ORIGINAL RESEARCH

Cost-sensitive performance metric for comparing multiple ordinal classifiers

Nysia I. George[1], Tzu-Pin Lu[1,2], Ching-Wei Chang*[1]

[1] Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, FDA, Jefferson, Arkansas, USA
[2] Institute of Epidemiology and Preventive Medicine, Department of Public Health, National Taiwan University, Taiwan

Received: October 29, 2015 **Accepted:** January 3, 2016 **Online Published:** January 15, 2016
DOI: 10.5430/air.v5n1p135 **URL:** http://dx.doi.org/10.5430/air.v5n1p135

ABSTRACT

The surge of interest in personalized and precision medicine during recent years has increased the application of ordinal classification problems in biomedical science. Currently, accuracy, Kendall's τ_b, and average mean absolute error are three commonly used metrics for evaluating the effectiveness of an ordinal classifier. Although there are benefits to each, no single metric considers the benefits of predictive accuracy with the tradeoffs of misclassification cost. In addition, decision analysis that considers pairwise analysis of the metrics is not trivial due to inconsistent findings. A new cost-sensitive metric is proposed to find the optimal tradeoff between the two most critical performance measures of a classification task – accuracy and cost. The proposed method accounts for an inherent ordinal data structure, total misclassification cost of a classifier, and imbalanced class distribution. The strengths of the new methodology are demonstrated through analyses of three real cancer datasets and four simulation studies. The new cost-sensitive metric proved better performance in its ability to identify the best ordinal classifier for a given analysis. The performance metric devised in this study provides a comprehensive tool for comparative analysis of multiple (and competing) ordinal classifiers. Consideration of the tradeoff between accuracy and misclassification cost in decisions regarding ordinal classification problems is imperative in real-world application. The work presented here is a precursor to the possibility of incorporating the proposed metric into a prediction modeling algorithm for ordinal data as a means of integrating misclassification cost in final model selection.

Key Words: Ordinal classification, Classification, Ordinal data, Misclassification, Performance metric, Cost-sensitive

1. BACKGROUND

Ordinal classification is a multiclass classification problem where objects are classified into groups that have an inherent, natural ordering. In recent years, the prevalence of ordinal classification problems in clinical cancer research has dramatically increased due to technological advancements in genomics research coupled with the rise of personalized and precision medicine.[1,2] For example, typically, real-valued attributes are used to classify a patient into one of several ranked target classes such as health status, cancer stage, tumor grade, risk prediction, or survival prognosis.

In practice, an integral task of a single classification study is comparing the utility of several classifiers based on a predefined performance metric. Choosing the best classifier for a given dataset often involves consideration of multiple input variables (or feature sets) and multiple machine learning algorithms that employ different decision criteria. Measures to assess the comparative performance of multiple

*Correspondence: Ching-Wei Chang; Email: wei0917@gmail.com; Address: 3900 NCTR Road HFT-20, Jefferson, AR 72079, USA.

classifiers for two-class problems are well defined. In supervised learning for multiclass prediction, typically performance metrics suitable for binary classification problems are modified to accommodate multiclass problems (see Refs.[3,4] for two independent reviews). In turn, performance measures for nominal classification problems are generally applied to datasets with ordinal class structures.[5,6] However, valuable information is contained in the implicit ordered relationship between classes and should not be disregarded in evaluating the comparative performance of multiple ordinal classifiers.

Classification accuracy (*Acc*), the proportion of correctly predicted objects/samples, is the most commonly used performance metric to evaluate multiple classifiers. However, accuracy alone is insufficient since it does not incorporate a penalty for decision-theory misclassification costs. Moreover, Acc as a single metric is not effective for datasets with imbalanced class distributions or ordinal datasets.[5,7] Alternatively, performance metrics that measure the degree of loss between predicted and true class membership have been proposed as a supplement to Acc. These include mean absolute error (MAE), mean squared error (MSE), and average MAE across classes (AMAE). Although each of the aforementioned statistics attributes a higher cost to misclassifying an object into a more distant class, each measure depends heavily on the values used to label each class. In order to avoid the quantitative influence of arbitrary class labels, Kendall's τ_b[8] is generally used to assess the nonparametric association between true and predicted class labels. Other proposed methods consist of modifications of the receiver operating characteristic (ROC) curves,[9] additional ordinal association coefficients,[10,11] and variants of rank-order correlation.[12]

Several algorithms have been developed to classify ordinal data;[2,13–17] some have been specifically devised to incorporate the cost of misclassification error.[18,19] While these methods may improve the precision of an ordinal classifier, appropriate evaluation metrics for comparative studies designed to determine the best classifier among multiple classifiers are still lacking. In this manuscript, we develop a new performance metric to evaluate the performance of multiple multiclass classifiers for a given ordinal dataset. The proposed metric, which incorporates a cost for misclassification and a weighting factor for class distribution, is guided by the tradeoff between accuracy and misclassification cost. Several real cancer datasets are used to present an analysis of the proposed metric compared to Acc, AMAE, and Kendall's τ_b as comparative performance metrics. In addition, a study of simulated confusion matrices with fixed accuracy is used to evaluate the effectiveness of the aforementioned ordinal metrics.

2. METHODS

2.1 Multiclass prediction performance metrics

There are many metrics to evaluate the efficiency of a multiclass classifier. The list of metrics suitable for ordinal data is much more limited. In general, three types of performance metrics are used to evaluate ordinal classifiers. These measures include assessing overall accuracy, misclassification error that accounts for the inherent order between classes, and rank association. The goal of each is to measure how well predicted class labels for N samples, $\{\hat{y}_1, \hat{y}_2 \cdots, \hat{y}_N\}$, correspond to true class labels, $\{y_1, y_2, \cdots, y_N\}$. The most popular metric for assessing a classifier's performance is Acc, which measures the proportion of correct classification:

$$Acc = \frac{1}{N}\sum_{i=1}^{N} I_{\{\hat{y}_i = y_i\}} \tag{1}$$

where $I_{\{\}}$ is the indicator function with value 1 if $\hat{y}_i = y_i$ and 0 otherwise and $0 \leq Acc \leq 1$. While accuracy is simplistic in nature and provides a general overview of a classifier's performance, it ignores the cost of misclassification. On the other hand, metrics such as MAE and MSE solely account for misclassification by measuring the degree of error between true and predicted labels through a loss function. If class sizes are imbalanced, which is typical in most cancer classification studies, computing a weighted average of MAE across all classes (AMAE)[5] is more robust.

$$AMAE = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{n_m}\sum_{i=1}^{n_m}|y_i - \hat{y}_i|\ I_{\{y_i \in class\, m\}} \tag{2}$$

where M is the number of classes, n_m is the number of samples in the m^{th} class, $I_{\{\}}$ is the indicator function with value 1 if $y_i \in classm$ and 0 otherwise, and $AMAE \in [0, M-1]$. Alternatively, the association between $\{y\}$ and $\{\hat{y}\}$ can be measured using rank order correlation statistics. The most widely used rank correlation coefficient is Kendall's correlation coefficient[8] $\tau_b \in [-1, 1]$:

$$\tau_b = \frac{C - D}{\sqrt{(C + D - T_t)(C + D - T_p)}} \tag{3}$$

where C = the number of concordant pairs, D = the number of discordant pairs, T_t = the number of tied pairs in the true class membership, and T_p = the number of tied pairs in the predicted class membership. Unlike metrics that measure error via a loss function, Kendall's τ_b simply assesses the order relation between true and predicted class labels. Thus, it is not affected by values chosen to represent class labels, which are often arbitrary.

2.2 Formulating the cost matrix

Let $C_{M\times M}$ denote the cost matrix associated with predicting the class membership of N test objects into one of M classes for a single experiment. Let the rows and columns of $C_{M\times M}$ denote the predicted and true class membership, respectively. Thus, an entry $c_{i,j}$ denotes the cost of misclassifying a "class j" object into "class i".

In this study, misclassification cost denotes the cost attributed to a single classifier rather than simply the penalty or cost of misclassifying a single object. Misclassification cost, as presented herein, uses information from the distribution of classes and domain knowledge about the ordinal class structure to derive the total misclassification cost associated with a classifier. More specifically, cost is measured by the product of two factors – the inverse probability of misclassifying an object into a specific class given that the object has been misclassified and absolute deviation between true and predicted class membership. The first component of the cost term ensures that cost is weighted by class size so that prediction error for rare classes is not masked by the majority class. The second component is a linear absolute value loss function, which measures the distance between predicted and true class labels. This component could also indicate the cost unit from domain knowledge and may be modified based on study need. No cost is attributed to correct classification; therefore, all diagonal entries of $C_{M\times M}$ are set to $c_{k,k}$ for $k = 1, \cdots, M$. The off-diagonal entries of $C_{M\times M}$ are computed as:

$$c_{i,j} = \frac{\sum_{i\neq j}^{M} n_i}{n_i}\left|j-i\right|, \quad \text{for } i \neq j,\ i = 1,\ldots,M, \text{ and } j = 1,\ldots,M \tag{4}$$

with n_i representing the number of training samples in "class i".

As an example, we present the cost matrix C of an imbalanced ordinal dataset with 3 classes of size 10, 20, and 70. C is computed as the element-wise product of a matrix denoting misclassification cost given the inverse probability of misclassifying an object into a specific class given that the object has been misclassified and a symmetric matrix representing the linear absolute value loss function:

$$C = \begin{bmatrix} 0 & 8 & 3 \\ 4.5 & 0 & 1.5 \\ 90/70 & 80/70 & 0 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{bmatrix} \approx \begin{bmatrix} 0 & 8 & 6 \\ 4.5 & 0 & 1.5 \\ 2.6 & 1.1 & 0 \end{bmatrix}$$

Based on the given class distribution, the probability of misclassification into a given class, and the distance between true and predicted class membership, entry $c_{1,2}$ is attributed highest cost.

2.2.1 *Total misclassification cost of a classifier*

Let $F_{M\times M}$ represent the confusion matrix of a single classifier for an M-class classification problem. The confusion matrix summarizes the performance of a classifier by displaying how all N samples are distributed across predicted (rows) and true (columns) class membership. For example, $f_{1,1}$ represents the total number of "class 1" samples that were correctly classified, whereas $f_{1,2}$ represents the total number of "class 2" samples that were incorrectly classified as "class 1". By considering the cost matrix and the confusion matrix for a classifier, total misclassification cost (TC) of a supervised learning algorithm is represented as:

$$TC = \sum_{i=1}^{M}\sum_{j=1}^{M} c_{i,j} * f_{i,j} = trace\left(CF^T\right) \tag{5}$$

2.2.2 *Estimating maximum total misclassification cost of a classifier*

We utilized ROC-based methods to construct the proposed method, which identifies the best classifier by considering the optimal tradeoff between the cost and accuracy of multiple classifiers. This approach requires transformation of total misclassification cost to the [0,1] domain, which can be achieved by reporting TC as a percentage of the maximum TC of a single classification problem. Maximum TC is computed by fixing the cost matrix C and allowing the confusion matrix F_x to be a random variable (in essence, allowing for varying observations of a classifier's performance). The value of maximum TC was obtained by optimizing the function:

$$\text{maxTC} = \arg\max_{F_{\mathbf{x}}} \sum_{i=1}\sum_{i=1} \mathrm{c}_{\mathrm{i,j}} * f_{\mathbf{x}_{\mathrm{i,j}}} \tag{6}$$

subject to the constraints:

$$\begin{aligned} &\sum_{i=1}^{M}\sum_{j=1}^{M} f_{\mathbf{x}_{\mathbf{i,j}}} = N \\ &\sum_{i=1}^{M} f_{\mathbf{x}_{\mathbf{i,j}}} = n_j, \quad j = 1,\ldots,M \\ &f_{\mathbf{x}_{\mathbf{i,j}}} \geq 0, \qquad \forall\, i \neq j \\ &f_{\mathbf{x}_{\mathbf{i,i}}} = 0, \qquad \forall\, i = j. \end{aligned}$$

The constrOptim[20] function in R (`http://www.r-project.org`) was used to determine the unique solution of maxTC subject to the specified constraints.

2.3 The tradeoff between accuracy and cost

Let Acc and $MC = TC/\max TC$ denote the accuracy and misclassification cost, respectively, of a single classifier. The $[0,1]\times[0,1]$ grid space for evaluating the comparative performance of multiple classifiers is provided in Figure 1. Ideally, the best classifier would achieve highest Acc and lowest MC, resulting in an (Acc, MC) coordinate closest to (1, 0). The distance between a classifier's (Acc, MC) coordinate and the ideal coordinate of (1, 0) is represented by:

$$d = \sqrt{(1-Acc)^2 + MC^2} \tag{7}$$

where $d \in [0, \sqrt{2}]$. In this study, the classifier that achieves minimum distance d is selected as the classifier with superior performance. In the case of tied d statistics, the superior classifier is identified as the point that maximizes the vertical distance from the line of random chance. This distanced is measured as $|Acc + MC - 1|/\sqrt{2}$.

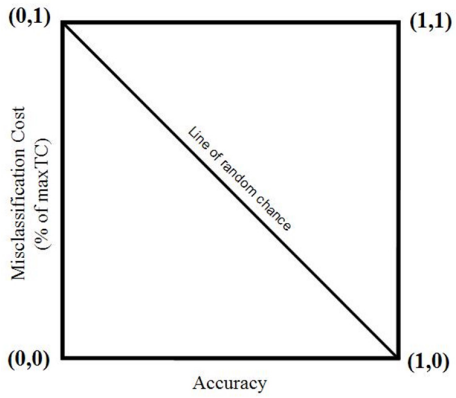

Figure 1. $[0,1]\times[0,1]$ performance grid for evaluating the comparative performance of multiple classifiers by balancing misclassification cost and accuracy

3. Results and Discussion

3.1 Analysis of three cancer datasets

We evaluated the performance of the proposed methodology using three cancer datasets with ordinal class structure. Each ordinal dataset had $3 \le M \le 5$ classes. In this paper, the three datasets are referred to as: colon cancer (downloaded from the GEO database [GSE17536]),[21] lung cancer (GSE19804),[22] and ovarian cancer[23] (obtained from the TCGA database). There are multiple and competing ways to derive a gene expression classifier. The subsequent analysis carried out for each dataset presents a different application of how the proposed methodology may be implemented in ordinal classification decision theory. It is important to note that for the purposes of this paper, we do not emphasize the methodology behind building a classifier or assert that the methods used in this paper are superior. Rather, the different applications simply provide a platform for comparing multiple ordinal classifiers, which is the primary focus of the current work.

3.1.1 *Colon cancer dataset*

Gene expression profiles from colon cancer patients categorized by 4 American Joint Committee on Cancer (AJCC) stages were downloaded from the Gene Expression Omnibus (GEO) database (GSE17536).[21] AJCC stages were available for 177 patients from the Moffitt Cancer Center, resulting in 24, 57, 57, and 39 patients with AJCC stages of I-IV respectively. For illustrative purposes, we applied a support vector machine (SVM) algorithm and two ordinal classification algorithms that were available in R to the dataset. The SVM model, which uses hyperplanes to optimize the linear separation between classes, was implemented using the e1071 package[24] in R. The two ordinal prediction models were constructed using the R packages glmnetcr[14] and rpartScore.[16] In glmnetcr, ordinal response data is modeled with an L_1 penalized continuation ratio model. rpartScore builds classification and regression trees for ordinal response categories. Classification trees are constructed by a user-specified ordinal impurity function and are pruned by a user-specified measure of predictive performance. Both ordinal classification algorithms were designed for high-dimensional data and can accommodate "large p, small n" datasets. All three classification models were built using a reduced dataset consisting of the 100 top-ranked genes determined by ANOVA F-test statistics. For simplicity, each prediction model was trained and tested on the same Moffitt Cancer Center samples. Confusion matrices presenting the results of each algorithm are presented in Table 1(a).

According to all ordinal performance metrics, comparative analyses of the three classifiers (presented in Table 2) demonstrate that SVM markedly dominates both ordinal classifiers. Not only does SVM achieve highest accuracy, but it also has the lowest AMAE and the highest rank correlation. Naturally, the identification of SVM as the superior classifier holds when assessing our proposed metric d, $i.e.$ SVM also attains minimum d (see Figure 2(a)).

3.1.2 *Lung cancer dataset*

Microarray gene expression from 56 lung adenocarcinoma patients (GEO; GSE19804)[22] with varying stages of disease was used to evaluate the performance of selected performance metrics. Stage information of these patients was encoded into three categories. Among the 56 patients, only one sample (103T) was stage 4 and thus was grouped with stage 3 patients. In total, there were 31 stage 1 patients, 12 stage 2 patients and 13 stage 3 patients. For each gene, a

linear regression model was performed to evaluate whether its expression level was associated with patients' stage. This analysis identified 24 significant genes ($p < .0001$), which were subsequently used to develop a classifier using the diagonal linear discriminant analysis[25] prediction algorithm. A leave-one-out cross-validation procedure was used to predict stage information for each patient. In addition, we shuffled the resulting confusion matrix to create a second set of predicted results with the same accuracy. To establish a null baseline for comparison, we randomly selected 24 genes from the original gene pool in GSE19804 and developed a corresponding prediction model to predict stage information for each patient.

Table 1. Confusion matrices of classifiers for (a) colon cancer patients classified by 4 AJCC stages (GSE17536), (b) lung adenocarcinoma patients classified by 3 ordered stages of disease (GSE19804), and (c) TCGA ovarian cancer patients classified by 5 stages of disease.

(a)

$$\text{svm} = \begin{bmatrix} 21 & 0 & 0 & 0 \\ 1 & 54 & 2 & 2 \\ 2 & 3 & 55 & 2 \\ 0 & 0 & 0 & 35 \end{bmatrix} \quad \text{glmnetcr} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 24 & 57 & 57 & 39 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{rpartScore} = \begin{bmatrix} 15 & 2 & 1 & 0 \\ 8 & 46 & 6 & 0 \\ 1 & 8 & 47 & 14 \\ 0 & 1 & 3 & 25 \end{bmatrix}$$

(b)

$$\text{Sig24} = \begin{bmatrix} 26 & 3 & 0 \\ 4 & 5 & 3 \\ 1 & 4 & 10 \end{bmatrix} \quad \text{Shuffle24} = \begin{bmatrix} 16 & 0 & 0 \\ 4 & 12 & 0 \\ 11 & 0 & 13 \end{bmatrix} \quad \text{Rand24} = \begin{bmatrix} 18 & 10 & 6 \\ 8 & 0 & 1 \\ 5 & 2 & 6 \end{bmatrix}$$

(c)

$$p\_1\text{E-}8 = \begin{bmatrix} 7 & 8 & 5 & 53 & 2 \\ 4 & 2 & 3 & 44 & 5 \\ 0 & 1 & 1 & 7 & 2 \\ 5 & 19 & 21 & 242 & 58 \\ 0 & 0 & 2 & 71 & 17 \end{bmatrix} \quad p\_1\text{E-}7 = \begin{bmatrix} 10 & 5 & 10 & 83 & 3 \\ 0 & 6 & 3 & 18 & 7 \\ 4 & 6 & 9 & 52 & 17 \\ 2 & 11 & 8 & 143 & 29 \\ 0 & 2 & 2 & 121 & 28 \end{bmatrix}$$

$$p\_1\text{E-}6 = \begin{bmatrix} 10 & 5 & 9 & 71 & 3 \\ 3 & 10 & 7 & 28 & 2 \\ 2 & 6 & 8 & 71 & 21 \\ 1 & 7 & 6 & 105 & 20 \\ 0 & 2 & 2 & 142 & 38 \end{bmatrix} \quad p\_1\text{E-}5 = \begin{bmatrix} 8 & 6 & 5 & 54 & 1 \\ 4 & 12 & 5 & 26 & 1 \\ 4 & 5 & 16 & 105 & 24 \\ 0 & 5 & 3 & 86 & 17 \\ 0 & 2 & 3 & 146 & 41 \end{bmatrix}$$

Confusion matrices resulting from prediction using the 24 significant predictor genes (Sig24), shuffling the confusion matrix of Sig24 (Shuffle24) but maintaining the same accuracy, and prediction using 24 randomly selected predictor genes (Rand24) are given in Table 1(b). As expected, the classifier generated by Rand24 performs worst according to all four performance metrics. Since accuracy is the same for Sig24 and Shuffle24, a comparative analysis of the two must be guided by some other performance metric. Unfortunately, AMAE and τ_b, two metrics often recommended for ordinal datasets, present conflicting results. Shuffle24 has smaller misclassification error; however, Sig24 has higher rank correlation. On the other hand, our proposed metric d, which assesses the balance between accuracy and misclassification cost, clearly indicates that Sig24 is the better classifier of the two (see Table 2(b), Figure 2(b)).

3.1.3 *Ovarian cancer dataset*

All four ordinal classification performance metrics were also evaluated on an ovarian cancer dataset downloaded from The Cancer Genome Atlas Database.[23] Data was obtained for 579 ovarian cancer patients who were classified into five different groups based on their stage information (Group 1: stage IA-IC, n=16; Group 2: stage IIA-IIC, n=30; Group 3: stage IIIA-IIIB, n=32; Group 4: stage IIIC, n=417; Group 5: stage IV, n=84). For each gene, an ANOVA test was performed to investigate whether its expression level was associated with stage of disease. Four different significant gene sets were identified by using four distinct p-value thresholds ($p < 10^{-8}, p < 10^{-7}, p < 10^{-6}$, and $p < 10^{-5}$). Similar to the analysis for the lung cancer dataset, a leave-one-out cross-validation procedure and the DLDA method were utilized to predict stage of ovarian cancer for each patient. Confusion matrices showing classification results for the four classifiers produced by the varying gene sets are presented in Table 1(c).

Similar to the lung cancer dataset, our proposed statistic, d, favors the classifier with highest accuracy (see Figure 2(c)). Interestingly enough, p_1E-8 has the highest AMAE and is tied with p_1E-7 for the lowest measure of τ_b, which would suggest that the classifier using a p-value threshold of p_1E-8 has poorest performance. However, p_1E-8 is a clear winner in this particular comparative analysis because not only does it attain highest accuracy, it also has the lowest misclassification cost. This assessment/decision is not trivial without the added information of misclassification cost.

3.2 Simulation study

Four simulation studies were conducted to further evaluate the comparative performance of d, AMAE, and Kendall's τ_b as metrics for comparing multiple classifiers. For each simulation study, we simulate the performance of a classification model by simulating confusion matrices (*i.e.* the classified outcome rather than raw data) and evaluate the effectiveness of each metric when comparing two confusion matrices with equal accuracy. Naturally, Acc ceases to be a useful metric under the setting of equal predictive accuracy. Moreover, the goal of the simulation study was to assess how often the ordinal performance metrics, d, AMAE, and Kendall's τ_b, coincided with the preferred (*i.e.* lower misclassification cost) classifier when accuracy was fixed. In general, 4 or 5 class confusion matrices were simulated based on designs of either fixed class-wise accuracy ($Acc_i = Acc$ for $i = 1, \cdots, M$) or fixed overall accuracy. A total of 10,000 pairs of confusion matrices were evaluated for each simulation scenario. We reported the percent of instances (out of 10,000) where d, AMAE, and Kendall's τ_b favor the confusion matrix with lower cost.

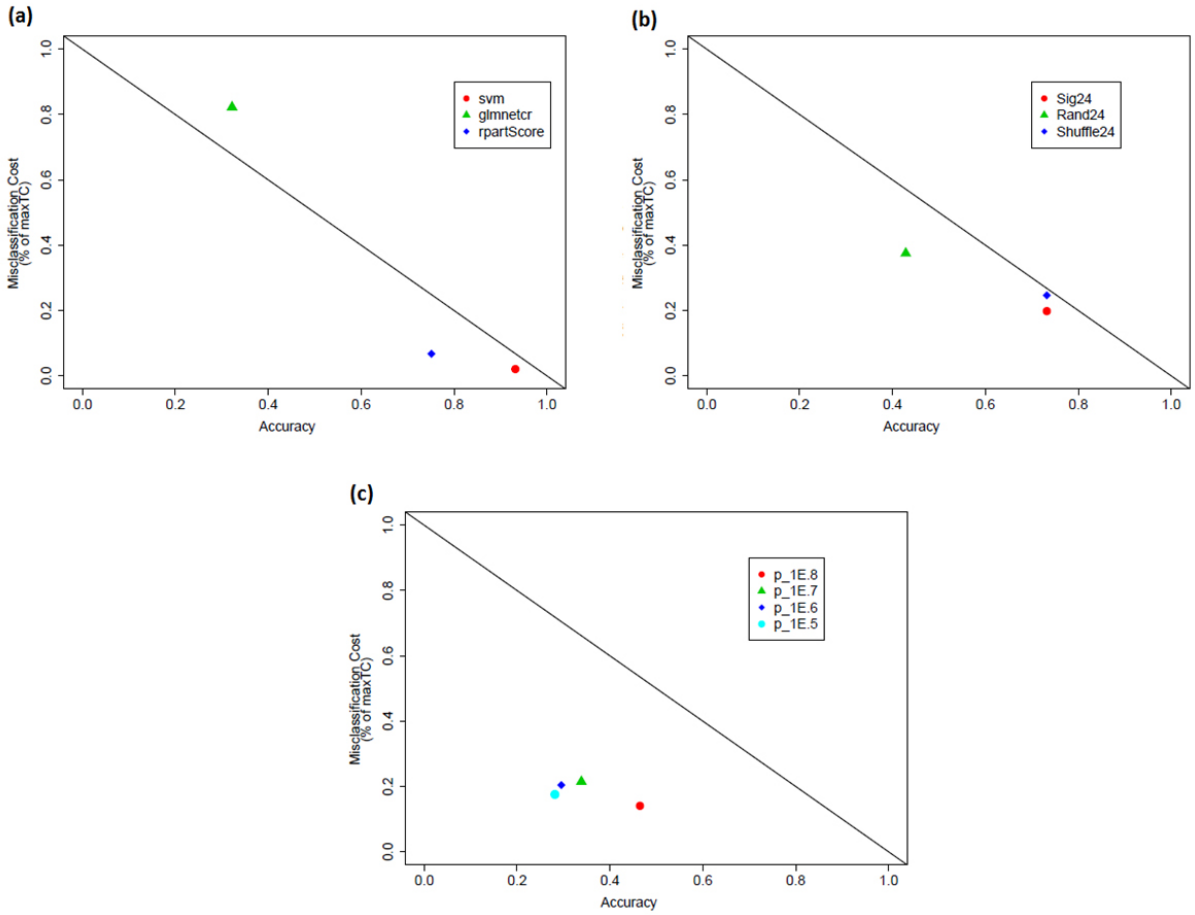

Figure 2. Performance grid of misclassification cost (MC) versus accuracy for the (a) colon cancer (GSE17536), (b) lung cancer (GSE19804), and (c) TCGA ovarian cancer datasets.

Table 2. Summary of performance metrics for the colon cancer, lung cancer, and ovarian cancer datasets. The arrow listed beside each performance metric indicates whether low values (↓) or high values (↑) are preferred.

Performance Metric	*Colon Cancer*			*Lung Cancer*			*Ovarian Cancer*			
	SVM	**Glmnetcr**	**RpartScore**	**Sig24**	**Shuffle24**	**Rand24**	**p_1E-8**	**p_1E-7**	**p_1E-6**	**p_1E-5**
$d \downarrow$	0.07	1.07	0.26	0.33	0.36	0.68	0.55	0.70	0.73	0.74
Accuracy ↑	0.93	0.32	0.75	0.73	0.73	0.43	0.46	0.34	0.30	0.28
AMAE ↓	0.11	1.00	0.29	0.34	0.28	0.86	1.15	1.1	0.97	0.87
$\tau_b \uparrow$	0.91	NA	0.79	0.72	0.49	0.08	0.19	0.19	0.25	0.28

In the first two simulations, 4×4 confusion matrices to represent four ordinal classes were simulated based on the following criteria: total sample size $N = 400$, accuracy set to 0.60 and 0.80, and proportion of the largest class to total sample size set to 0.25 (balanced design), 0.50, and 0.75. Confusion matrices were simulated with fixed class-wise accuracy in Simulation 1 and with fixed overall accuracy in Simulation 2. As an extension of Simulation 2, we increase the total sample size to $N = 800$ in Simulation 3, all other settings held constant. In Simulation 4, 5×5 confusion matrices to represent five ordinal classes were simulated based on the following criteria: total sample size $N = 1,000$, fixed overall accuracy set to 0.60 and 0.80, and proportion of largest class to total sample size set to 0.25 (balanced design), 0.50, and 0.75.

The study of simulated confusion matrices reveals a number of key findings (see Table 3). Since the proposed methodology is the only one that incorporates cost, our method scores perfectly across all simulation settings. In a com-

pletely balanced design, the cost matrix (as specified in the current manuscript) is equal to a scalar times AMAE. Thus, AMAE and d perform identically for most of the simulation study designed for balanced class sizes. When the number of classes increases and fixed overall accuracy is set to 80%, as in Simulation 4, a small number of simulated confusion matrices contained no entries for predicted class 1. In this case, AMAE chose the classifier with a higher misclassification cost, which explains the slight deviation of AMAE from d in Simulation 4 for $Acc = 80\%$.

Table 3. Summary of comparative performance metrics for four simulation studies, each was carried out under two predictive accuracy settings – 0.60 and 0.80.

Simulation		*Acc*=60%			*Acc*=80%		
		Bal	**0.50**	**0.75**	**Bal**	**0.50**	**0.75**
1	d	1.000	1.000	1.000	1.000	1.000	1.000
	AMAE	1.000	0.616	0.591	1.000	0.609	0.571
	τ_b	0.960	0.708	0.673	0.980	0.726	0.726
2	d	1.000	1.000	1.000	1.000	1.000	1.000
	AMAE	1.000	0.571	0.646	1.000	0.592	0.666
	τ_b	0.905	0.671	0.606	0.956	0.703	0.443
3	d	1.000	1.000	1.000	1.000	1.000	1.000
	AMAE	1.000	0.570	0.651	1.000	0.572	0.661
	τ_b	0.908	0.683	0.606	0.961	0.724	0.443
4	d	1.000	1.000	1.000	1.000	1.000	1.000
	AMAE	1.000	0.601	0.686	0.996	0.599	0.717
	τ_b	0.909	0.667	0.662	0.953	0.708	0.516

In Table 3, Simulation 1 has 4 classes with fixed class-wise accuracy. Simulation 2 is a study of 4 classes with fixed overall accuracy. Simulation 3 is an extension of Simulation 2 with double the total sample size. Simulation 4 has 5 classes with fixed overall accuracy. Each simulation study is evaluated for balanced class distribution (Bal), largest class size 50% of total sample size (0.50), and largest class size 75% of total sample size (0.75).

Kendall's τ_b performs best when the class sizes are balanced and class-wise accuracy is fixed, a scenario that is not often encountered in real-world application. In every simulation scenario, Kendall's τ_b monotonically decreases as the imbalance between class sizes increases. In general, as the opportunity for misclassification error increases, Kendall's τ_b is more severely impacted. In Simulations 2 – 4, we observed a much larger gap between largest class percentages of 0.50 and 0.75 when Acc is fixed at 0.80. In the former largest class percentage, there is less of an imbalance in class size so when accuracy is high, a number of classes can be perfectly classified. This, in turn, reduces misclassification error and increases the potential association between predicted and true class membership.

In contrast to Kendall's τ_b when overall Acc is fixed, AMAE shows a consistent v-shaped pattern across increased class distribution imbalance. Performance measures are lower for the largest class percentage of 0.50 due to the increased range of possible absolute error. A property observed across all ordinal metrics is the invariance of performance to total sample size and number of classes.

In our analysis of real data, the proposed methodology often selected the classifier that achieved highest accuracy, which is reasonable since a classifier with high accuracy is more likely to also have low cost. The decision to favor a model with better accuracy is intuitive when there is a large discrepancy in predictive accuracy. Thus, to study the tradeoff between accuracy and cost and evaluate the ability of the selected metrics to identify a lower accuracy/lower cost classifier, we carried out an additional simulation study with fixed overall accuracy set to 0.6, 0.7, and 0.8. Confusion matrices were simulated for a 4 class study with a total sample size of 800 and largest class percentage of 0.75. A total of 10,000 confusion matrices were simulated for each accuracy setting. We performed a pair-wise analysis of the simulated confusion matrices for $Acc = 0.60$ *vs.* $Acc = 0.70$ and for $Acc = 0.70$ *vs.* $Acc = 0.80$ and identified pairs of matrices where the proposed metric favored the lower accuracy/lower cost confusion matrix. A representative sample of the hits from each accuracy comparison is provided in Table 4.

In the first comparison, the classifier with lower accuracy is preferred by d although AMAE and Kendall's τ_b favor the classifier with higher accuracy. However, MC is at least two folds higher for the higher accuracy classifier (0.15 *vs.* 0.33 for $Acc = 0.60$ *vs.* $Acc = 0.70$). Thus, we are willing to forsake a 10% increase in accuracy for a 50% increase in cost, which is a decision the other metrics never consider. In the second comparison (see Table 4(b)), we again observe that the classifier with higher accuracy also has higher cost. The value of MC for confusion matrices generated with $Acc = 0.70$ and $Acc = 0.80$ are 0.13 and 0.26, respectively. Although this comparison also demonstrates that the price to pay for 10% increase in accuracy is double the value in cost, analysis of the d metric shows almost no distinction between the two. When this happens, we base our decision on the distance from the line of random chance to the coordinates for each classifier. These distances are 0.12 and 0.04 for $Acc = 0.70$ and $Acc = 0.80$, respectively, which indicates that the classifier with 70% accuracy is farther from the line of random chance in the performance grid (see Figure 1) and is thus the superior classifier.

Table 4. Representative confusion matrices and ordinal performance metrics for evaluating simulated data with (a) $Acc = 0.60$ *vs.* $Acc = 0.70$ and (b) $Acc = 0.70$ *vs.* $Acc = 0.80$

(a) $Acc_{0.60} = \begin{bmatrix} 444 & 37 & 27 & 57 \\ 144 & 25 & 18 & 0 \\ 2 & 2 & 8 & 7 \\ 10 & 2 & 14 & 3 \end{bmatrix}$	$Acc_{0.70} = \begin{bmatrix} 553 & 8 & 18 & 2 \\ 18 & 1 & 43 & 53 \\ 15 & 37 & 4 & 11 \\ 14 & 20 & 2 & 2 \end{bmatrix}$
d =0.428 AMAE = 1.222 τ_b = 0.138 MC = 15%	d =0.449 AMAE = 1.113 τ_b = 0.637 MC = 33%
(b) $Acc_{0.70} = \begin{bmatrix} 539 & 53 & 33 & 52 \\ 19 & 3 & 24 & 0 \\ 42 & 2 & 3 & 0 \\ 0 & 8 & 7 & 15 \end{bmatrix}$	$Acc_{0.80} = \begin{bmatrix} 589 & 4 & 3 & 1 \\ 0 & 4 & 25 & 17 \\ 4 & 14 & 6 & 8 \\ 7 & 44 & 33 & 41 \end{bmatrix}$
d =0.329 AMAE = 1.256 τ_b = 0.242 MC = 13%	d =0.331 AMAE = 0.820 τ_b = 0.843 MC = 26%

4. Conclusions

Predictive accuracy for multiclass classification problems has a number of drawbacks when evaluating data with imbalanced class distribution. For example, if the largest class represents a large proportion of all the samples, the classifier can achieve high accuracy by simply classifying each sample into the majority class. However, the cost associated with misclassifying a sample can be detrimental when classes are ordered and there is an increasing probability of risk. The most efficient classifier should consider these costs along with predictive accuracy.

An interesting observation in this study is that the best cost-sensitive model was not consistent with pairing *Acc* with AMAE or Acc with Kendall's τ_b. In other words, in some cases the model with better accuracy had higher misclassification error and sometimes the model with higher accuracy had lower rank correlation. This is problematic if either ordinal metric is used to judge the performance of a prediction model in lieu of accuracy or taken together with *Acc*. Naturally, if two classifiers have the same predictive accuracy, then the classifier associated with lower cost is the better model. Our analysis of simulated confusion matrices with fixed accuracy settings demonstrates that although AMAE is more consistent with the proposed method when class size is balanced, neither AMAE nor Kendall's τ_b consistently correlated with the classifier associated with lower cost under imbalanced class distribution. This was true for varying sample sizes, class distributions, and number of classes.

In this study, we incorporate domain knowledge by assuming a linear loss function in the cost matrix for ordinal classification to determine the optimum tradeoff between cost and accuracy. This imposes a constraint on evaluation that may not be ideal. For example, in an application of cancer prediction, the cost associated with classifying a normal patient into an advanced cancer category is not equivalent to the cost of mislabeling a patient with advanced cancer as normal. Since it is difficult to estimate the true cost matrix, a linear loss function offers some guidance in quantifying a cost-sensitive measure. On the other hand, if a researcher has additional information that would suggest a different or more accurate cost function, then the proposed methodology can be easily modified by changing the cost matrix.

Previous studies have evaluated accuracy and alternative performance metrics for ordinal data and have compared them under different settings and for different purposes.[26,27] In this study, we introduce a new performance metric for comparing ordinal classifiers and use the ordered relationships between classes to attribute cost to each error. We make no presumptions about the predictive modeling algorithm (*e.g.* details about feature selection, incorporating misclassification cost in the classifier-building algorithm, final model selection, *etc.*) nor do we stress how the resulting classification was derived; although most algorithms are guided by accuracy. Subsequent to selecting suitable classifiers for a given dataset, the current work provides researchers an improved metric for evaluating the comparative performance of multiple ordinal classifiers. Alternatively, it may prove beneficial to utilize the proposed metric in the development of an ordinal classifier through the process of integrating cost as a means of selecting the final classifier model.

Acknowledgements

The views presented in this paper are those of the authors and do not necessarily represent those of the U.S. Food and Drug Administration.

Conflicts of Interest Disclosure

The authors declare that they have no competing interests.

REFERENCES

[1] Claggett B, Tian L, Castagno D, *et al*. Treatment selections using risk–benefit profiles based on data from comparative randomized clinical trials with multiple endpoints. Biostatistics. 2014: kxu037.

[2] Leha A, Jung K, Beißbarth T. Utilization of ordinal response structures in classification with high-dimensional expression data. Proceedings of the German Conference on Bioinformatics; 2013. p. 90-100.

[3] Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Information Processing & Management. 2009; 45(4): 427-37. http://dx.doi.org/10.1016/j.ipm.2009.03.002

[4] Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognition Letters. 2009; 30(1): 27-38. http://dx.doi.org/10.1016/j.patrec.2008.08.010

[5] Baccianella S, Esuli A, Sebastiani F, editors. Evaluation measures for ordinal regression. Intelligent Systems Design and Applications, 2009 ISDA'09 Ninth International Conference on; 2009: IEEE.

[6] Gaudette L, Japkowicz N. Evaluation methods for ordinal classification. Advances in Artificial Intelligence: Springer; 2009. p. 207-10.

[7] Sánchez-Mondero J, Gutiérrez P, Fernández-Navarro F, *et al*. Weighting efficient accuracy and minimum sensitivity for evolving multi-class classifiers. Neural Processing Letters. 2011; 34(2): 101-16. http://dx.doi.org/10.1007/s11063-011-9186-9

[8] Kendall M. Rank Correlation Methods. 3rd edition, New York: Hafner Press; 1962.

[9] Waegeman W, De Baets B, Boullart L, editors. A Comparison of Different ROC Measures for Ordinal Regression. In: Proceedings of the 3rd International Workshop on ROC Analysis in Machine Learning; 2006.

[10] Goodman L, Kruskal W. Measures of association for cross classifications. Journal of the American Statistical Association. 1954; 49: 732-64.

[11] Somers R. The rank analogue product-moment partial correlation and regression with application to manifold, ordered contingency tables. Biometrika. 1955; 46: 241-6. http://dx.doi.org/10.1093/biomet/46.1-2.241

[12] Pinto da Costa J, Alonso H, Cardoso J. The unimodal model for the classification of ordinal data. Neural Networks. 2008; 21(1): 78-91. PMid:18093801. http://dx.doi.org/10.1016/j.neunet.2007.10.003

[13] Archer J. rpartOrdinal: An R package for deriving a classification tree for predicting an ordinal response. J Stat Softw. 2010; 34: 7. PMid:20625561.

[14] Archer K, Williams A. L1 penalized continuation ratio models for ordinal response prediction using high-dimensional datasets. Stat Med. 2012; 31(14): 1464-74. PMid:22359384. http://dx.doi.org/10.1002/sim.4484

[15] Chu W, Keerthi S. Support vector ordinal regression. Neural Computation. 2007; 19(3): 792-815. PMid:17298234. http://dx.doi.org/10.1162/neco.2007.19.3.792

[16] Galimberti G, Soffritti G, Di Maso M. Classification trees for ordinal responses in R: the rpartScore package. J Stat Softw. 2012; 47(10): 1-25.

[17] Hechenbickler H, Schliep K. Weighted k-nearest-neighbor techniques and ordinal classification. In Discussion Paper 399, SFB 386 2006.

[18] Kotsiantis S, Pintelas P. A cost sensitive technique for ordinal classification problems. SETN. 2004: 220-9.

[19] Lin HT, Li L. Reduction from cost-sensitive ordinal ranking to weighted binary classification. Neural Computation. 2012; 24(5): 1329-67. PMid:22295981. http://dx.doi.org/10.1162/NECO_a_00265

[20] Lange K. Numerical Analysis for Statisticians: Springer: p185ff; 2001.

[21] Smith JJ, Deane NG, Wu F, *et al*. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology. 2010; 138(3): 958-68. PMid:19914252. http://dx.doi.org/10.1053/j.gastro.2009.11.005

[22] Lu TP, Tsai MH, Lee JM, *et al*. Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. Cancer Epidemiol Biomarkers Prev. 2010; 19(10): 2590-7. PMid:20802022. http://dx.doi.org/10.1158/1055-9965.EPI-10-0332

[23] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474(7353): 609-15. PMid:21720365. http://dx.doi.org/10.1038/nature10166

[24] Meyer D, Dimitriadou E, Hornik K, *et al*. e1071: Misc functions of the Department of Statistics (e1071), Tu Wien. 2014.

[25] Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc. 2002; 97(457): 77-87. http://dx.doi.org/10.1198/016214502753479248

[26] Cardoso J, Sousa R. Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence. 2011; 25(8): 1173-95. http://dx.doi.org/10.1142/S0218001411009093

[27] Cruz-Ramírez M, Hervás-Martínez C, Sánchez-Monedero J, *et al*. A preliminary study of ordinal metrics to guide a multi-objective evolutionary algorithm. ISDA. 2011: 1176-81.