree is less influenced by the decision threshold. It reaches the maximum at about  $\tau = 0.8$ ;

the concordances are fairly constant for the range  $\tau$  between 0.2 and 0.9. Weighted  $k$ -NN improves the concordance gradually as  $\tau$  increases; the concordance reaches the maximum at  $\tau = 0.9$ .

Among the four classification methods, FLDA has the best concordance 74.3% (but, only slightly higher than the naive procedure), while the CTree has the lowest concordance, 64.3%. The concordances at the  $p_0$  threshold  $\tau = 0.717$  range from 64.1% (CTree) to 74.2% (FLDA). These values are less than 2% different from their respective maximums. As might be anticipated, the maximum concordances are accompanied by very low sensitivity (50% or less) and high specificity. In the context of predictive toxicology, it is important to have a high sensitivity because of health concerns.

### **NCTR Estrogen Activity Data Set**

The NCTR estrogen activity data set consists of 232 structurally diverse chemicals, of which 131 chemicals exhibit estrogen receptor binding activity and 101 are inactive in a competitive estrogen receptor binding assay [23]. The ratio of negatives to positives is 1:1.3. This data set has 202 descriptors (predictors) generated using the Cerius2 software for each chemical. This data set has been used to develop SAR models for predicting estrogen binding for prioritizing the chemicals for further testing.

Table 3 shows the means and the standard deviations of sensitivity, specificity and concordance for the four classification methods. Again, sensitivity decreases and specificity increases as the threshold increases. However, the ranges of SN and SP are much narrower than the ranges obtained from the liver tumor data set. Furthermore, the concordances in a classification method do not vary much. The ranges of the concordances from  $\tau = 0.4$  to 0.5 are less than 1%. Both  $p_0$  threshold, 0.435, and  $p_1$  threshold, 0.565, are in the interval of [0.4,0.6]. The concordances from the two thresholds are very close, but,  $p_1$  threshold does provide a better balance between sensitivity and specificity. Among the four methods, weighted  $k$ -NN appears to outperform the other three methods. Finally, when  $\tau$  is between 0.55 and 0.6, the weighted  $k$ -NN method will give at least 85% sensitivity and 75% specificity. In particular,  $\tau$  is at the  $p_1$  threshold, the weighted  $k$ -NN has SN = 85.8%, SP = 75.1%, and CC = 81.1%.

### **Colon Data Set**

The colon tumor data set [5] consists of 22 normal and 40 colon tumor tissue samples on 2000 human genes with highest minimal intensity across the 62 samples. The goal of the analysis is to discriminate the normal samples from the cancer samples based on the gene expression profiles. The colon cancer

data set has negative-to-positive ratio of about 1:1.8. There were more positive than negative samples. Table 4 shows the sensitivities, specificities, and concordances from the four classification methods.

The results for the colon data set are generally similar to the results from the liver tumor and estrogen data sets. The concordances do not vary much in the interval near the maximum concordance. The width of the interval varies among the classification methods. The concordances from logistic regression and FLDA are almost constant. Logistic regression, CTree, and FLDA have the maximum concordance at (or near) the  $p_0$  threshold. The weighted  $k$ -NN method has the best performance, and the CTree has the poorest performance. For  $\tau$  between 0.5 and 0.7, the weighted  $k$ -NN method will give at least 85% sensitivity and 85% specificity. For  $\tau = 0.5$ , the weighted  $k$ -NN has SN = 92.0%, SP = 86.5%, and CC = 90.1%. For  $\tau = 0.645$ , it has SN = 85.7%, SP = 90.6%, and CC = 87.4%.

Figure 4 is the ROC plots of four classification methods for the three data set. It can be seen that logistic regression and FLDA are the dominating models for the NCTR liver cancer data set since their ROC curves are completely above the ROC cruves for CTree and Weighted  $k$ -NN. Similarly, the weighted  $k$ -NN is superior for the NCTR estrogen and colon data sets. Note that the  $x$ -axis represents 1-specificity. Each point corresponds to a decision threshold from a classification method; different classification methods would have different decision thresholds.

For a comparison between the weighted  $k$ -NN and the standard  $k$ -NN based on the majority voting, the standard  $k$ -NN method has SN = 32.9%, SP = 84.7%, and CC = 70.1% for the NCTR liver data, has SN = 89.8%, SP = 60.0%, and CC = 76.8% for the NCTR estrogen data set, and has SN = 90.0%, SP = 90.9%, and CC = 90.3% for the colon data set.

## DISCUSSION AND CONCLUSION

This paper investigates the use of decision threshold to improve the performances on four classification methods. The performance of a classification method depends on the feature selection method, the number of predictors and selected predictors, and the classification method. Regardless of the feature selection method, different numbers of predictors will give different classification results. There is no theoretical estimation of the optimal number of selected predictors even for a given specific classification method. Thus, the optimal predictor set may depend on a classification method and can vary from data set to data set. Selection of predictor set can be conducted before the building of a

classification model, such as FLDA and  $k$ -NN, or be incorporated into model building, such as CTree and step-wise logistic regression. Cross-validation can be performed prior to feature selection (external cross-validation) or after feature selection (internal cross-validation). In the internal cross-validation, the same selected predictor set is used in each of training samples. On the other hand, in the external cross-validation, a new predictor set is selected for each training sample set. Ambroise and McLachlan [24] argued that the cross validation should include the feature selection in the training phase (external cross validation) to avoid selection bias in estimating prediction accuracy. However, the purpose of this paper is to investigate the effects of the decision threshold on the performances of the four classification methods. The same set of predictors is used in the evaluation. The cross-validation restricts to the selected 32 predictors.

Standard classification algorithms generally use a default decision threshold 0.5 and/or based on maximization of the classification concordance. This performance measure might be inappropriate if the sample class sizes are unequal or misclassification costs are different (Table 2). This paper considers a simple modification of the standard algorithm by changing the decision threshold in assigning class memberships for four classification methods. The simulation and example results show that the sensitivity and specificity decreases and increases, respectively, with the decision threshold, as expected. The concordance does not vary much in an interval near the maximum concordance. Thus, a change of decision threshold simply makes a tradeoff between the number of true positive and the number of true negative predictions. It has limited effects on the concordance in the interval near the maximum concordance. When the class sample sizes are approximately equal, the optimal decision threshold and balanced decision threshold are close to 0.5. The default threshold of 0.5 should have high concordance with a balance between sensitivity and specificity. When the class sizes are unequal, the interval between two sample ratios, which covers 0.5, appears generally to have the maximum concordance with the balanced sensitivity and specificity.

Decision threshold approach can only be applied to the classification methods that produce a quantitative output (e.g., logistic regression) from which different thresholds can be applied to assign class membership. Classification methods, such as  $k$ -NN or SVM, that produce a binary outcome to determine class membership, cannot be used for threshold adjustment. We used a weighted  $k$ -NN classification method by estimating the probability of the test sample in each class. The probability is calculated based on the relative distances between the test sample and the class samples in the nearest

neighbor. One challenge is the choice of distance metrics. The Gaussian kernel distance function is used in this paper based on empirical comparisons. Currently, we are developing a generalization of the SVM classifiers to allow for a decision threshold adjustment.

## References

- [1] Helma C and Kramer S. A survey of the predictive toxicology challenge 2000-2001. *Bioinformatics* 2003; 19: 1179-1182.
- [2] Rosenkranz HS and Cunningham AR. SAR modeling of unbalanced data sets. *SAR QSAR Environ Res* 2001; 12: 267-274.
- [3] Rosenkranz HS. SAR modeling of genotoxic phenomena: the consequence on predictive performance of deviation from a unity ratio of genotoxicants/non-genotoxicants. *Mutat Res* 2004; 559: 67-71.
- [4] Tong W, Xia Q, Hong H, Shi L, Fang H, and Perkins R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ Health Perspect* 2004; 112: 1249-1254.
- [5] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, and Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci* 1999; 96: 6745-6750.
- [6] Golub T, Slonim D, Tamayo P, Huard C, Gassenbeek M, Mesirov J, Coller H, Loh M, Downing J, Caligiuri M, Bloomfield C, and Lander E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science* 1999; 286: 531-537.
- [7] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, and Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* 2002; 359: 572-577.

- [8] Kubat M and Matwin S. Addressing the curse of imbalanced data sets: one-sided sampling, *Proceedings of the 14th International conference on Machine Learning* (Morgan Kaufmann); 1997. p.179-186.
- [9] Provost F. Machine learning from imbalanced data sets 101. *The AAAI'2000 Workshop on Imbalanced Data Sets*; Technical Report WS-00-05, 2000.
- [10] Chawla NV, Bowyer KW, Hall LO, and Kegelmeyer WP. Smote: synthetic minority over-sampling technique. *J Artif Intell Res* 2002; 16: 321-357.
- [11] Ling C and Li C. Data mining for direct marketing problems and solutions. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (KDD-98). New York, NY: AAAI Press; 1998. p.73-79.
- [12] Chen JJ, Tsai CT, Young JF, and Kodell RL. Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR and QSAR in Environ Res*; 2005, to appear.
- [13] Provost F, Fawcett T, and Kohavi R. The case against accuracy estimation for comparing classifiers. In *Proceedings of the Fifteenth International Conference on Machine Learning* (ICML-98); 1998. p.445-453.
- [14] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 1997; 30(6): 1145-1159.
- [15] Provost F and Fawcett T. Robust classification for imprecise environments. *Machine Learning* 2001; 42(3): 203-231.
- [16] Hastie T, Tibshirani RT, and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer; 2001.
- [17] Brieman L, Friedman J, Olshen RA, Stone CJ, Steinberg D, and Colla P. *CART: Classification and Regression Trees*. Stanford, CA, 1995.
- [18] Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; 240(4857): 1285-1293.
- [19] Zweig MH and Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993; 39(4): 561-577.

- [20] Huberty CJ. Applied Discriminant Analysis. John Wiley & Sons; 1994.
- [21] Gold LS, Slone TH, Manley NB, Garfinkel GB, Hudes ES, Rohrbach L, and Ames BN. The Carcinogenic Potency Database: Analyses of 4000 chronic animal cancer experiments published in the general literature and by the U.S. National Cancer Institute/National Toxicology Program. *Environ Health Perspect* 1991; 96:11-15.
- [22] Young JF, Tong W, Fang H, Xie Q, Pearce B, Hashemi R, Beger RD, Cheeseman MA, Chen JJ, Chang YI, and Kodell RL. Building an organ-specific carcinogenic database for SAR analyses. *J Toxicol Environ Health, Part A* 2004; 67:1363-1389.
- [23] Blair R, Fang H, Branham WS, Hass B, Dial SL, Moland CL, Tong W, Shi L, Perkins R, and Sheehan DM. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol Sci* 2000; 54:138-153.
- [24] Ambroise C and McLachlan G (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci* 2002; 99(10):6562-6566.

Table 1: For a given decision threshold  $\tau$ , the performance of a classification algorithm is summarized by the  $2 \times 2$  confusion matrix. The sensitivity is  $TP(\tau)/n_1$  and specificity is  $TN(\tau)/n_0$ . As  $\tau$  increases the sensitivity decreases and the specificity increases.

	Predicted Negative	Predicted Positive	Total
True Negative ( $Y = 0$ )	$TN(\tau)$	$FP(\tau)$	$n_0$
True Positive ( $Y = 1$ )	$FN(\tau)$	$TP(\tau)$	$n_1$
Total	$PN(\tau)$	$PP(\tau)$	$n$

Table 2: Sensitivity<sup>1</sup> (SN), specificity (SP), and concordance (CC) for the NCTR liver cancer data set from four classification methods using the decision thresholds ( $\tau$ ) of 0.1-0.9,  $n_0/n$  and  $n_1/n$ , where  $n_0 = 714$  and  $n_1 = 282$  are the numbers of normal and cancer chemicals, respectively.

$\tau$	Logistic Regression				CTree				FLDA				Weighted $k$ -NN			
	SN	SP	CC		SN	SP	CC		SN	SP	CC		SN	SP	CC	
.1	93.5 (0.64)	16.6 (0.54)	38.4 (0.04)		54.8 (3.37)	61.6 (3.27)	59.7 (1.91)		99.8 (0.20)	2.6 (0.23)	30.1 (0.18)		65.3 (1.02)	53.8 (0.88)	57.1 (0.84)	
.2	80.8 (0.89)	42.6 (0.65)	53.5 (0.05)		50.7 (2.76)	67.9 (1.71)	63.1 (1.30)		96.6 (0.58)	12.9 (0.46)	36.6 (0.37)		58.3 (1.72)	62.1.1 (0.73)	61.0 (0.88)	
.3	61.4 (1.12)	66.6 (0.69)	65.1 (0.06)		50.1 (2.71)	68.6 (1.58)	63.4 (1.26)		89.6 (0.57)	28.4 (0.58)	45.7 (0.43)		51.2 (1.49)	67.1 (0.80)	62.6 (0.57)	
.4	40.2 (1.02)	85.0 (0.47)	72.3 (0.04)		50.0 (2.69)	68.7 (1.58)	63.4 (1.26)		78.2 (0.99)	47.4 (0.71)	56.1 (0.56)		45.2 (1.45)	72.4 (0.93)	64.7 (0.67)	
.5	23.5 (0.94)	93.4 (0.36)	73.6 (0.04)		50.0 (2.73)	68.9 (1.60)	63.5 (1.27)		61.2 (1.03)	66.4 (0.65)	64.9 (0.58)		38.9 (0.61)	77.6 (0.73)	66.6 (0.56)	
.6	11.5 (0.70)	96.2 (0.26)	72.2 (0.03)		50.0 (2.61)	69.0 (1.49)	63.6 (1.25)		43.9 (0.91)	82.8 (0.50)	71.8 (0.48)		33.5 (1.62)	82.0 (0.91)	68.2 (0.74)	
.7	4.8 (0.52)	98.2 (0.20)	71.8 (0.02)		49.3 (2.66)	69.4 (1.50)	62.7 (1.23)		28.0 (0.97)	92.5 (0.46)	74.3 (0.48)		27.1 (1.12)	85.8 (0.94)	69.2 (0.86)	
.8	1.7 (0.47)	99.5 (0.16)	71.8 (0.02)		48.3 (2.81)	70.6 (1.47)	64.3 (1.20)		11.9 (0.68)	96.2 (0.26)	72.3 (0.28)		20.7 (0.56)	89.4 (0.59)	69.9 (0.38)	
.9	0.1 (0.01)	99.9 (0.10)	71.6 (0.01)		45.3 (2.93)	73.2 (1.53)	63.4 (1.23)		2.8 (0.47)	99.0 (0.16)	71.8 (0.17)		14.8 (0.89)	93.5 (0.49)	71.2 (0.45)	
$p_0$	4.1 (0.41)	98.5 (0.22)	71.8 (0.02)		48.8 (2.67)	70.1 (1.49)	64.1 (1.19)		25.2 (0.10)	93.6 (0.36)	74.2 (0.37)		26.5 (1.44)	86.1 (0.61)	69.3 (0.52)	
$p_1$	65.3 (1.03)	62.8 (0.71)	63.5 (0.06)		50.0 (2.70)	68.6 (1.58)	63.4 (1.26)		91.0 (0.56)	25.5 (0.57)	44.0 (0.43)		52.6 (1.45)	66.7 (0.94)	62.7 (0.96)	

1. Based on 10-fold cross validation with 100 different partitions

2.  $p_0$  threshold  $n_0/n = 0.717$

3.  $p_1$  threshold  $n_1/n = 0.283$

Table 3: Sensitivity<sup>1</sup> (SN), specificity (SP), and concordance (CC) for the NCTR estrogen data set from four classification methods using the decision thresholds ( $\tau$ ) of 0.1-0.9,  $n_0/n$  and  $n_1/n$ , where  $n_0 = 101$  and  $n_1 = 131$  are the numbers of normal and cancer chemicals, respectively.

$\tau$	Logistic Regression				CTree				FLDA				Weighted $k$ -NN			
	SN	SP	CC	SN	SP	CC	SN	SP	CC	SN	SP	CC	SN	SP	CC	
.1	83.7 (4.48)	54.3 (5.92)	70.9 (2.14)	84.2 (4.14)	56.9 (7.74)	74.5 (3.54)	96.4 (0.70)	36.0 (1.87)	70.1 (0.84)	94.7 (0.72)	53.8 (2.06)	76.9 (1.00)				
.2	81.8 (3.89)	58.5 (4.95)	71.7 (2.18)	82.2 (4.66)	61.3 (6.91)	74.8 (3.77)	93.2 (0.90)	46.0 (2.17)	72.7 (1.12)	93.7 (1.02)	62.7 (2.00)	80.2 (1.01)				
.3	80.4 (3.51)	61.9 (3.94)	72.3 (1.97)	80.8 (5.03)	62.8 (7.23)	74.4 (3.97)	90.6 (1.13)	55.0 (2.20)	75.1 (1.10)	92.7 (1.12)	65.1 (1.88)	80.7 (1.15)				
.4	79.4 (3.41)	64.6 (3.20)	73.0 (1.95)	80.6 (5.16)	64.9 (6.65)	75.1 (3.91)	86.8 (1.22)	62.9 (1.91)	76.4 (1.11)	89.3 (1.03)	70.0 (1.94)	80.7 (1.06)				
.5	78.0 (3.34)	67.2 (3.25)	73.3 (2.11)	80.6 (5.16)	65.0 (6.76)	75.1 (3.90)	82.8 (1.28)	68.4 (1.81)	76.5 (1.13)	87.0 (1.35)	73.2 (1.52)	81.0 (1.09)				
.6	76.1 (3.02)	69.4 (3.47)	73.2 (2.19)	80.6 (5.16)	65.0 (6.76)	75.1 (3.90)	78.6 (1.43)	73.7 (1.49)	76.5 (1.02)	85.5 (1.57)	75.6 (1.56)	81.2 (1.10)				
.7	74.1 (3.01)	71.9 (4.10)	73.1 (2.22)	80.5 (5.12)	65.0 (6.76)	75.0 (3.87)	72.0 (1.58)	77.8 (1.66)	74.5 (1.06)	82.2 (1.27)	78.3 (1.85)	80.5 (1.16)				
.8	70.2 (3.39)	74.2 (4.24)	72.0 (2.03)	80.5 (5.12)	65.0 (6.76)	75.0 (3.87)	62.7 (1.39)	83.7 (1.83)	71.9 (1.08)	78.0 (1.28)	80.9 (1.80)	79.2 (1.14)				
.9	64.4 (4.77)	76.8 (4.80)	69.8 (2.22)	79.5 (5.47)	66.0 (6.11)	74.7 (4.05)	44.6 (1.84)	91.1 (1.60)	64.9 (1.21)	73.7 (1.47)	83.6 (1.85)	78.0 (1.13)				
$p_0$	79.0 (3.43)	65.5 (3.31)	73.1 (2.01)	80.6 (5.13)	64.6 (6.91)	74.9 (3.97)	85.2 (1.06)	65.0 (1.88)	76.4 (1.03)	88.5 (1.17)	70.7 (1.70)	80.8 (0.99)				
$p_1$	76.8 (3.09)	68.6 (3.27)	73.2 (2.14)	80.6 (2.52)	65.0 (2.93)	75.0 (1.96)	80.1 (1.14)	71.9 (1.36)	76.5 (1.03)	85.8 (1.51)	75.1 (1.69)	81.1 (1.06)				

1. Based on 10-fold cross validation with 100 different partitions

2.  $p_0$  threshold  $n_1/n = 0.565$

3.  $p_1$  threshold  $n_0/n = 0.435$

Table 4: Sensitivity<sup>1</sup> (SN), specificity (SP), and concordance (CC) for the colon cancer data set from four classification methods using the decision thresholds ( $\tau$ ) of 0.1-0.9,  $n_0/n$  and  $n_1/n$ , where  $n_0 = 22$  and  $n_1 = 40$  are the numbers of inactive and active chemicals, respectively.

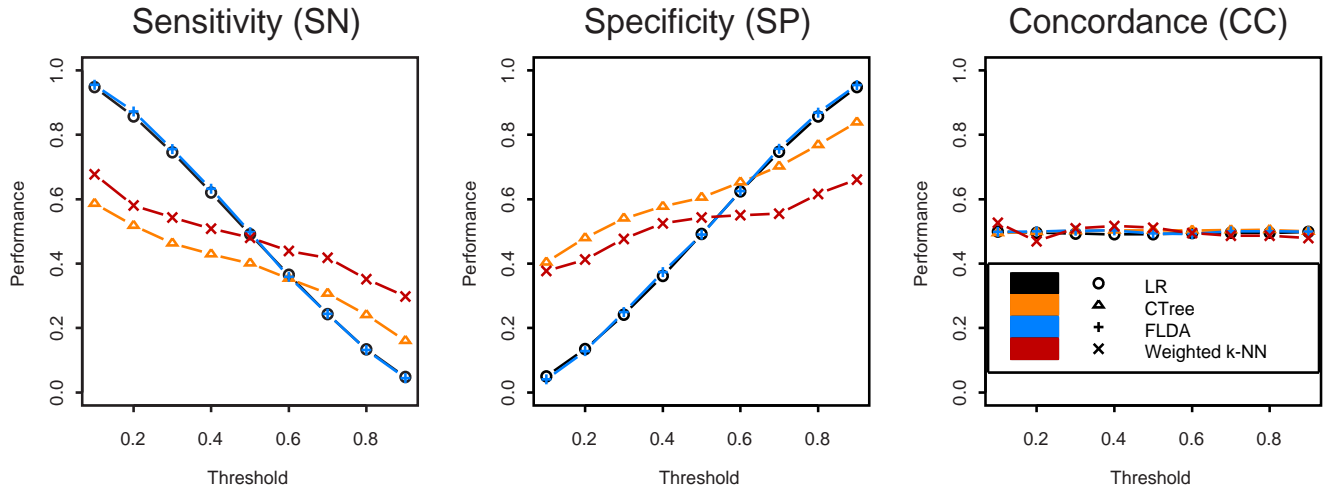
$\tau$	logistic Regression						C'Tree			FLDA			Weighted $k$ -NN		
	SN	SP	CC	SN	SP	CC	SN	SP	CC	SN	SP	CC	SN	SP	CC
.1	83.7 (4.46)	77.9 (7.51)	81.7 (3.98)	83.7 (4.39)	57.7 (7.87)	74.5 (3.53)	87.4 (3.22)	72.4 (6.32)	82.1 (3.35)	92.6 (0.43)	30.1 (5.24)	70.4 (1.87)			
.2	83.3 (4.48)	78.5 (7.53)	81.6 (3.92)	81.5 (4.86)	62.0 (6.94)	74.5 (3.77)	86.3 (3.29)	75.4 (6.84)	82.4 (3.33)	92.5 (0.)	55.2 (4.24)	79.3 (1.50)			
.3	83.1 (4.52)	78.8 (7.73)	81.5 (3.97)	80.3 (5.09)	63.8 (7.32)	74.4 (3.95)	85.2 (3.31)	76.6 (6.54)	82.1 (2.94)	92.5 (0.)	77.6 (3.84)	87.2 (1.36)			
.4	82.6 (4.56)	79.2 (7.70)	81.4 (4.00)	80.2 (5.13)	65.3 (6.70)	74.9 (3.84)	85.5 (3.23)	76.3 (6.50)	82.2 (3.16)	92.0 (1.10)	82.8 (2.30)	88.4 (1.15)			
.5	82.3 (4.62)	79.5 (7.71)	81.3 (4.06)	80.2 (5.13)	65.3 (6.70)	74.9 (3.84)	83.8 (3.36)	78.1 (6.38)	81.8 (3.26)	92.0 (1.01)	86.5 (0.78)	90.1 (0.73)			
.6	82.0 (4.58)	80.0 (7.86)	81.3 (4.03)	80.2 (5.10)	65.3 (6.70)	74.9 (3.81)	82.9 (3.45)	79.0 (7.04)	81.5 (3.58)	89.8 (1.04)	90.1 (1.72)	89.9 (0.94)			
.7	81.6 (4.69)	80.3 (7.81)	81.1 (4.08)	80.2 (5.10)	65.3 (6.70)	74.9 (3.81)	82.5 (3.58)	80.4 (6.58)	81.8 (3.49)	85.8 (1.31)	90.4 (1.48)	87.4 (0.98)			
.8	81.1 (4.69)	80.7 (7.90)	81.0 (4.10)	79.1 (5.64)	66.0 (6.41)	74.5 (4.03)	81.6 (3.45)	82.0 (6.88)	81.8 (3.46)	85.7 (1.10)	90.8 (1.36)	87.5 (0.91)			
.9	80.5 (4.80)	81.6 (7.63)	80.9 (3.96)	69.1 (9.13)	69.6 (7.19)	69.3 (5.32)	80.3 (3.62)	84.4 (5.38)	81.7 (3.12)	83.3 (1.63)	95.2 (1.00)	87.6 (1.08)			
$p_0$	82.9 (4.62)	79.1 (7.62)	81.5 (4.05)	80.2 (5.10)	65.3 (6.70)	74.9 (3.81)	85.5 (3.37)	76.3 (6.75)	82.2 (3.31)	92.5 (1.01)	78.7 (0.31)	87.6 (1.09)			
$p_1$	81.9 (4.63)	80.1 (7.82)	81.2 (4.02)	80.2 (5.10)	65.3 (6.70)	74.9 (3.81)	83.1 (3.46)	79.9 (6.69)	81.9 (3.49)	85.7 (1.22)	90.6 (1.17)	87.4 (0.92)			

1. Based on 10-fold cross validation with 100 different partitions

2.  $p_0$  threshold  $n_1/n = 0.645$

3.  $p_1$  threshold  $n_0/n = 0.355$

M0: Equal class size:  $n_0=56$  and  $n_1=56$



M0: Unequal class size:  $n_0=112$  and  $n_1=56$

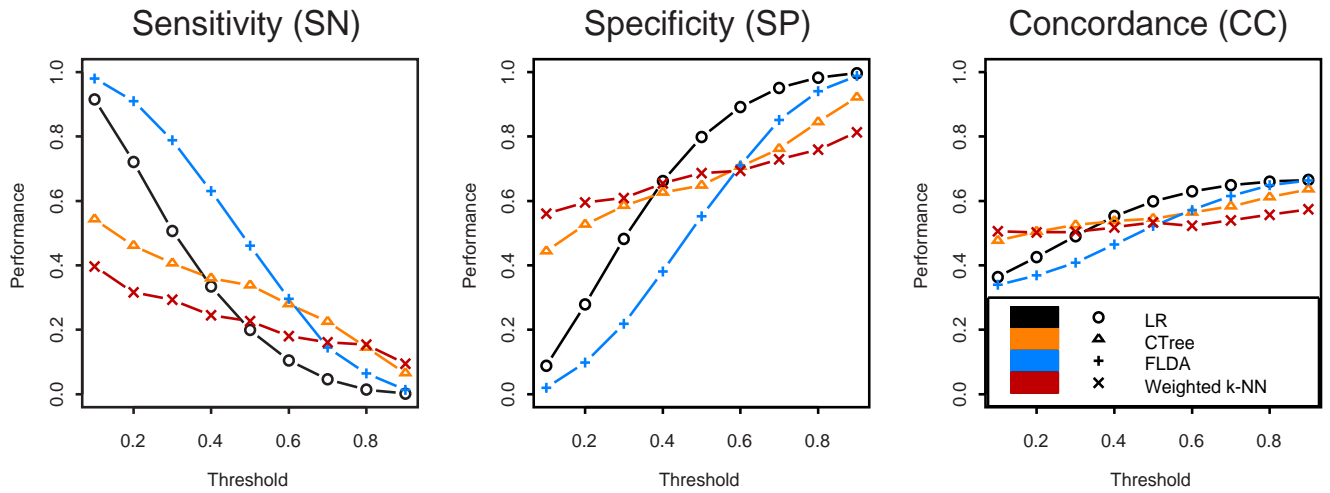
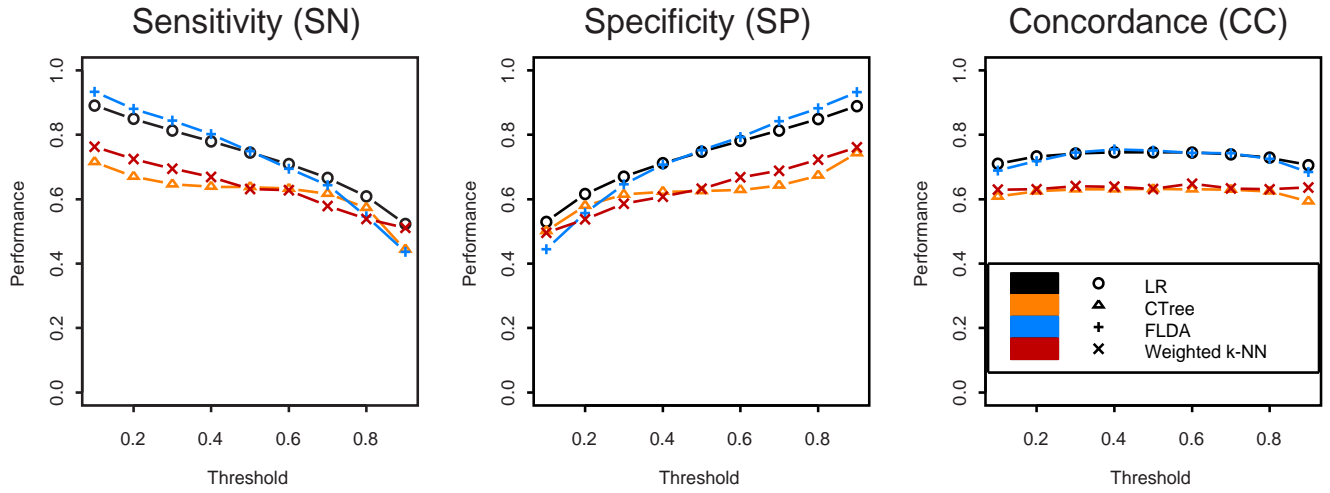


Figure 1: Plots of sensitivity, specificity, and concordance of M0 for the equal class size (upper panel) and unequal class sizes (lower panel), where M0: class 0  $\sim N(0, .2^2)$  and class 1  $\sim N(0, .2^2)$ .

M1: Equal class size:  $n_0=56$  and  $n_1=56$



M2: Equal class size:  $n_0=56$  and  $n_1=56$

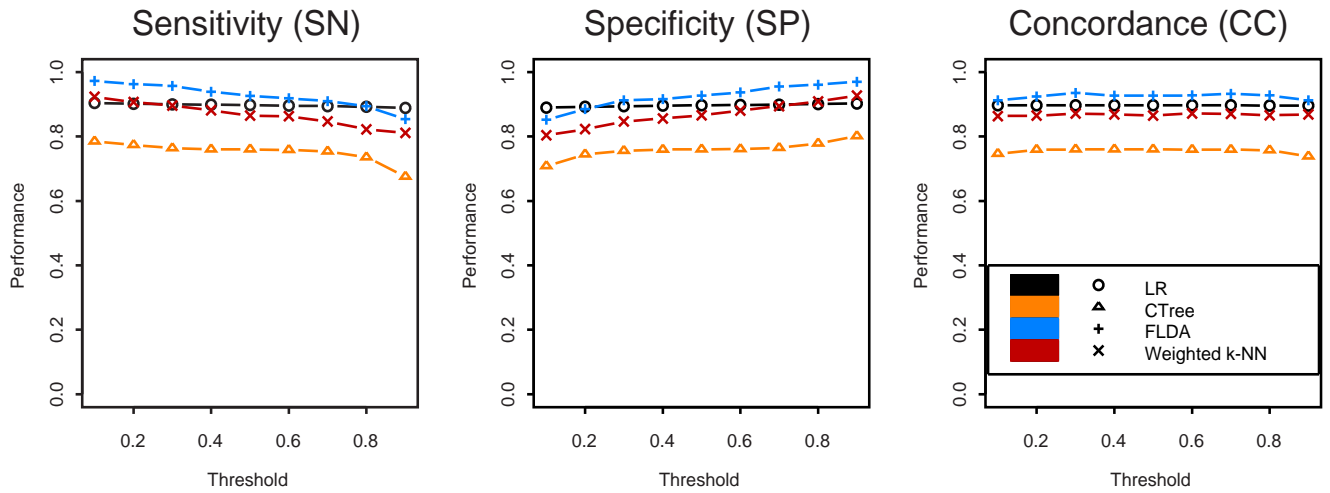
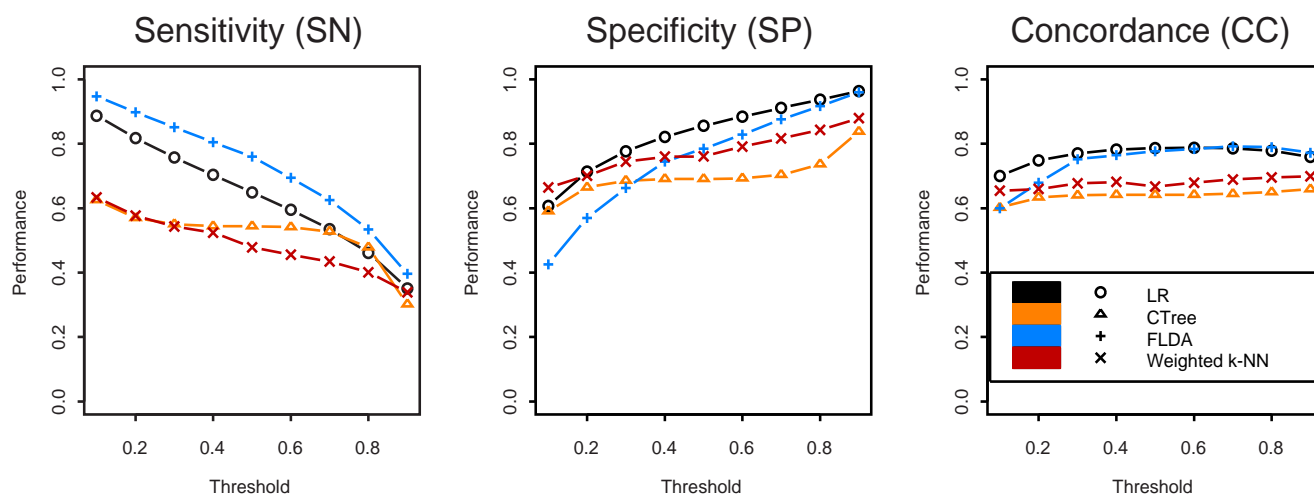


Figure 2: Plots of sensitivity, specificity, and concordance of M1 (upper panel) and M2 (lower panel) for equal class size, where M1: class 0  $\sim N(0, .2^2)$  and class 1  $\sim N(.1, .2^2)$  and M2: class 0  $\sim N(0, .2^2)$  and class 1  $\sim N(.2, .2^2)$ .

M1: Unequal class size:  $n_0=112$  and  $n_1=56$



M2: Unequal class size:  $n_0=112$  and  $n_1=56$

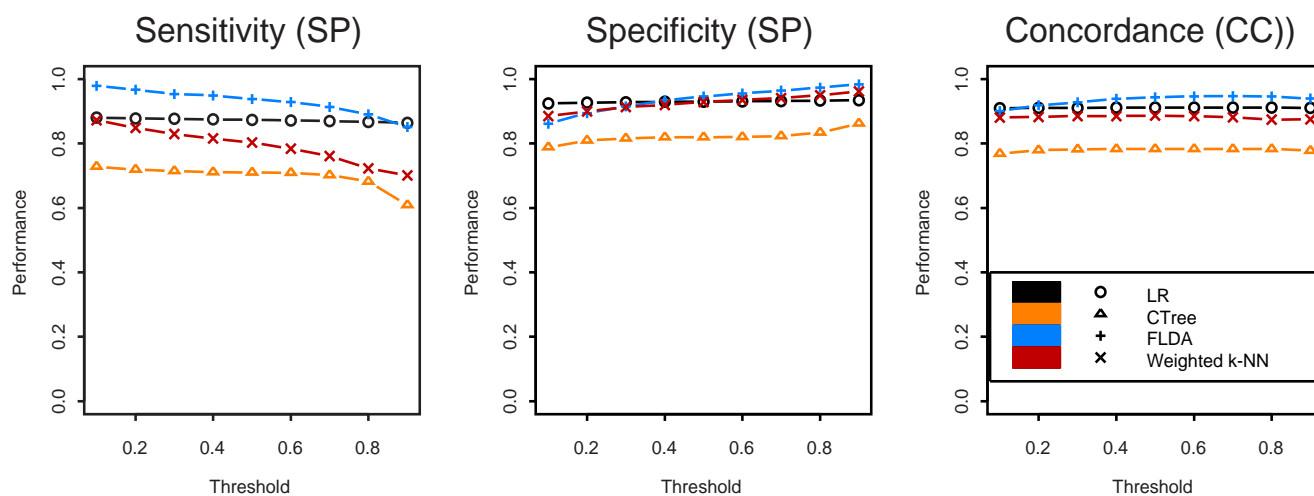


Figure 3: Plots of the sensitivity, specificity, and concordance of M1 (upper panel) and M2 (lower panel) for unequal class sizes, where M1: class 0  $\sim N(0, .2^2)$  and class 1  $\sim N(.1, .2^2)$  and M2: class 0  $\sim N(0, .2^2)$  and class 1  $\sim N(.2, .2^2)$ .

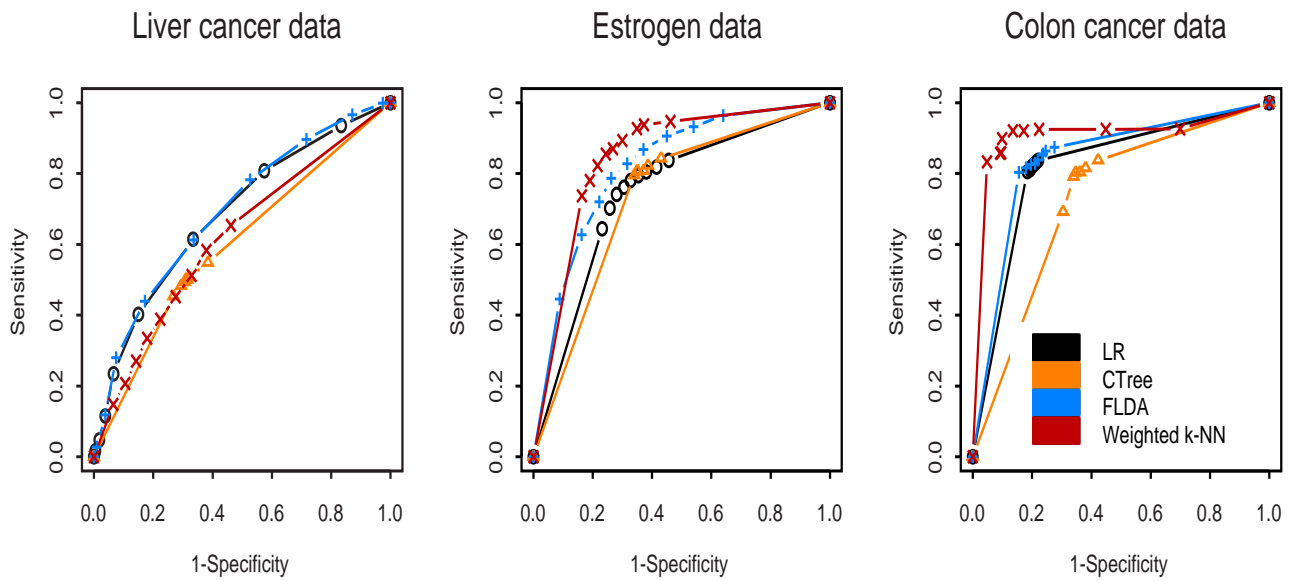


Figure 4: ROC Plots of the NCTR liver tumor, estrogen activity, and colon cancer data sets.