Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Pattern Recognition Letters 29 (2008) 1–9

Pattern Recognition Letters

www.elsevier.com/locate/patrec

ROC analysis in ordinal regression learning

Willem Waegeman [a,*], Bernard De Baets [b], Luc Boullart [a]

[a] *Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium*
[b] *Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium*

Received 22 November 2006; received in revised form 25 March 2007
Available online 14 September 2007

Communicated by W. Pedrycz

Abstract

Nowadays the area under the receiver operating characteristics (ROC) curve, which corresponds to the Wilcoxon–Mann–Whitney test statistic, is increasingly used as a performance measure for binary classification systems. In this article we present a natural generalization of this concept for more than two ordered categories, a setting known as ordinal regression. Our extension of the Wilcoxon–Mann–Whitney statistic now corresponds to the volume under an r-dimensional surface (VUS) for r ordered categories and differs from extensions recently proposed for multi-class classification. VUS rather evaluates the ranking returned by an ordinal regression model instead of measuring the error rate, a way of thinking which has especially advantages with skew class or cost distributions. We give theoretical and experimental evidence of the advantages and different behavior of VUS compared to error rate, mean absolute error and other ranking-based performance measures for ordinal regression. The results demonstrate that the models produced by ordinal regression algorithms minimizing the error rate or a preference learning based loss, not necessarily impose a good ranking on the data.

Keywords: ROC analysis; Ranking; Ordinal regression; Unbalanced learning problems; Performance measures; Machine learning

1. Introduction

In multi-class classification, labels are picked from a finite set of unordered categories. In metric regression, labels might take an infinite number of continuous values. Ordinal regression can be located in between these learning problems because here labels are chosen from a finite set of ordered categories. Applications of ordinal regression frequently arise in domains where humans participate in the data generation process. Humans prefer to choose a label from a (usually) small set of alternatives when they assess objects for their beauty, quality, suitability or any other characteristic. In essence, they prefer to quantify objects with ordinal labels instead of continuous scores, although often an underlying and unobserved continuous variable is assumed. This kind of data is for example found in information retrieval, when users of recommender systems express on a scale from one to five stars to what extent they like items like movies or songs. Machine learning techniques are then applied to predict the ratings of new users on these items or to recommend new items to existing users of the system. Another application is quality control, where human experts frequently evaluate products with linguistic terms, varying from ‘very bad’ to ‘very good’ for example. Also in medicine and social sciences, where many data sets originate by interaction with humans, ordinal regression models can be employed.

In these applications of ordinal regression one is often primarily interested in a subset of the classes. In many cases these classes of interest are the ‘extreme’ categories, such as the documents with the highest relevance to the query or the products with the lowest quality and, consequently, we would like to associate a higher misclassification cost with these objects. Moreover, there is often an unequal number of training objects for the different categories in real-world ordinal regression problems. In other words, we are often dealing with a skew class and/or cost distribution.

[*] Corresponding author. Tel.: +32 9 2645586; fax: +32 9 2645839.
E-mail address: Willem.Waegeman@UGent.be (W. Waegeman).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.07.019

The overall classification rate or mean absolute error, which are commonly used for evaluating ordinal regression models, do not fully represent the desired performance of our system. In these situations, we are more interested in criteria that quantify the ranking on the data imposed by the classifier. We will directly explain how this relates to ROC analysis.

This article aims to discuss possible extensions of ROC analysis for ordinal regression. It is organized as follows. In Section 2 we briefly describe the main concepts of binary and multi-class ROC analysis in the framework of machine learning. This gives us the opportunity to define an extension of the Wilcoxon–Mann–Whitney statistic and its geometrical interpretation in Section 3 and, subsequently, the properties of this performance estimator and related measures are compared. In Section 4 all measures are empirically analyzed on synthetic and real-world data. Finally, in Section 5 we formulate a conclusion and present some ideas for future work.

2. Overview of existing work

In machine learning one often assumes that examples are identically and independently drawn according to an unknown distribution $\mathscr{D}$ over $\mathscr{X} \times \mathscr{Y}$ with $\mathscr{X}$ the object space and $\mathscr{Y}$ the set of labels. In binary classification two types of objects are observed and $\mathscr{Y} = \{\bar{y}_-, \bar{y}_+\}$. This extends to $\mathscr{Y} = \{\bar{y}_1, \ldots, \bar{y}_r\}$ when more than two types of objects have to be distinguished (for r categories). In the case of ordinal regression there is a linear order relation $\leqslant_{\mathscr{Y}}$ defined on the elements of $\mathscr{Y}$. When the order relation is absent, one speaks of multi-class classification. Furthermore, we define a data set of size n as $D = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$, so $D \subseteq \mathscr{X} \times \mathscr{Y}$. Sometimes we will also need the conditional distribution of a data object from $\mathscr{X}$ given that it belongs to category $\bar{y}_k$, which will be denoted by $\mathscr{D}_k$, and the marginal distribution on $\mathscr{X}$ will be referred to as $\mathscr{D}_{\mathscr{X}}$. The number of data objects in the data set D with label $\bar{y}_k$ will be denoted by n_k. In the binary case (when $r = 2$) we will use the notations $\mathscr{D}_-, \mathscr{D}_+, n_-$ and n_+.

The main concept of binary ROC analysis says that one can consider two kinds of errors when two types of objects have to be distinguished. Since long the method is widely applied in medicine when diseased subjects have to be separated from healthy cases. Quite recently ROC analysis has been introduced in the machine learning domain where the area under the ROC curve is increasingly used as a performance measure for classification systems.

A ROC curve is created by plotting the *true positive rate* (TPR) versus the *false positive rate* (FPR). The TPR (or *sensitivity*) and the FPR (also known as 1 – *specificity*) are computed from the confusion matrix or contingency table (shown in Table 1). Sensitivity is defined as the number of positive predicted examples from the positive class TP divided by the number of positive examples n_+ and specificity is defined as the number of negative predicted examples TN from the negative class divided by the number of negative examples n_-:

Table 1
Confusion matrix for a two class classification problem of size n with y the true labels and $\hat{y}$ the predicted labels

	$\hat{y} = -1$	$\hat{y} = 1$	
$y = -1$	TN	FP	n_-
$y = 1$	FN	TP	n_+
	NP	PP	n

$$\text{Sens} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

$$\text{Spec} = \text{TNR} = 1 - \text{FPR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2}$$

With a classifier that estimates a continuous function $f : \mathscr{X} \to \mathbb{R}$, the class prediction $h : \mathscr{X} \to \mathscr{Y}$ for an object x is obtained by the following rule:

$$h(\boldsymbol{x}) = \text{sgn}(f(\boldsymbol{x}) + b) \tag{3}$$

with b a real number. The points defining the ROC curve can then be computed by varying the threshold b from the most negative to the most positive function value and the *area under the ROC curve* (AUC) gives an impression of the quality of the classifier. It has been shown (Cortes and Mohri, 2003; Yan et al., 2003) that the AUC is equivalent to the Wilcoxon–Mann–Whitney statistic:

$$\text{AUC} = \widehat{A}(f, D) = \frac{1}{n_- n_+} \sum_{y_i < y_j} I_{f(\boldsymbol{x}_i) < f(\boldsymbol{x}_j)} \tag{4}$$

with I the indicator function.[1]

As evaluation criterion, the area under the ROC curve offers advantages over accuracy when the class distributions are unbalanced or when different misclassification costs can be assigned to the different classes. The impact of the skewness of the class or cost distributions can be efficiently analyzed with ROC curves (Flach, 2003). Cortes and Mohri (2003) studied in detail the relationship between accuracy and the AUC and concluded that both measures will reveal separate characteristics of a classifier. In particular, they derived an exact expression for the expected value and the variance of the AUC for a fixed error rate and showed that classifiers with the same (low) error rate can exhibit noticeably different AUC values.

Another difference is that the mean zero-one error (1 – accuracy) directly evaluates the performance of the decision function h, while the AUC quantifies the ranking imposed by the function f on the data without taking the predicted labels into account. Therefore, f will be further referred to as a *ranking* function. Agarwal et al. (2005) define in this context the term *expected ranking accuracy*.

[1] The value of the indicator function I will be one when its argument, written as subscript, is true and zero otherwise.

Definition 1. Let $f : \mathcal{X} \to \mathbb{R}$ be a ranking function on $\mathcal{X}$, then the expected ranking accuracy, denoted by $A(f)$, is defined as

$$A(f) = P_{X_-\sim\mathcal{D}_-, X_+\sim\mathcal{D}_+}\{f(X_-) < f(X_+)\}. \tag{5}$$

The expected ranking accuracy thus stands for the probability that an object randomly drawn from the negative class is assigned a smaller value by the ranking function than an object randomly drawn from the positive class. The choice of notation is guided by the property that (4) is an unbiased nonparametric estimator of (5) on a data set $D \in (\mathcal{X} \times \mathcal{Y})^n$, which can be easily seen from the formulas.

3. ROC measures for ordinal regression

Recently, different approaches have been proposed to extend ROC analysis for multi-class classification, see e.g. Hand and Till (2001), Ferri et al. (2003), Flach (2004), Fieldsend and Everson (2006). In the most general case, the volume under the ROC surface (VUS) has to be maximized in multi-class classification. The ROC surface can be seen as a Pareto front, where each objective corresponds to one dimension. In case there are more than two classes (say r), then the number of objectives depends on the multi-class method that is used:

- For a *one-versus-all* method, r functions f_k are estimated that try to separate objects of class $\bar{y}_k$ from the other classes. As a consequence misclassification costs for each class are fixed and the corresponding ROC surface will have r dimensions representing the true positive rates TPR_k for each class (Flach, 2004). The true positive rate of class $\bar{y}_k$ is defined as the fraction of the instances classified as class $\bar{y}_k$ that really belong to class k, divided by the number of data objects of class $\bar{y}_k$. ROC points are here obtained by varying the thresholds b_k in the prediction rule $h(\boldsymbol{x}) = \text{argmax}_{\bar{y}_k} f_k(\boldsymbol{x}) + b_k$.
- For a *one-versus-one* method, a function f_{kl} is estimated for each pair of classes, which allows to specify the cost for a misclassification of an object of class $\bar{y}_k$ predicted as class $\bar{y}_l$. The corresponding ROC space is in this case spanned by $\frac{r(r-1)}{2}$ objectives (Hand and Till, 2001; Ferri et al., 2003). A prediction for new instances is done by majority voting over all $\frac{r(r-1)}{2}$ classifiers based on the outcomes $\text{sgn}(f_{kl}(\boldsymbol{x}) + b_{kl})$.

In ordinal regression the picture is slightly different. The vast majority of existing ordinal regression models, like traditional statistical models (Agresti, 2002), kernel methods (Shashua and Levin, 2003; Chu and Keerthi, 2005), perceptron based algorithms (Crammer and Singer, 2001), ordinal regression trees (Kramer et al., 2000) and bayesian approaches (Chu and Ghahramani, 2005), can be represented in the following general form:

$$h(\boldsymbol{x}) = \begin{cases} \bar{y}_1, & \text{if } f(\boldsymbol{x}) < b_1 \\ \bar{y}_k, & \text{if } b_{k-1} < f(\boldsymbol{x}) \leqslant b_k,\ k = 2,\ldots,r-1 \\ \bar{y}_r, & \text{if } f(\boldsymbol{x}) > b_{r-1} \end{cases} \tag{6}$$

with $b_1 < \cdots < b_{r-1}$ free parameters and $f : \mathcal{X} \to \mathbb{R}$ a continuous (ranking) function.

As mentioned before, in multi-class classification more than one ranking function is used to derive a classification rule h. Because we are dealing here with a single ranking function, the model is more restricted. The following definition extends (5) to more than two classes.

Definition 2. Let $f : \mathcal{X} \to \mathbb{R}$ be the ranking function of an ordinal regression model of the form (6), then the expected ranking accuracy, denoted by $U(f)$, is defined as

$$U(f) = P_{X_k\sim\mathcal{D}_k}\{f(X_1) < \cdots < f(X_r)\}. \tag{7}$$

Now the expected ranking accuracy measures the probability that a random sequence of one data object of each category is correctly ranked by the ranking function f. It is estimated from a finite data set D by counting the number of sequences of r objects, one of each class, that are correctly ranked by the ranking function, i.e.

$$\widehat{U}(f,D) = \frac{1}{\prod_{k=1}^{r} n_k} \sum_{y_{j_1}<\cdots<y_{j_r}} I_{f(\boldsymbol{x}_{j_1})<\cdots<f(\boldsymbol{x}_{j_r})}. \tag{8}$$

It is easy to see that $\widehat{U}(f,D)$ is an unbiased estimator of $U(f)$. Furthermore, $\widehat{U}(f,D)$ has a geometrical interpretation, which can be summarized in the following theorem.

Theorem 3.1. *Given a ranking function $f : \mathcal{X} \to \mathbb{R}$ that imposes a ranking over a data set $D \in (\mathcal{X} \times \mathcal{Y})^n$, then $\widehat{U}(f,D)$ corresponds to the volume under the r-dimensional ROC surface (VUS) spanned by the true positive rates of each class.*

Proof. For a given ranking function f, the true positive rate of each class only depends on the threshold vector $\mathbf{b} = (b_0,\ldots,b_r)$ as defined in (6) with $b_0 = -\infty$ and $b_r = +\infty$, i.e.

$$TPR_k(\mathbf{b}) = \frac{1}{n_k} \sum_{y_i=\bar{y}_k} I_{b_{k-1}<f(\boldsymbol{x}_i)\leqslant b_k}. \tag{9}$$

The ROC surface represents all optimal models for different cost distributions and each model corresponds to a unique vector of true positive rates. We can collect all possible values for the true positive rates by linking a particular element of the data set with each threshold, i.e.

$$b_k = f(\boldsymbol{x}_{j_k}), \tag{10}$$

with $k = 1,\ldots,r-1$ and $j_k = 1,\ldots,n$. Let us therefore consider the set $\mathcal{B}$ containing all such vectors $\mathbf{b}$ that uniquely define a point on the convex hull of the ROC surface, then without loss of generality the volume under the ROC surface can be written as:

$$\text{VUS} = \frac{1}{|\mathcal{B}|} \sum_{\mathbf{b}\in\mathcal{B}} \text{TPR}_r(\mathbf{b}). \tag{11}$$

We will now write out this set $\mathcal{B}$ by looking at all possible vectors $\mathbf{b}$ that lead to different points on the convex hull. By definition, the cardinality of $\mathscr{B}$ is upper bounded by n^{r-1} because only $b_1,\ldots,b_{r-1}$ can vary.

Nevertheless, the cardinality will be smaller because the ordering of the thresholds enforces an additional constraint on $\mathbf{b}$:

$$f(\boldsymbol{x}_{j_1}) \leqslant \cdots \leqslant f(\boldsymbol{x}_{j_r}) \tag{12}$$

Now suppose that the data set was sorted according to f. We will count all elements of $\mathscr{B}$ starting with all thresholds at the first element of the data set: $b_1 = \cdots = b_{r-1} = f(\boldsymbol{x}_1)$, so all instances are then classified into the last category and hence $\mathrm{TPR}_r = 1$. This value will gradually decrease when the thresholds are shifted to the end of the ranking. We will first count all contributions while the last threshold is moved up. When the position j_{r-1} increases one step to the next element in the ordered data set, the true positive rates TPR_{r-1} and TPR_r can change. There are three possibilities:

- $y_{j_{r-1}} = \bar{y}_{r-1}$: TPR_{r-1} will increase, leading to a new element of $\mathscr{B}$ (we shift in a horizontal direction to a new point on the ROC-surface).
- $y_{j_{r-1}} = \bar{y}_r$: TPR_r will decrease, but this setting for b_{r-1} does not correspond to a new point on the surface (we shift in a vertical direction to a new point lying beneath the convex hull of all ROC points).
- $y_{j_{r-1}} \notin \{\bar{y}_{r-1}, \bar{y}_r\}$: there is no contribution to the sum because TPR_r and TPR_{r-1} remain unchanged (we stay at the same point on the ROC surface).

As a consequence, the sum over all values for b_{r-1} reduces to a sum over all elements of class $\bar{y}_{r-1}$ and recursively the same reasoning holds for each b_k. Thus, $\mathscr{B}$ contains $\prod_{k=1}^{r-1} n_k$ elements and

$$\mathrm{VUS} = \frac{1}{\prod_{k=1}^{r-1} n_k} \sum_{\substack{b_k = f(\boldsymbol{x}_{j_k}) \\ y_{j_1} < \ldots < y_{j_{r-1}}}} \mathrm{TPR}_r(\mathbf{b}) \tag{13}$$

For all positions $(j_1,\ldots,j_{r-1})$ of the thresholds the true positive rate of class $\bar{y}_r$ can be seen as

$$\mathrm{TPR}_r(\mathbf{b}) = \frac{1}{n_r} \sum_{y_{j_r} = \bar{y}_r} I_{f(\boldsymbol{x}_{j_{r-1}}) < f(\boldsymbol{x}_{j_r})} \tag{14}$$

This value is counted over all objects of the other $r-1$ classes. Combining (13) and (14) and moving constraint (12) from the sum to the indicator function leads to expression (8). □

In statistics there has been some related work on this topic. Dreiseitl et al. (2000) derive formulas for the variance of $\widehat{U}(f,D)$ and the covariance between two volumes in the three-class case. This work has been extended to the general r-class case by Nakas and Yiannoutsos (2004). They conclude that bootstrapping is preferred over U-statistics[2] to compare more than two diagnostic tests because the computation of the exact variance and covariance estimators become intractable for large values of n and r. In this article we focus more on the use of $\widehat{U}(f,D)$ as performance measure for ordinal regression problems.

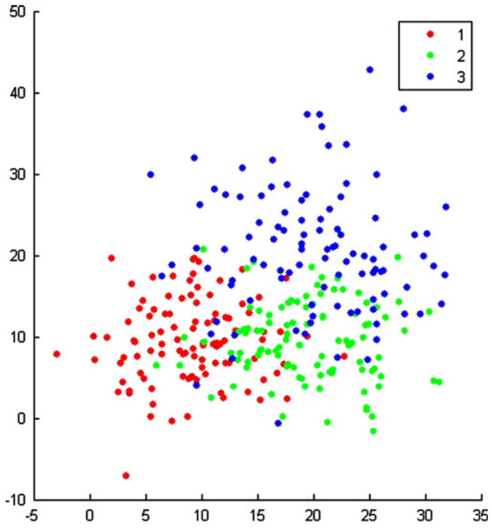


Fig. 1. A synthetic data set with three bivariate Gaussian clusters representing three ordered classes with respective means (10,10), (20,10) and (20,20). The standard deviation was set to (5,5) for the first two clusters and to (7,7) for the last cluster, while ρ was fixed to 0.

For three ordered classes the ROC surface can be visualized. We have constructed this ROC surface for a synthetic data set. We sampled $3*100$ instances from 3 bivariate Gaussian clusters representing consecutive classes. The mean of the clusters was set to (10,10), (20,10) and (20,20) respectively, σ_1 and σ_2 were set to 5 for the first two clusters and were set to 7 for the last cluster. ρ was fixed to 0. This data set is visualized in Fig. 1. We used the support vector ordinal regression algorithm of Chu and Keerthi (2005) to estimate the ranking function f, without looking at the thresholds. The obtained ROC surface is shown in Fig. 2.

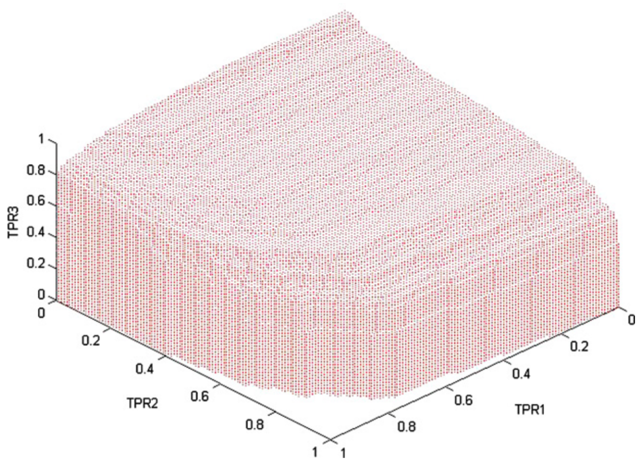


Fig. 2. The ROC surface obtained for the synthetic data set of Fig. 1 and the ranking returned by a support vector ordinal regression algorithm.

[2] A U-statistic is a class of nonparametric statistics. See for example Lehmann (1975) for more information.

Let us now associate a cost function $c : \mathscr{Y} \to \mathbb{R}$ with each category of $\mathscr{Y}$. In other words, $c(\bar{y}_k)$ defines the penalty of misclassifying an object of class $\bar{y}_k$. As in the binary case, the convex hull of the r-dimensional ROC surface represents the set of optimal classifiers for any particular choice of cost. With known costs and a predefined loss function l we can search for a classifier which minimizes

$$\sum_{i=1}^{n} c(y_i) l(h(\boldsymbol{x}_i), y_i) \tag{15}$$

on a training data set when the costs are provided before training. However, the costs will in many cases still vary after training and then the optimal classifier can be selected from the ROC surface.

The volume under the ROC surface gives us the opportunity to compare the quality of two different ROC surfaces for varying costs. Because $\widehat{U}(f, D)$ is a nonparametric estimator of the expected ranking accuracy, it quantifies the ranking imposed by an ordinal regression model. We will now discuss how $\widehat{U}(f, D)$ relates to previous work on ordinal regression and multi-class classification.

Herbrich et al. (2000) propose a ranking-based framework for ordinal regression based on structural risk minimization. Guided by preference learning, their algorithm optimizes the number of correctly ranked object pairs, i.e.

$$\widehat{U}_{\text{pairs}}(f, D) = \frac{1}{\sum\limits_{\bar{y}_k < \bar{y}_l} n_k n_l} \sum\nolimits_{y_i < y_j} I_{f(\boldsymbol{x}_i) < f(\boldsymbol{x}_j)} \tag{16}$$

together with a regularization term in the form of a convex quadratic program.

As stressed out at the beginning of this section, ROC analysis for ordinal outcomes can be seen as a simplified version of multi-class ROC analysis since we are dealing with a single ranking function. Nevertheless, the approximation presented by Hand and Till (2001) for one-versus-one multi-class classification could easily be modified for our purpose, i.e.

$$\begin{aligned} \widehat{U}_{\text{ovo}}(f, D) &= \frac{2}{r(r-1)} \sum_{k<l} \widehat{A}_{kl}(f, D) \\ \widehat{A}_{kl}(f, D) &= \frac{1}{n_k n_l} \sum_{y_i = \bar{y}_k} \sum_{y_j = \bar{y}_l} I_{f(\boldsymbol{x}_i) < f(\boldsymbol{x}_j)} \end{aligned} \tag{17}$$

They construct a two-dimensional ROC curve for each pair of classes $\{\bar{y}_k, \bar{y}_l\}$, in which only the objects of these classes are taken into account. Instead of multi-classification, we use the same ranking function for each two-dimensional ROC curve and take the sum of the areas under these curves as a measure for the quality of the ranking. In nonparametric statistics $\widehat{U}_{\text{ovo}}(f, D)$ is known as the *Jonckheere–Terpstra* test, a more powerful alternative to the *Kruskal–Wallis* test for simultaneously testing whether more than two ordered populations significantly differ (Higgins, 2004).

Another approximation $\widehat{U}_{\text{cons}}(f, D)$ is directly deduced from (6). With a function f and $r - 1$ thresholds one could envisage threshold b_k as providing the separation between the consecutive ranks $\bar{y}_k$ and $\bar{y}_{k+1}$. Varying this threshold will change the proportion between objects predicted lower than or equal to class k and objects predicted higher than class k. This corresponds to measuring the non-weighted sum of $r - 1$ two-dimensional ROC curves representing the trade-off between consecutive classes:

$$\begin{aligned} \widehat{U}_{\text{cons}}(f, D) &= \frac{1}{r-1} \sum_{k=1}^{r-1} \widehat{A}_k(f, D) \\ \widehat{A}_k(f, D) &= \frac{1}{\sum_{i=1}^{k} n_i \sum_{j=k+1}^{n} n_j} \sum_{y_i \leqslant \bar{y}_k} \sum_{y_j > \bar{y}_k} I_{f(\boldsymbol{x}_i) < f(\boldsymbol{x}_j)} \end{aligned} \tag{18}$$

$\widehat{U}_{\text{pairs}}(f, D)$, $\widehat{U}_{\text{ovo}}(f, D)$ and $\widehat{U}_{\text{cons}}(f, D)$ all compare pairs of objects instead of sequences of r objects as in VUS. These measures can be seen as unbiased estimators of probabilities, other than the expected ranking accuracy. This is summarized in the following theorem.

Theorem 3.2. *Let $\mathscr{D}_{k,l}$ be the conditional distribution of an object of $\mathscr{X}$ given that its label lies between categories $\bar{y}_k$ and $\bar{y}_l$ (inclusive) and let*

$$U_{\text{pairs}}(f) = P_{X_1, X_2 \sim \mathscr{D}_{\mathscr{X}}}\{f(X_1) < f(X_2)\}$$

$$U_{\text{ovo}}(f) = \frac{2}{r(r-1)} \sum_{k<l} P_{X_1 \sim \mathscr{D}_k, X_2 \sim \mathscr{D}_l}\{f(X_1) < f(X_2)\}$$

$$U_{\text{cons}}(f) = \frac{1}{r-1} \sum_{k=1}^{r-1} P_{X_1 \sim \mathscr{D}_{1,k}, X_2 \sim \mathscr{D}_{k+1,r}}\{f(X_1) < f(X_2)\}$$

then

$$E[\widehat{U}_{\text{pairs}}(f, D)] = U_{\text{pairs}}(f)$$

$$E[\widehat{U}_{\text{ovo}}(f, D)] = U_{\text{ovo}}(f)$$

$$E[\widehat{U}_{\text{cons}}(f, D)] = U_{\text{cons}}(f).$$

The proof directly follows from the definitions. $\widehat{U}_{\text{pairs}}(f, D)$, $\widehat{U}_{\text{ovo}}(f, D)$ and $\widehat{U}_{\text{cons}}(f, D)$ assess the ranking of an ordinal regression model in another way than VUS. They can be considered as approximations of VUS because they only count object pairs instead of sequences, which is computationally more efficient. However, they also turn out to behave differently than VUS.

Lemma 3.3. *Let f_R be a random ranking function of an ordinal regression model with r classes, thus f_R randomly assigns continuous outputs to data objects, then*

(i) $U(f_R) = \frac{1}{r!}$.
(ii) $U_{\text{pairs}}(f_R) = U_{\text{ovo}}(f_R) = U_{\text{cons}}(f_R) = \frac{1}{2}$.

Proof. (i) There are $r!$ different ways of ordering a sequence of r objects and f_R randomly picks one of these rankings. (ii) The other measures only compare object pairs. □

One must be careful in interpreting the values of different measures for a model. With a relatively high number of categories (say $r > 4$), a value of 0.5 for the volume under the ROC surface indicates that the ordinal regression model is able to rank the data well. Conversely, for few classes and especially in the binary classification case, a value of 0.5 for VUS alludes to absolutely no discriminative power for the ranking function. $\widehat{U}_{\text{pairs}}(f,D)$, $\widehat{U}_{\text{ovo}}(f,D)$ and $\widehat{U}_{\text{cons}}(f,D)$ all follow the behavior of the latter case and, hence, for poor ranking models and r relatively high, noticeably different values will be found in practice between VUS and its approximations (see also the following section). Additionally, the following relationship is observed between $\widehat{U}_{\text{pairs}}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$:

$$\widehat{U}_{\text{pairs}}(f,D) = \frac{1}{\sum_{\bar{y}_k<\bar{y}_l} n_k n_l} \sum_{k<l} n_k n_l \widehat{A}_{kl}(f,D) \tag{19}$$

As a result, both estimators are equal for balanced data sets. In the following section we will demonstrate that they can significantly differ for unbalanced data sets.

4. Experiments

Three kinds of experiments were set up to reveal the characteristics of $\widehat{U}(f,D)$ and to make a comparison with the approximations and the standard measures 'mean zero-one error' and 'mean absolute error'. First we show by means of simulations on synthetic ranking functions that the relationship between $\widehat{U}(f,D)$ and its approximations appears to be non-linear. In a second simulation experiment we prove that in general there will be no monotone relationship between $\widehat{U}(f,D)$ and other ranking-based measures. We also investigate the distribution of $\widehat{U}(f,D)$. In a last experiment we use real-world data to demonstrate that ranking optimization and error rate minimization are conflicting objectives for a classifier system when the data set is unbalanced. Therefore, all measures serve as fitness scores in a multi-objective stochastic search procedure concentrating on the region of the search space that represents good classifiers.

4.1. Simulations

In the first experiment we wanted to find out which values are obtained for different levels of separability and for an increasing number of classes. Therefore we assume that the function values of the model f can be represented by a distribution with cdf $F(\boldsymbol{x})$, in which the function values for the objects of class $\bar{y}_k$ are distributed with cdf $F_k(\boldsymbol{x}) = F(\boldsymbol{x} - kd)$. Furthermore, we sample from a Gaussian distribution with standard deviation $\sigma = 1$. So the function values conditioned on the labels are normally distributed with equidistant ordered means. Repeatedly 100 data points were sampled from each class while we increased the distance d between the means of consecutive clusters. We started at $d = 0$ (random classifier) and stopped at $d = 5$ (as good as perfect separation) with step size 0.25.

The results obtained for $\widehat{U}(f,D)$, $\widehat{U}_{\text{cons}}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$ are graphically compared. In this simulation all classes have the same prior of occurring, so $\widehat{U}_{\text{ovo}}(f,D)$ and $\widehat{U}_{\text{pairs}}(f,D)$ will always have the same value due to (19). Hence, the results for $\widehat{U}_{\text{pairs}}(f,D)$ are omitted. The relationship between $\widehat{U}(f,D)$ and $\widehat{U}_{\text{cons}}(f,D)$ on the one hand and between $\widehat{U}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$ on the other hand are respectively shown in Figs. 3 and 4. One can see that, as expected, these relationships are without doubt non-linear. As discussed at the end of the previous section, the average value of $\widehat{U}(f,D)$ heavily depends on the number of classes, while this is not the case for the approximations. The approximations all take an average over a set of two-dimensional ROC-curves, so their average value is never lower than a half, irrespective of the number of classes. Nevertheless, one can also see that $\widehat{U}(f,D)$ converges rapidly to one when the distance between the subsequent means increases. In addition, $\widehat{U}_{\text{cons}}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$ behave quite similarly in this simulation. This is also shown in Fig. 5. Their observed values become more dissimilar when the number of classes increases.

In a second simulation we wanted to investigate whether all ranking-based performance measures are pairwisely

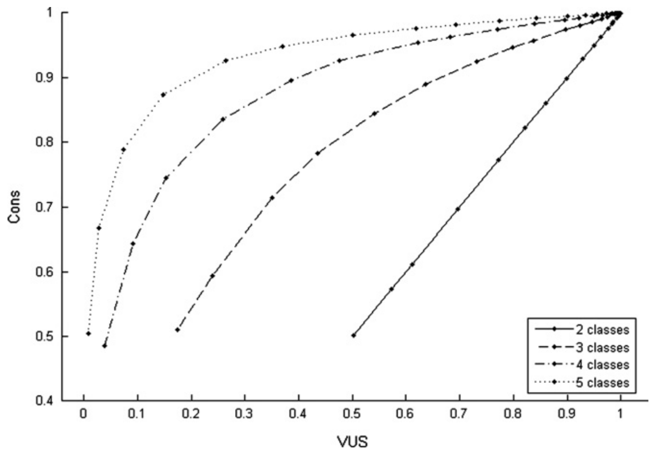


Fig. 3. Relationship between $\widehat{U}(f,D)$ and $\widehat{U}_{\text{cons}}(f,D)$ for $r = 2,\ldots,5$ and $d = 0,\ldots,5$ with step size 0.25. The values are averaged over 20 runs.

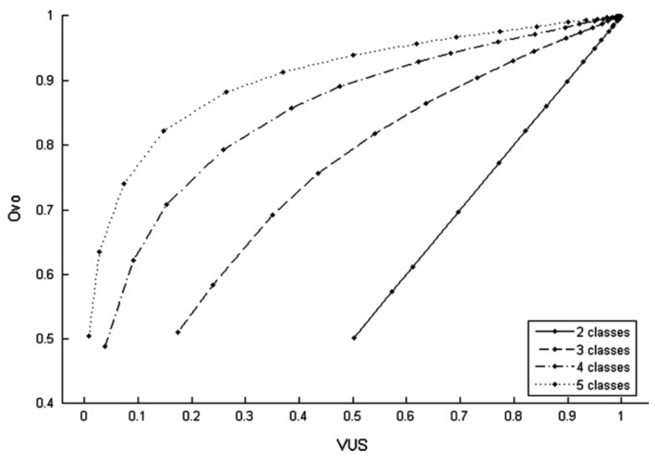


Fig. 4. Relationship between $\widehat{U}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$ for $r = 2,\ldots,5$ and $d = 0,\ldots,5$ with step size 0.25. The values are averaged over 20 runs.

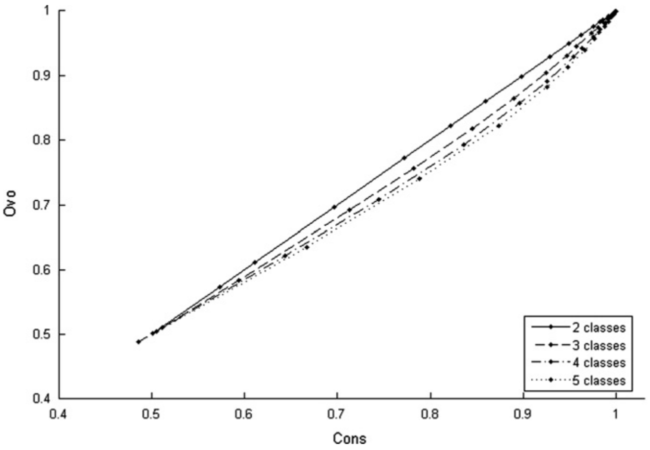


Fig. 5. Relationship between $\widehat{U}_{\text{cons}}(f,D)$ and $\widehat{U}_{\text{ovo}}(f,D)$ for $r = 2,\ldots,5$ and $d = 0,\ldots,5$ with step size 0.25. The values are averaged over 20 runs.

comonotone associated. A comonotone association between two measures M_1 and M_2 means that for any two functions f and f^* part of a given hypothesis space $\mathscr{F}$:

$$M_1(f) < M_1(f^*) \iff M_2(f) < M_2(f^*) \tag{20}$$

$$M_1(f) = M_1(f^*) \iff M_2(f) = M_2(f^*). \tag{21}$$

To test whether this property holds for all four measures we looked at a large number of rankings for a synthetic data set. All measures only quantify the quality of the ordering of a data set for a function f. For a data set of size n there are n! possible rankings of the objects, so evaluating them all is computationally intractable. Therefore, we sampled randomly 1000 rankings from all possible orderings of the data set. We assumed we had 50 samples per class with four ordered classes, resulting in a sample size of 200 objects and 200! possible rankings. The results are given in Fig. 6, which shows the distributions of all measures together with pairwise scatter plots. All classes again have the same prior of occurring, so $\widehat{U}_{\text{ovo}}(f,D)$ and $\widehat{U}_{\text{pairs}}(f,D)$ have a perfect correlation. This is however not true for the other measures. One can clearly see that for no other pair of measures conditions (20) or (21) hold. Thus, for many types of ordinal regression systems, two models can be found for which the first model dominates the second model in terms of measure $\widehat{U}(f,D)$, while the latter dominates the first one in terms of another performance measure. As a result, the question arises whether the same model will be obtained when optimizing $\widehat{U}(f,D)$ or another ranking-based performance measure. This question will be answered in the next experiment (Section 4.2).

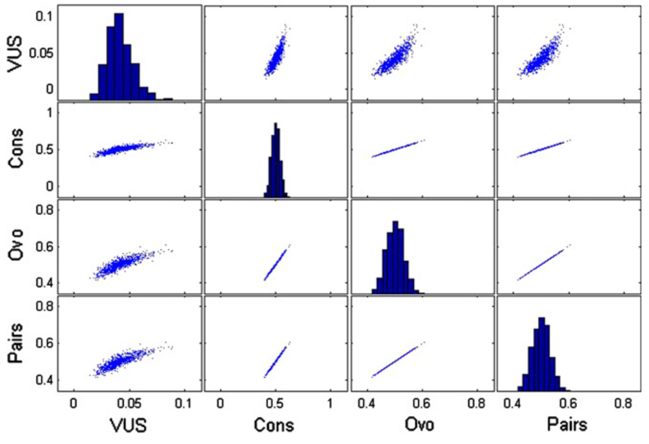
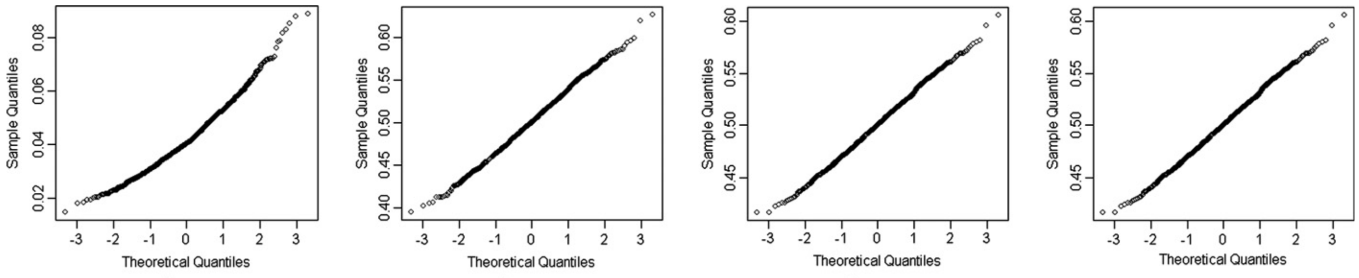


Fig. 6. Histograms and pairwise scatter plots obtained for all ranking-based measures by randomly sampling 1000 rankings of a synthetic balanced four-class data set of size 200.

When observing the histograms, the skewness of the distribution of $\widehat{U}(f,D)$ also draws the attention. This phenomenon does not occur for the other performance measures, which raises the suspicion of non-normality of the distribution of $\widehat{U}(f,D)$. For completeness, the qq-plots of the quantiles of the empirical distributions of all four measures versus the quantiles of a normal distribution are shown in Fig. 7. The figure clearly indicates that the observed distribution of $\widehat{U}(f,D)$ differs from the normal distribution. Notwithstanding the rather small sample size, this finding was confirmed by a Kolmogorov–Smirnov test ($\alpha = 0.05$). On the other hand, the qq-plots for the other three measures demonstrate the normality of their empirical distributions.

4.2. Multi-objective optimization

In this experiment we wanted to find out whether optimizing the volume under the ROC surface will lead to different ordinal regression models compared to minimizing the error rate or mean absolute error. Contrary to the previous simulations real data was analyzed this time. We picked the *Boston housing* data set from the UCI Machine learning repository. This data set consists of 506 instances with 13 features and continuous labels. In previous studies on ordinal regression (Frank and Hall, 2001; Chu and Keerthi, 2005; Chu and Ghahramani, 2005) these continuous labels were adjusted to an ordinal value by subdividing the original data set into equal frequency bins after sorting. This gave us the opportunity to control for the class frequencies and precisely because we were interested in unraveling the behavior of $\widehat{U}(f,D)$ and other measures for unbalanced data, we chose a setting with five ordinal levels and a skew class distribution: $\pi_1 = 0.3$, $\pi_2 = 0.2$, $\pi_3 = 0.1$ and $\pi_4 = \pi_5 = 0.05$ with π_k the prior probability of observing an object of class $\bar{y}_k$. As ordinal regression model (6) a simple linear model was considered, i.e. $f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x}$. Together with four thresholds this resulted in a model with 17 free parameters. To discover the optimal values of $w_1,\ldots,w_{13}$ and $b_1,\ldots,b_4$ for the various performance measures a simple multi-objective stochastic algorithm, namely particle swarm optimization (MOPSO), was implemented. MOPSO is a relatively new multi-objective optimization technique inspired by the way large bird flocks navigate through the air and searches for a set of non-dominated solutions, the so-called Pareto front. The algorithm maintains a population of particles $\vec{\mathbf{p}} = (w_1,\ldots,w_{13},b_1,\ldots,b_4)$



(a) $\widehat{U}(f,D)$ (b) $\widehat{U}_{cons}(f,D)$ (c) $\widehat{U}_{ovo}(f,D)$ (d) $\widehat{U}_{pairs}(f,D)$

Fig. 7. The quantiles of the normal distribution (x-axis) plotted versus the quantiles of the empirical distribution observed by randomly sampling 1000 rankings from a synthetic balanced four-class data set of size 200 (y-axis).

which will be adjusted in successive iterations according to the following update rules:

$$\vec{\mathbf{v}}_j \leftarrow \left(\omega - \frac{\rho_1}{10}\right)\vec{\mathbf{v}}_j + \rho_2(\vec{\mathbf{p}}_j^L - \vec{\mathbf{p}}_j) + \rho_3(\vec{\mathbf{p}}_j^G - \vec{\mathbf{p}}_j) \tag{22}$$

$$\vec{\mathbf{p}} \leftarrow \vec{\mathbf{p}}_j + \mathrm{trunc}(\vec{\mathbf{v}}_j) \tag{23}$$

The velocities $\vec{\mathbf{v}}$ allow the particles to move to new positions in successive iterations and they are guided towards the local best solution $\vec{\mathbf{p}}_j^L$, which is one of the non-dominated solutions found by particle $\vec{\mathbf{p}}_j$ so far, and towards the global best solution $\vec{\mathbf{p}}_j^G$, which is selected from the repository of non-dominated solutions found so far by all particles according to a selection scheme which assigns solutions lying in little explored regions higher chances of being selected. Besides, the function trunc truncates the components of $\vec{\mathbf{v}}$ exceeding the interval $[0,1]$, ω is an inertia weight (typically 0.4) and $\rho_1,\ldots,\rho_3$ represent random numbers selected from a uniform distribution on the interval $[0,1]$. Nowadays many methods in machine learning and statistics optimize a weighted sum of the loss function and a regularization term to control the complexity of the model and to prevent overfitting on training data. For example in a linear support vector machine and in ridge regression the norm of $\boldsymbol{w}$ serves as regularization term (see for example Cristianini and Shawe-Taylor, 2000; Hastie et al., 2001; Schölkopf and Smola, 2002 for a detailed discussion on this subject). Here the random number ρ_1, which is specific for our problem setting, acts as a regularization term controlling the complexity of the fitted models by pushing particles back towards the center of the search space (corresponding to a null model).

Six different objectives were considered, namely accuracy (or equivalently '1 – mean zero-one error'), '1 – mean absolute error', $\widehat{U}(f,D)$, $\widehat{U}_{\mathrm{cons}}(f,D)$, $\widehat{U}_{\mathrm{ovo}}(f,D)$ and $\widehat{U}_{\mathrm{pairs}}(f,D)$. The algorithm was executed 20 times for 100 iterations with a population of 500 particles and different seeds for the random generator. In all runs the non-dominating solutions found during the search were stored in a

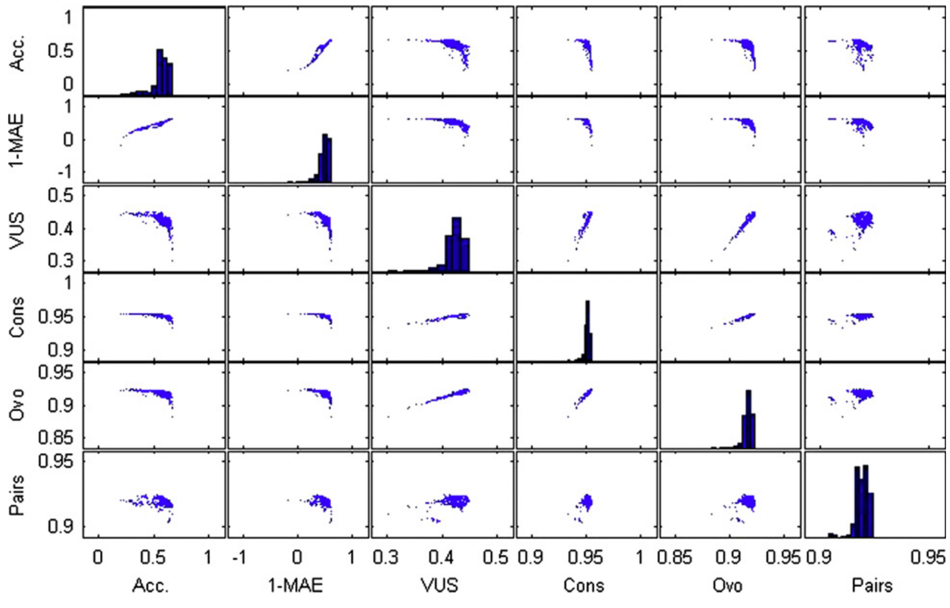


Fig. 8. The set of non-dominated solutions aggregated from 20 runs of the MOPSO-algorithm. The six-dimensional Pareto front is plotted as a matrix of two-dimensional scatter plots showing the trade-off for each pair of objectives.

repository and afterwards the global non-dominated set of these 20 repositories was computed. This set is visualized by a matrix of two-dimensional Pareto fronts in Fig. 8.

One can easily see that none of the six measures manifests a comonotone relationship with another. Accuracy and mean absolute error on the one hand and the ranking-based measures on the other hand exhibit a relatively large trade-off, as almost all solutions lie on the two-dimensional front. The ranking-based performance measures give also rise to trade-offs, but here the monotonic association is more prominent. The multi-class approaches $\widehat{U}_{\mathrm{cons}}(f,D)$ and $\widehat{U}_{\mathrm{ovo}}(f,D)$ turn out to approximate the behavior of $\widehat{U}(f,D)$ better than simply counting all correctly ordered pairs. Apparently, for $\widehat{U}_{\mathrm{pairs}}(f,D)$ the optimal models are biased towards correctly ranking the biggest classes (due to the skew class distribution the data set contains only 640 object pairs of classes $\bar{y}_4$ and $\bar{y}_5$ compared to more than 15000 object pairs of the biggest classes $\bar{y}_1$ and $\bar{y}_2$). Methods minimizing the error rate or the number of incorrect instance pairs hence will both overfit on the biggest classes.

5. Conclusion

In this article we analyzed ranking-based ordinal regression models. We argued that evaluating the ranking returned by an ordinal regression model is often more appropriate than looking at ‘mean zero-one error’ or ‘mean absolute error’, especially with skew class or cost distributions. To that end, we extended the concept of expected ranking accuracy for ordinal labeled data and showed that a nonparametric unbiased estimator $\widehat{U}(f,D)$ of this quantity corresponds to the volume under the ROC surface spanned by the true positive rates of each class. Moreover, we revealed the relationship between $\widehat{U}(f,D)$ and previous ranking-based performance measures, which can be considered as approximations of this statistic. The volume under the ROC surface and related measures do not manifest a comonotone relationship and they also have a different distribution. Consequently, algorithms optimizing different criteria will lead to different models.

These observations were confirmed with experiments on synthetic and real data. In particular, a large trade-off was discovered between ‘mean zero-one error’ and ‘mean absolute error’ on the one hand and ranking-based measures on the other hand. The latter mutually displayed smaller trade-offs, but among them $\widehat{U}_{\mathrm{pairs}}(f,D)$ turned out to concentrate too much on the biggest classes.

We conclude that all existing methods for ordinal regression, which typically minimize a loss based on error rate or the number of incorrectly ranked object pairs, might not construct appropriate models when the class or cost distributions are skew. ROC analysis offers in this case a valuable alternative allowing to pick a classifier from the surface for a specific setting of cost and the volume under the ROC surface gives a good overall indication of the quality of the model for different costs without favoring the majority classes.

Acknowledgements

We would like to thank Bruno Bauwens for useful discussions. Willem Waegeman was supported by a grant of the “Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)”.

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., Roth, D., 2005. Generalization bounds for the area under the ROC curve. J. Machine Learn. Res. 6, 393–425.

Agresti, A., 2002. Categorical Data Analysis, 2nd version, John Wiley and Sons.

Chu, W., Ghahramani, Z., 2005. Gaussian processes for ordinal regression. J. Machine Learn. Res. 6, 1019–1041.

Chu, W., Keerthi, S., 2005. New approaches to support vector ordinal regression. In: Proc. Internat. Conf. on Machine Learn., Bonn, Germany, pp. 321–328.

Cortes, C., Mohri, M., 2003. AUC optimization versus error rate minimization. In: Adv. Neural Inform. Process. Systems, vol. 16. The MIT Press, Vancouver, Canada.

Crammer, K., Singer, Y., 2001. Pranking with ranking. In: Proc. Conf. on Neural Inform. Process. Systems, Vancouver, Canada, pp. 641–647.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines. Cambridge University Press.

Dreiseitl, S., Ohno-Machado, L., Binder, M., 2000. Comparing three-class diagnostic tests by three-way ROC analysis. Med. Decis. Making 20, 323–331.

Ferri, C., Hernandez-Orallo, J., Salido, M., 2003. Volume under ROC surface for multi-class problems. In: Proc. European Conf. on Machine Learn., Dubrovnik, Croatia, pp. 108–120.

Fieldsend, J., Everson, M., 2006. Multi-class ROC analysis from a multi-objective optimization perspective. Pattern Recognition Lett. 27, 918–927.

Flach, P., 2003. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In: Proc. Internat. Conf. Machine Learn., Washington, DC, USA.

Flach, P., August 2004. The many faces of ROC analysis in machine learning. In: Tutorial Presented at the European Conf. on Machine Learn., Valencia, Spain.

Frank, E., Hall, M., 2001. A simple approach to ordinal classification. In: Proc. European Conf. Machine Learn., pp. 146–156.

Hand, D., Till, R., 2001. A simple generalization of the area under the ROC curve for multiple class problems. Machine Learn. 45, 171–186.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

Herbrich, R., Graepel, T., Obermayer, K., 2000. Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers. The MIT Press, pp. 115–132.

Higgins, J., 2004. Introduction to Modern Nonparametric Statistics. Duxbury Press.

Kramer, S., Widmer, G., Pfahringer, B., Degroeve, M., 2000. Prediction of ordinal classes using regression trees. Fundam. Inform. 24, 1–15.

Lehmann, L., 1975. Nonparametrics: Statistical Methods based on Ranks. Holden Day.

Nakas, C., Yiannoutsos, C., 2004. Ordered multiple-class ROC analysis with continuous measurements. Statist. Med. 22, 3437–3449.

Schölkopf, B., Smola, A., 2002. Learning with Kernels, Support Vector Machines, Regularisation, Optimization and Beyond. The MIT Press.

Shashua, A., Levin, A., 2003. Ranking with large margin principle: Two approaches. Proc. Internat. Conf. Neural Inform. Process. Systems. The MIT Press, Vancouver, Canada, pp. 937–944.

Yan, L., Dodier, R., Mozer, M., Wolniewiecz, R., 2003. Optimizing classifier performance via an approximation to the Wilcoxon–Mann–Whitney statistic. In: Proc. Internat. Conf. on Machine Learn., Washington DC, USA, pp. 848–855.