Journal of Artificial Intelligence Research 49 (2014) 241-267 Submitted 08/13; published 02/14

An Empirical Evaluation of Ranking Measures With Respect to Robustness to Noise

Daniel Berrar BERRAR.D.AA@M.TITECH.AC.JP
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama 226-8502, Japan

Abstract

Ranking measures play an important role in model evaluation and selection. Using both synthetic and real-world data sets, we investigate how different types and levels of noise affect the area under the ROC curve (AUC), the area under the ROC convex hull, the scored AUC, the Kolmogorov-Smirnov statistic, and the H-measure. In our experiments, the AUC was, overall, the most robust among these measures, thereby reinvigorating it as a reliable metric despite its well-known deficiencies. This paper also introduces a novel ranking measure, which is remarkably robust to noise yet conceptually simple.

1. Introduction

Various metrics exist to evaluate the performance of a predictive model, but it is often not so clear which one we should actually choose for a concrete problem at hand (Hand, 2006; Prati, Batista, & Monard, 2011; Hernández-Orallo, Flach, & Ferri, 2012; Bradley, 2013; Parker, 2013). It is also known that different metrics quantify different aspects of a model (Caruana & Niculescu-Mizil, 2004; Ferri, Hernández-Orallo, & Modroiu, 2009). In practice, the fair and objective comparison of predictive models is therefore not trivial, particularly when the data are affected by noise. Here, we consider the problem of binary classification and ranking problems, which are pervasive in numerous applications (Prati et al., 2011), ranging from web-based recommender systems and search engines to biomedical classifiers.

The goal of this study is to investigate how robust various ranking measures are to different types and different levels of noise. Particularly, we are interested in the robustness of the widely used AUC and whether recently proposed alternatives, such as the H-measure (Hand, 2009), are indeed preferable. In addition, we present a novel performance measure, called the *truncated average Kolmogorov-Smirnov statistic* (taKS). This measure is derived from the distance between the true positive rate (TPR) and false positive rate (FPR) curves, similarly to the "classic" Kolmogorov-Smirnov statistic (KS).

Surprisingly few experimental studies focused on the comparison of performance measures for predictive models. Of course, different measures may quantify different aspects of a model, and it may make little sense to compare them across different classes. But it does make sense to compare metrics within the same class, for example, ranking measures and their robustness to noise. To our knowledge, the most comprehensive study to date was carried out by Ferri et al. (2009) who investigated the relations between various performance measures and observed that these measures essentially measure quite different aspects; an observation also made by Caruana and Niculescu-Mizil (2004). A recent comparative study

by Parker (2013) focused on ranking measures; this study recommends the H-measure and advises against the AUC. A more theoretical approach to the comparison of performance measures can be found in the work of Flach (2003). Hernández-Orallo et al. (2012) provide a comprehensive view of how the different measures are related to each other.

In summary, our main insights are the following. First, among the conventional measures, the AUC was arguably the most robust across a wide range of noise levels and types. This result confirms that in fact, the AUC can be a reliable measure, although it has been criticized as incoherent and potentially misleading (Hilden, 1991; Lobo, Jiménez-Valverde, & Real, 2008; Hand, 2009; Hand & Anagnostopoulos, 2013; Parker, 2013). Therefore, our experiments lend further empirical support to the AUC as a robust measure for model evaluation and selection. Second, overall, the magnitude of the differences in robustness between the commonly used measures were not that dramatic for relatively low noise levels. Third, the proposed new measure, taKS, was also remarkably robust to noise, and it is conceptually simple with a neat geometrical interpretation.

This paper is organized is follows. First, we briefly review the ranking measures that we included in our comparative study. Then, we give the rationale for the new measure, beginning with an introductory example and then describing the formal details. In Section 4, we report the results of the comparative study involving both synthetic and real-world data sets. Section 5 concludes the paper with a discussion.

2. A Brief Review of the Investigated Ranking Measures

Let a data set contain k instances (or cases) $\mathbf{x}_i$, $i = 1..k$. With each case, exactly one class y is associated, i.e., $(y, \mathbf{x}_i)$, $y \in \{0, 1\}$, where 1 denotes the positive class and 0 denotes the negative class. Commonly, predictive models generate a numeric score s for each $\mathbf{x}_i$, which quantifies the degree of class membership of that case to a class, for example, a class posterior probability. If the data set contains only positive and negative examples, then a predictive model can either be used as a *ranker* or as a *classifier*. If the scores are expressed on an ordinal scale, the model can use the scores to rank the cases from the most to the least likely positive. By setting a threshold t on the ranking score, $s(\mathbf{x})$, such that $C\{s(\mathbf{x}) \geq t\} = 1$, we can turn the ranker into a (crisp) classifier.

2.1 Area Under the ROC Curve (AUC)

Arguably the most commonly used ranking measure is the AUC. It has been used for model selection in various applications, ranging from data mining competitions to biomedical tests (Berrar & Flach, 2012). The AUC is the area under the ROC curve, which depicts the trade-offs between the false positive rate (or 1 minus specificity, depicted on the x-axis) and the true positive rate (or sensitivity, depicted on the y-axis). These trade-offs correspond to all possible binary classifications that any dichotomization of the continuous outputs of a model would allow. The AUC is equivalent to a Wilcoxon rank-sum statistic (Bamber, 1975; Hanley & McNeil, 1983) and can be interpreted as a conditional probability: given any randomly selected positive and negative case, the AUC is the probability that the classifier assigns a higher score to the positive case (i.e., ranks it before the negative case).

Let P denote the probability that a randomly selected actual positive case, $\mathbf{x}_+$, has a higher ranking score, s_+, than a randomly selected negative case, $\mathbf{x}_-$, i.e., $s_+ > s_-$. Here,

a *higher* ranking score means that $\mathbf{x}_+$ is ranked *before* $\mathbf{x}_-$, and $f(s_+)$ and $g(s_-)$ are the distribution functions of these scores (Hilden, 1991). Following Hilden's notation, the AUC can then be written as

$$\text{AUC} = \Pr\{s_+ > s_-|\mathbf{x}_+ \text{ and } \mathbf{x}_-\} = \iint\limits_{s_+>s_-} f(s_+)\mathrm{d}s_+ \; g(s_-)\mathrm{d}s_- = \int F(s_-) \; \mathrm{d}G(s_-). \quad (1)$$

The AUC can be calculated in different ways from an empirical ROC curve; for a practical guide, see the tutorial by Fawcett (2004). ROC analysis is now an integral part of the evaluation of machine learning algorithms (Bradley, 1997). Whereas ROC curves are widely (and rightly so) considered useful, both theoretical and practical shortcomings of the AUC have been pointed out (Hilden, 1991; Adams & Hand, 1999; Bengio, Mariéthoz, & Keller, 2005; Webb & Ting, 2005; Lobo et al., 2008; Hand, 2009; Hanczar, Hua, Sima, Weinstein, Bittner, & Dougherty, 2010; Hand & Anagnostopoulos, 2013; Parker, 2013). A particular problem of the AUC is that it can be incoherent, in the sense that it assumes different cost distributions for different classifiers (Hand, 2009). One of the first criticisms, with an insightful example showing that area comparisons can be misleading, can be found in the work of Hilden (1991). Hand (2009), too, considers the AUC fundamentally incoherent. Recently, however, Hand and Anagnostopoulos (2013) showed that the AUC can be a coherent measure, but only under certain assumptions that may not hold for real applications.

2.2 Scored Area Under the ROC Curve (sAUC)

The AUC measures only how well positive and negative cases are ranked relative to each other, but it does not consider the actual ranking scores. This means that the margin between scores is irrelevant. Intuitively, however, it seems reasonable to take the scores somehow into account. Various alternatives of the AUC have been suggested that do just that; an example is the scored AUC (sAUC) (Wu & Flach, 2005; Wu, Flach, & Ferri, 2007).

Let n_+ denote the total number of positive cases and n_- denote the total number of negative cases. Let $\{s_{1+}, ...s_{n+}\}$ denote the predicted ranking scores for the positive cases and $\{s_{1-}, ...s_{n-}\}$ denote the scores for the negative cases, where $s_{1+} \geq ... \geq s_{n+}$ and $s_{1-} \leq ... \leq s_{n-}$. Both s_{i+} and s_{j-} are assumed to be normalized between $[0, 1]$, with $i = 1, ...n_+$ and $j = 1, ...n_-$. Let $I(\cdot)$ be an indicator function with $I(true) = 1$ and $I(false) = 0$. The sAUC is then defined as

$$\text{sAUC} = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} (s_{i+} - s_{j-}) I(s_{i+} > s_{j-}). \quad (2)$$

The indicator function $I(s_{i+} > s_{j-})$ assesses the ranking ability, while the factor $(s_{i+} - s_{j-})$ evaluates the score differences. Without this factor, Equation 2 is equivalent to the AUC. There exist further variants such as the soft-AUC (Calders & Jaroszewicz, 2007) and the probabilistic AUC (pAUC) (Ferri, Flach, Hernández-Orallo, & Senad, 2005); however, according to Vanderlooy and Hüllermeier (2008), none of the proposed alternatives to the AUC are effective for model evaluation. We therefore do not consider any further variants in our comparative analysis.

2.3 Area Under the ROC Convex Hull (AUCH)

The ROC convex hull is defined as the convex hull that encloses the operating points of the ROC curve (Provost & Fawcett, 2001; Flach, 2010). Note, that a curve is called convex if any straight line interpolating between two points on the curve is never above the curve.[1] The ROC convex hull results from the interpolation between the following k points, which are ordered based on increasing values of their abscissa: the origin $(x_i, y_i) = (0, 0)$, the minimum set of points spanning the concavities, and the point (1,1). The area under the ROC convex hull, AUCH, is always at least as large as the AUC. The AUCH can be calculated as shown in Equation 3.

$$\text{AUCH} = \sum_{i=1}^{k-1} y_i(x_{i+1} - x_i) + 0.5(y_{i+1} - y_i)(x_{i+1} - x_i). \tag{3}$$

2.4 H-measure

In order to address the incoherence of the AUC, Hand (2009) proposed the H-measure. Let t denote the classification threshold, and let $\text{TPR}(t)$ and $\text{FPR}(t)$ denote the corresponding true positive and false positive rate, respectively. The overall misclassification loss is then $c_+\pi_+(1-\text{TPR}(t))+c_-\pi_-(\text{FPR}(t))$, where c_+ is the cost associated with the misclassification of a positive case and c_- is the cost associated with the misclassification of a negative case, and π_+ and π_- are the prior probabilities of positive and negative cases, respectively.

$$\text{H}-\text{measure} = 1 - \frac{\int Q(T(c), c)u(c)\text{d}c}{\pi_+ \int_0^{\pi_-} cu(c)\text{d}c + \pi_- \int_{\pi_-}^{1} (1-c)u(c)\text{d}c}, \tag{4}$$

with $c = c_+/(c_+ + c_-)$; $T(c) = \arg\min_t\{c\pi_+(1-\text{TPR}(t)) + (1-c)\pi_-\text{FPR}(t)\}$; $Q(t, c) = \{c\pi_+(1-\text{TPR}(t)) + (1-c)\pi_-\text{FPR}(t)\}(c_+ + c_-)$; $u(c) = c(1-c)/\int_0^1 c(1-c)\text{d}c$.

The H-measure is a measure of the overall misclassification loss and a function of both the class-specific misclassification costs and the prior probabilities of positive and negative cases (Hand, 2009). By contrast, the AUC measures the ranking performance over the entire space of classification thresholds and is independent of costs and class priors. Flach et al. (2011) showed that, with two small variations, the H-measure is a linear transformation of the area under the cost curve, which was proposed by Drummond and Holte (2006).

2.5 Kolmogorov-Smirnov Statistic (KS)

The Kolmogorov-Smirnov goodness-of-fit test for a single sample is a test of ordinal data and assesses whether the distribution of n scores follows a specific theoretical or empirical distribution (Sheskin, 2007). The test statistic, KS, is defined as the maximum value of the absolute difference between two cumulative distributions. When we assess the ranking

1. In mathematical terms, this curve is considered "concave", whereas the standard machine learning terminology uses "convex".

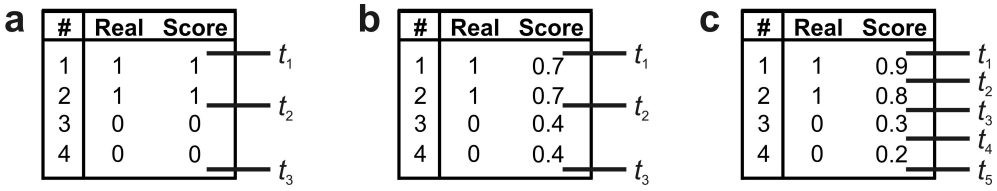

Figure 1: (a) Classification result on a toy data set of 4 cases, (a) optimal model with 3 possible thresholds, (b) another model with 3 possible thresholds; and (c) yet another model, with 5 possible thresholds (Real: real class label, 1 = positive class; Score: predicted score for the positive class).

ability of a classifier, these distributions are given by the true positive rates and false positive rates for all classification thresholds. The KS statistics has a simple geometrical interpretation as the maximum distance between the TPR and FPR curves,

$$\mathrm{KS} = \max\{|\mathrm{TPR}_i - \mathrm{FPR}_i|\}, \tag{5}$$

where TPR_i and FPR_i denote the true positive rate and false positive rate for the i^{th} threshold, respectively (see Figure 2 for an example).

3. Truncated Average Kolmogorov-Smirnov Statistic (taKS)

In this section, we propose a new ranking measure, called the *truncated average Kolmogorov-Smirnov* statistic (taKS).

3.1 Introductory Example

Let us consider the prediction results on an arbitrary test set comprising k cases, which belong to either the positive or the negative class. The optimal model will assign a score of 1 to each positive and a score of 0 to each negative case. Consider now another model that assigns the scores $s_1, s_2 ... s_k$ to the k test cases, where $s_1 \geq s_2 \geq ... \geq s_k$, which also happen to lead to a perfect ranking. Figure 1 illustrates this idea using a toy data set with $k = 4$ cases. All models in this example have indeed the same ranking performance, and consequently, ranking measures like the AUC do not distinguish between them. There is nothing wrong with that – all what matters is the relative ordering of the cases, irrespective of the actual ranking scores. Note, that the optimal model always allows exactly three possible thresholds: (1) one threshold separating the positive and negative cases (t_2 in Figure 1a); (2) one "top" threshold (i.e., TPR = 0 and FPR = 0; t_1 in Figure 1a); and (3) one "bottom" threshold (i.e., TPR = 1 and FPR = 1; t_3 in Figure 1a).

Assume now that these three models enter a data mining competition. Let us further assume that we, the judges, can only see the final predictions as shown in Figure 1. We do not have any other information about the models such as the calibration of the scores, except that higher scores reflect more relative confidence that a case belongs to the positive class. No other assumptions shall be allowed for now.

We can probably agree that model (a) is the winner, but which one should be the runner-up, (b) or (c)? To answer this question, we could consider the ranking scores, for example, by calculating the sum of squared errors (SSE) or a related measure such as the Brier score. This approach would tell us that model (c) with SSE = 0.18 is preferable to model (b) with SSE = 0.50. However, this approach makes the tacit assumption that both models, (b) and (c), produce comparable scores in the same range, maybe posterior probabilities in $[0, 1]$. This is of course often a reasonable assumption, but it does not necessarily have to be the case. In addition, any such assumption was actually not allowed.

What if we allowed assumptions about the calibration? Let us speculate that – by design – model (b) could not have produced probabilities larger than 0.7 or smaller than 0.4. Based on minimum message length theory, it may indeed make sense to prevent a model from making overly confident predictions, for example, by limiting the estimates to a specific range only (Korb, Hope, & Hughes, 2001).[2] Under this particular assumption, we may look at the performance of model (b) in a new light. In fact, the difference between the scores in (a) and (b) then reduces merely to a different scaling. Could model (b) be such a "careful" model? Granted, the assumption that all scores are comparable (e.g., from $[0, 1]$) is more plausible. But the point is that if we take the actual scores into account, then we have to make some assumptions about the models' calibrations. Furthermore, the scores of model (c) have a higher level of granularity than those of model (b), so we might say that model (c) is more refined than the "coarser" model (b). But does this refinement necessarily indicate the "better" model?

Let us now look at another difference between the models: the number of possible thresholds. This number depends on the refinement of the scores, but not on the actual values. The idea is to combine this number and the ranking performance.

3.2 Formal Details

Visually, we can represent the class discrimination by plotting both the true positive and the false positive rates as a function of the threshold in the same diagram. By interpolating the points, we obtain the corresponding TPR and FPR curves (henceforth referred to as *TPR-FPR plot*, also known as *Kolmogorov-Smirnov chart*).

Figure 2 illustrates the TPR-FPR curves for a toy data set. Each test case above the threshold is classified as a positive case. For instance, the threshold t_6 leads to 4 true positive and 1 false positive classifications because 4 positive cases and 1 negative case are located above the threshold; the corresponding rates are TPR $= \frac{4}{5}$ and FPR $= \frac{1}{5}$ (Figure 2b). Class discrimination could now be quantified in terms of the *area between the curves* (ABC).

Definition 1. Area between the curves
The area between the curves (ABC) is the area between the TPR curve and the FPR curve, which result from interpolating the true positive and false positive rates based on the n

2. The reason is that overly confident predictions that turn out to be incorrect lead to a dramatic information-theoretic penalty; for example, the penalty for an incorrect prediction with confidence 1 is $-\infty$. Korb et al. (2001), for example, allowed the range $[\min, \max] = [0.5(n+1)^{-1}, (n+0.5)(n+1)^{-1}]$, where n is the number of test cases.

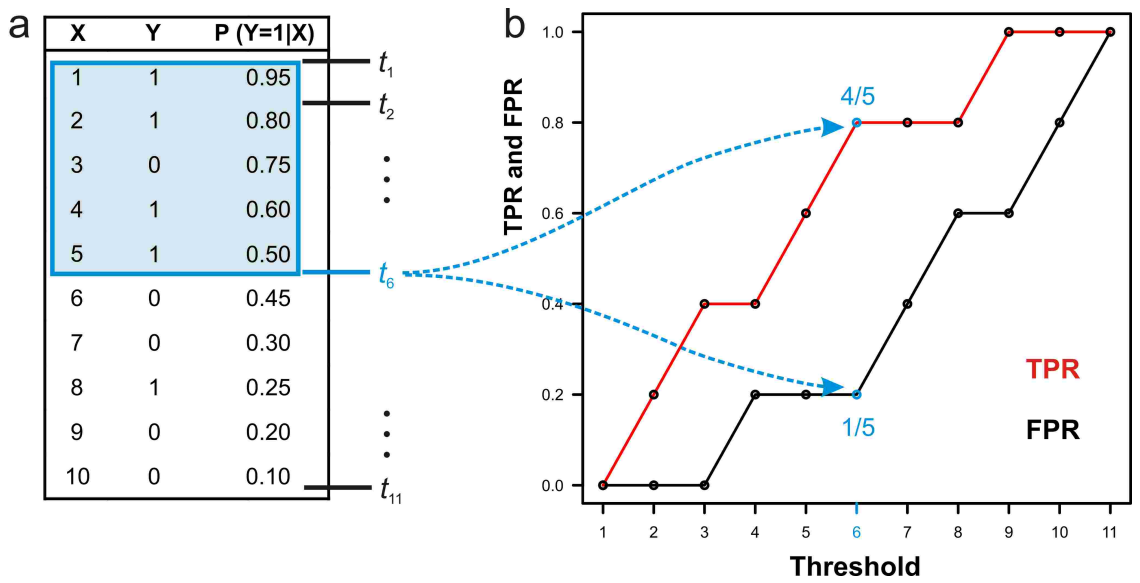

Figure 2: (a) An example of a binary classification task involving ten test cases. (b) The ranking scores allow 11 classification thresholds, each corresponding to one TPR and one FPR point. Interpolation through these points gives the TPR and FPR curves. The distance between the curves is maximal for t_6.

possible thresholds,

$$\mathrm{ABC} = |\int_1^n \mathrm{TPR}(x)\mathrm{d}x - \int_1^n \mathrm{FPR}(x)\mathrm{d}x|. \tag{6}$$

Note, that the absolute value is necessary if we accept that a model could also perform worse than random guessing, which means that the FPR curve could be above the TPR curve, thereby leading to a negative value of ABC. In the remainder of the paper, we will work with the signed ABC, though. Applying the trapezoidal rule, we obtain

$$\begin{aligned}
\mathrm{ABC} &= \sum_{i=1}^{n-1} \mathrm{TPR}_i(x_{i+1} - x_i) + \frac{1}{2}(\mathrm{TPR}_{i+1} - \mathrm{TPR}_i)(x_{i+1} - x_i) \\
&\quad - \sum_{i=1}^{n-1} \mathrm{FPR}_i(x_{i+1} - x_i) + \frac{1}{2}(\mathrm{FPR}_{i+1} - \mathrm{FPR}_i)(x_{i+1} - x_i) \\
&= \sum_{i=1}^{n-1} (x_{i+1} - x_i) \left[\mathrm{TPR}_i + \frac{1}{2}(\mathrm{TPR}_{i+1} - \mathrm{TPR}_i) - \mathrm{FPR}_i - \frac{1}{2}(\mathrm{FPR}_{i+1} - \mathrm{FPR}_i)\right] \\
&= \frac{1}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i) \left[(\mathrm{TPR}_i - \mathrm{FPR}_i) + (\mathrm{TPR}_{i+1} - \mathrm{FPR}_{i+1})\right],
\end{aligned}$$

where TPR_i and FPR_i denote the true positive rate and false positive rate for the i^{th} threshold, respectively. We now require that the thresholds on the abscissa be equidistant in $[0, 1]$, where the first threshold is linearly mapped to 0 and the last threshold, n, is mapped to 1 (this condition will be relaxed later). Then $(x_{i+1} - x_i) = \frac{1}{n-1}$. Denoting $j = i + 1$, we obtain

$$\text{ABC} = \frac{1}{2n-2}\sum_{i=1}^{n-1}(\text{TPR}_i - \text{FPR}_i) + \frac{1}{2n-2}\sum_{j=2}^{n-1}(\text{TPR}_j - \text{FPR}_j).$$

Note, that the true positive and false positive rates are always zero for the first threshold, $\text{TPR}_1 = \text{FPR}_1 = 0$, so we obtain

$$\text{ABC} = \frac{1}{n-1}\sum_{i=2}^{n-1}(\text{TPR}_i - \text{FPR}_i) \tag{7}$$

Consider the optimal model that assigns the score 1 to all cases of class 1 and the score 0 to all cases of class 0. Here, $n = 3$, and $\text{ABC} = \frac{1}{2} \cdot (1 - 0) = \frac{1}{2}$. It is now possible that another, suboptimal model has a *larger* ABC. For example, assume that the ranking score of the $i = 2$ ranked positive case is identical to the score of the $i = 1$ ranked positive case. Figure 3c shows such an example of a suboptimal model with $\text{ABC} > 0.5$. Thus, for model evaluation and selection, the ABC should not be used directly.

However, we can use the ABC to *derive* a performance measure. Consider the factor $\frac{1}{n-1}$ in Equation 7. By replacing -1 in the denominator with -2, we obtain the average of the distances between the points spanning the TPR and FPR curves, excluding the start-point and end-point. If we consider now this new measure, then the value for the optimal model is always 1, and no other model can score higher. This is immediately obvious because the thresholds on the abscissa are scaled from $[0, 1]$, and the FPR and TPR on the ordinate range from 0 to 1. Any area within these boundaries cannot be larger than 1. Thus, the distance between any pair (TPR_i,FPR_i) cannot be larger than 1, and therefore the average of the distances cannot be larger than 1.

Definition 2. Truncated average Kolmogorov-Smirnov statistic (taKS)
Let a model produce ranking scores $s_a \in \mathbb{R}$, $a = 1..k$ for k test cases belonging to either the positive or the negative class, and $s_{a-1} \geq s_a \geq s_{a+1}$ and $\exists s_a, s_b : a \neq b \wedge s_a \neq s_b$. Let $n > 2$ denote the number of possible thresholds t_i, $t_{i-1} < t_i < t_{i+1}$, that these scores allow. Let TPR_i and FPR_i denote the true positive and false positive rates, respectively, which result from a particular threshold t_i. The taKS is then defined as the average of the distances between the true positive rates and the false positive rates, excluding the points (0,0) and (1,1),

$$\text{taKS} = \frac{1}{n-2}\sum_{i=2}^{n-1}(\text{TPR}_i - \text{FPR}_i). \tag{8}$$

Note, that we can now relax the condition that the thresholds on the x-axis should be equidistant in $[0, 1]$. All what matters for the taKS are the distances between the points spanning the TPR and FPR curves; the scaling of the x-axis is irrelevant. A pseudocode for deriving taKS is given in Appendix A (algorithm 1).

3.3 Illustration of taKS

Figure 3 shows the TPR-FPR plots with the resulting taKS for nine different prediction results. The scores of the models in Figure 3a-b are different, but the relative order of the cases is the same, so the models have the same ranking performance. Here, the taKS is 1.0 for the optimal model (Figure 3a), 0.556 for the model allowing 11 thresholds (Figure 3b), and 0.600 for the model allowing 10 thresholds. These examples also illustrate that using the ABC can be misleading: we would erronously prefer the model in Figure 3c with ABC = 0.533 over the optimal model with ABC = 0.500.

The scores in (b) and (c) are different only for case #1. Model (b) predicted 0.95 whereas model (c) predicted 0.80. Provided that both models are equally calibrated, one could argue that (b) is better (has a smaller SSE), and that taKS is therefore misleading.[3] This reasoning is plausible, but the opposite could also have happened. Suppose that the model in (c) produces 0.95 for cases #1 and #2. The value of taKS remains the same (i.e., 0.60), but now it points us to the better model. Either scenario is equally likely a priori, so correct and incorrect decisions should balance, on average, for this example. This means that about half the time, taKS points us to the better model. Note, that the AUC is indifferent in these examples, making no difference between (b) and (c) in either scenario. If we used the AUC, then we could only guess which one is better, (b) or (c); thus, we would be correct about half the time, too. Consequently, compared with the AUC (or any other ranking measure, for that matter), taKS does not provide any advantage in this example – but it does not provide any disadvantage, either.

Figure 3d-f shows three models that make some ranking errors. These models have the same ranking performance, but they score a different taKS. The model in Figure 3e scores a larger taKS than the model in Figure 3f. The model in Figure 3e assigned the same score (0.80) to the cases #1 and #2; the model in Figure 3f assigned the same score (0.60) to the cases #4 and #5. While both models allow for the same number of thresholds (i.e., 10), the taKS identifies the model in Figure 3e as the preferable one because of its larger ABC.

Figure 3g-i show three particular models. The model in Figure 3g is the perfect "anti"-model that predicts like the mirror image of the optimal model. Consequently, the resulting taKS is -1. Another particular model is shown in Figure 3h. This model assigns the same score to all cases; thus, it allows only two possible thresholds, so that the TPR and FPR curves fall onto the same line. As the ABC is then not defined, the taKS is not defined, either. Figure 3i shows the results of a random prediction, leading to a taKS and ABC of zero. Note, that the AUC is defined for Figure 3h; the AUC is here 0.5 and the same as the AUC in Figure 3i.

3. Thanks to an anonymous reviewer for pointing this out!

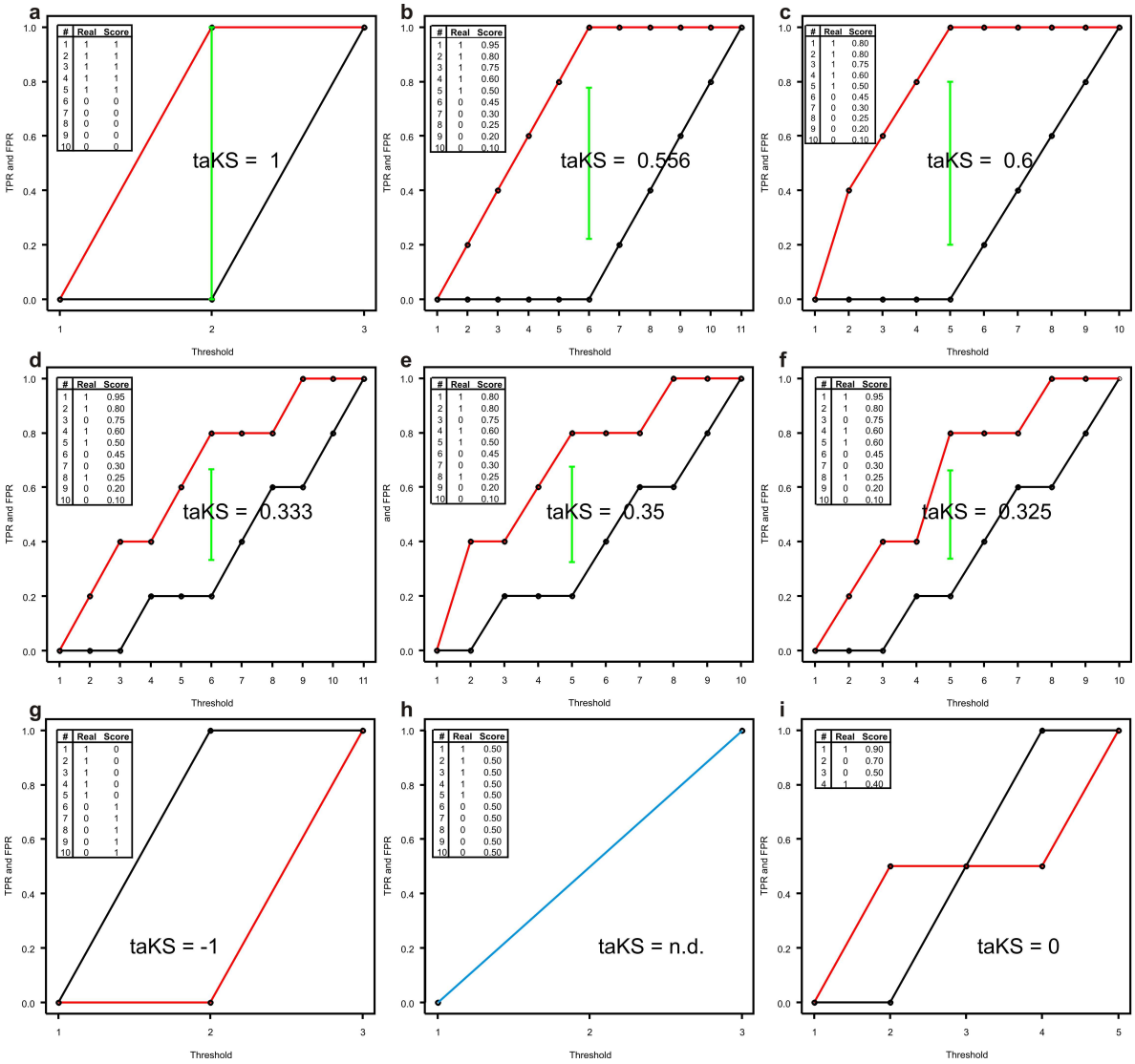

Figure 3: TPR (red) and FPR (black) curves and taKS for nine classification results (cf. inset table). (a) Best possible predictions with 3 possible thresholds; (b) perfect ranking with 11 possible thresholds; (c) perfect ranking with 10 possible thresholds; (d) prediction with ranking errors and 11 possible thresholds; (e) prediction with ranking errors and 10 possible thresholds; (f) prediction with ranking errors and 10 possible thresholds; (g) worst possible prediction with 3 possible thresholds; (h) all cases have the same ranking score, and neither ABC nor the taKS is defined; (i) random prediction with 5 possible thresholds.

3.4 Some Further Notes on taKS

As the name implies, the taKS is closely related to the Kolmogorov-Smirnov (KS) statistic, i.e., the maximum distance between the TPR and FPR curves (cf. Equation 5). By contrast, the taKS is the average distance between these curves, excluding start- and endpoint at (0,0) and (1,1), respectively. In the introductory example (Figure 2), we have KS = $\frac{3}{5}$ (for threshold t_6) and taKS = $\frac{1}{3}$.

For the optimal model, which scores 1 for all positive cases and 0 for all negative cases, taKS = 1. Also, each model that assigns the same score s_+ to all positive and the same score s_- to all negative cases has taKS = 1. Thus, the taKS does not distinguish between the optimal model and another model that, say, assigns 0.7 to all positive cases and 0.4 to all negative cases. For the worst-possible model (i.e., one that assigns 1 to all cases of class 0 and 0 to all cases of class 1), taKS = -1. The expected value of taKS for a random model is 0. If a model assigns the same score to all cases, then taKS is not defined because the number of possible thresholds is then $n = 2$, which would lead to a division by zero in Equation 8. Graphically, the TPR and FPR curves are then straight lines through (0,0) and (1,1). Note, that conventional ranking measures such as the AUC are defined in this case.

Like the AUC, taKS is an aggregate measure of performance for a final classification result. An advantage of ROC plots is that they can visualize the performance of more than just one classifier in the same diagram, in contrast to the TPR-FPR plots as used in this paper. A further limitation of taKS is the following. If two models are equally calibrated, then in some scenarios taKS can be larger for a model with a larger SSE, thereby leading us to the potentially inferior model (for an examle, cf. Figure 3b-c). On average, however, such scenarios should balance against scenarios where the opposite is the case.

4. Robustness Analysis

We considered both synthetic and real-world data sets to study the robustness of the ranking measures to various types and levels of noise. We adopted an approach similiar to the one described by Ferri et al. (2009). The main idea is the same: we consider two models, C_1 and C_2, where C_1 is the truly better model. Then, we progressively added noise. The question is whether the performance measures can still identify C_1 as the better model. A measure X_i can be considered more robust than another measure X_j if X_i is less affected by the increasing levels of noise. Below, we describe the different types of noise that we investigated in our experiments. All experiments are described in pseudocode in Appendix A and were carried out in R2.10.1 (R Development Core Team, 2009).

4.1 Synthetic Data Sets

We considered the predictions of two classifiers, C_1 and C_2, on a hypothetical test set comprising 100 cases, as described by Ferri et al. (2009). The number of test cases has arguably little influence on the experiments, provided that it is not too small. We then generated a vector of 100 real numbers by randomly sampling from a uniform distribution $[0, 1]$, which represent the ranking scores for the positive class. These membership scores can be interpreted as class posterior probabilities. The positive class label was then assigned

to all numbers ≥ 0.5, and the negative class label was assigned to the remaining numbers. Next, we randomly selected 10 scores and replaced them by a real number, which was again randomly sampled from $[0, 1]$. The resulting numbers represented the predictions of C_1. For a threshold of 0.5, the expected accuracy of C_1 is 95% because we expect that half of the new scores (i.e., 5 of 10) are wrong.

The predictions of C_2 were the same as those of C_1, except that we selected 10 further predictions at random and replaced them by a real number, again randomly sampled from $[0, 1]$. Thus, without noise, C_2 was expected to perform worse than C_1, with expected accuracy of only 90%. The difference between the two classifiers, however, was expected to become blurred for increasing levels of noise.

We considered three types of noise: misclassification noise, probability noise, and class proportion noise (described below). For each level of noise, we generated each model, C_1 and C_2, $n = 10000$ times, each time evaluating its performance based on the different ranking measures. For each measure, we then counted how many times it erroneously indicated that C_2 was better than C_1. Let $X(\cdot)$ denote the value of a performance measure X for a classifier. The error rate of a measure X is then given by

$$\epsilon(X) = \frac{1}{n}\sum_{i=1}^{n}\delta(X), \text{with} \quad \delta(X) = \begin{cases} 1 & \text{if } X(C_2) > X(C_1) \\ 0.5 & \text{if } X(C_2) = X(C_1) \\ 0 & \text{if } X(C_2) < X(C_1) \end{cases} \tag{9}$$

By plotting $\epsilon(X)$ for the measures as a function of the noise level, we can compare their resilience to noise. For example, if a measure X_i is more robust than a measure X_j, then the error rate of X_i should be consistently lower; hence, the curve for X_i should be below the curve for X_j.

4.1.1 Misclassification Noise

First, we considered noise that affects the class labels. This experiment evaluates how sensitive the measures are with respect to mislabelings. We investigated noise levels ranging from 0% (i.e., no class label was altered) to 100% (i.e., each class label was altered and determined by the flip of a coin). For a noise level of 0%, we expect that all error scores are 0 because C_1 is clearly better than C_2. By contrast, if all class labels are random, then we expect no difference between the models, so the error score should be around 0.5.

For a class label noise between 0% and about 70% (Figure 4), the error rates of the H-measure and the Kolmogorov-Smirnov statistic are slightly below those of the other measures. The AUCH is slightly more robust than the AUC and the taKS but less robust than the H-measure and the Kolmogorov-Smirnov statistic. sAUC is the least robust in this experiment. Overall, however, we do not see dramatic differences between the measures. As expected, all measures are around 0.5 for a noise level of 100%. The error rates of the AUC and the taKS are virtually identical in this experiment.

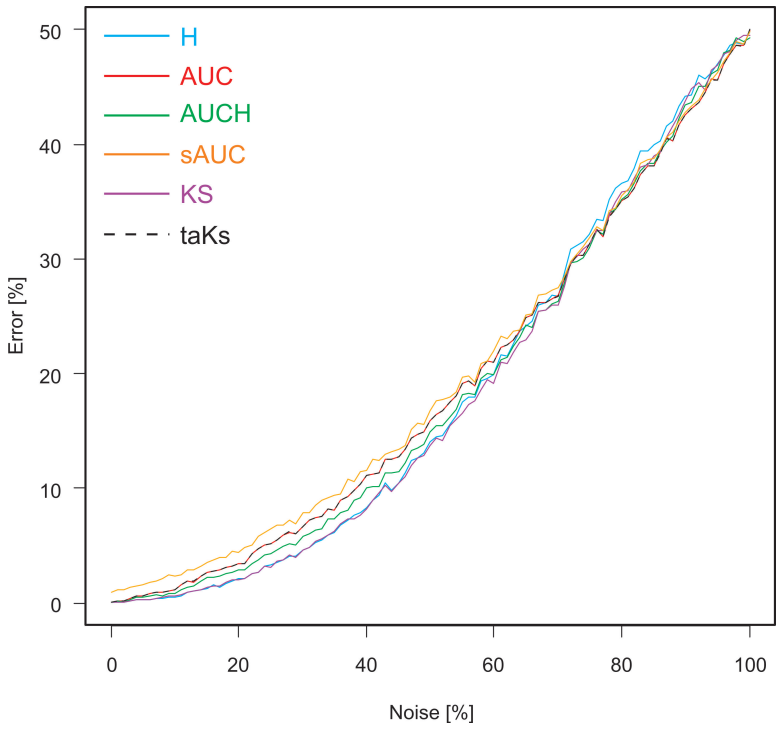

Figure 4: Synthetic data, experiment #1. Robustness to misclassification noise.

4.1.2 Probability Noise

This noise affects the class membership scores. This experiment evaluates how sensitive the measures are when the posterior class probabilities are not well estimated. The noise was randomly sampled from a uniform distribution $[-x, x]$, where x ranged from 0 (i.e., no noise) to 0.5 (i.e., 100% noise) in a stepsize of 0.005. The noise was added to all ranking scores.

When the noise affects the class posterior probabilities (Figure 5), the sAUC performs the worst. The Kolmogorov-Smirnov statistic is the next least robust, followed by the H-measure. The AUC and the taKS are the most robust in this experiment; their error rates are again almost identical. The AUCH is slightly less robust than these two measures.

4.1.3 Class Proportion Noise

This noise affects the class proportions. The experiment evaluates how sensitive the measures are to changes in the class distribution drifts. We changed the class frequencies by progressively deleting x% of the cases of the positive class. The noise x% ranged from 5% to 95%.

When the noise affects the class frequencies (Figure 6), all measures except sAUC perform very similarly. For AUC, AUCH, H-measure, KS, and taKS, the error rates are remarkably low for noise levels up to around 80%. Thus, in contrast to sAUC, these measures can cope quite well even when the classes are heavily imbalanced.

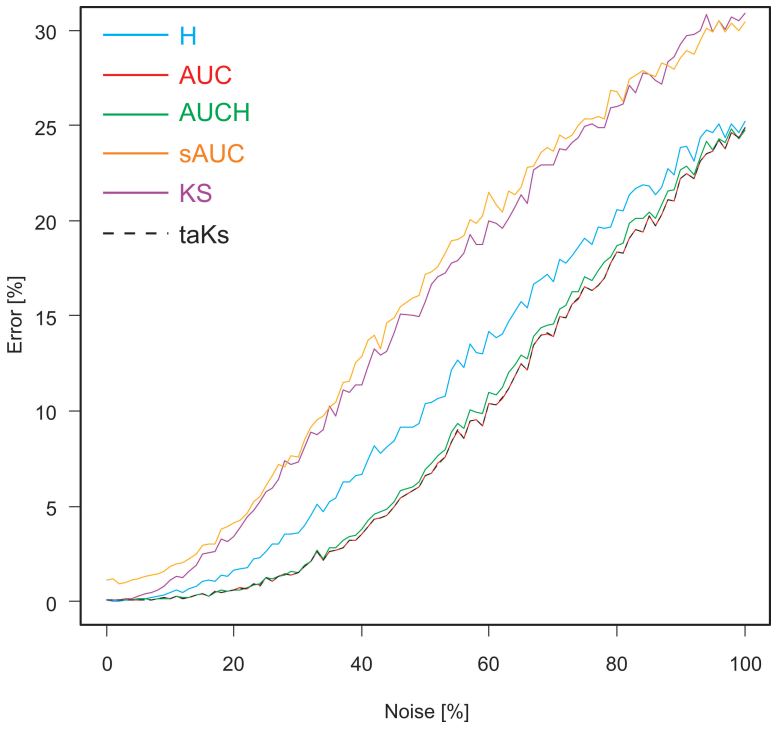

Figure 5: Synthetic data, experiment #2. Robustness to probability noise.

4.2 Real-World Data Sets

In the experiments with the synthetic data sets, we investigated a wide range of noise levels, including some that are arguably unrealistically high for real-world data sets. Therefore, we limited the next experiments to a noise level that was neither too small to cause any noticeable effect nor too large to be unrealistic. We assumed that a noise level of 10% would meet this requirement.

In the experiments with the synthetic data sets, we observed that the class proportion noise has very little effect on the performance measures, except for unrealistically high noise levels. Therefore, we excluded this type of noise in the following experiments. Instead, we considered a new type of noise that we could not study before: *attribute noise*, which affects the attributes either in the training set or the entire data set.

We used naive Bayes learning to construct our base classifier. The concrete learning algorithm was assumed to have little influence on the experimental results. We denoted the predicted scores of this classifier as C_1. We then randomly selected 10% of these scores and replaced each score by a random number, which was uniformly sampled from $[0, 1]$. The result was C_2. Without noise, C_1 is clearly better than its corrupted competitor, C_2. We used ten benchmark data sets from the UCI repository (Bache & Lichman, 2013).

Experiment #1: Misclassification Noise Affecting the Entire Data Set
In the first experiment, we investigated the resilience to noise affecting the class labels of the entire data set. For each data set, we selected 10% of the class labels and randomly

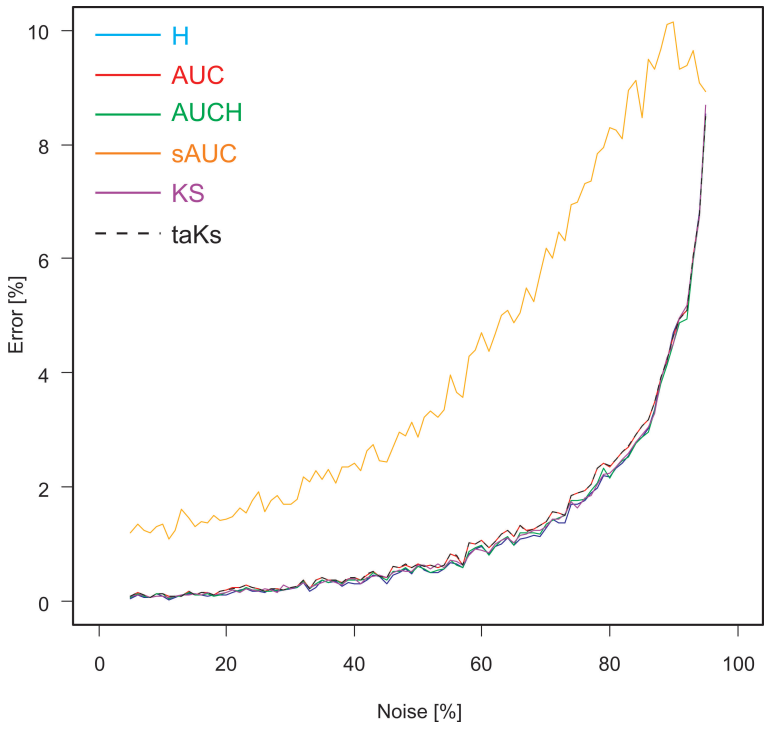

Figure 6: Synthetic data, experiment #3. Robustness to class proportion noise.

assigned either a positive or a negative label. Then, we compared the performance of C_1 and C_2 in 10-fold cross-validation. We repeated this experiment 1000 times and recorded how many times C_2 was declared the better model by the respective ranking measure (see Appendix A, algorithm 2).

Experiment #2: Misclassification Noise Affecting the Training Set
In the second experiment, we investigated the resilience to noise affecting the class labels of only the training set. For each training set, we selected 10% of the class labels and randomly assigned either a positive or a negative label. Then, we compared the performance of C_1 and C_2 in 10-fold cross-validation. We repeated this experiment 1000 times and recorded how many times C_2 was declared the better model by the respective ranking measure (see Appendix A, algorithm 3).

Experiment #3: Attribute Noise Affecting the Entire Data Set
In the third experiment, we investigated the resilience to noise affecting the attribute values in the entire data set. For each data set and each attribute, we selected 10% of the values and randomly permuted them. Then, we compared the performance of C_1 and C_2 in 10-fold cross-validation. We repeated this experiment 1000 times and recorded how many times C_2 was declared the better model by the respective ranking measure (see Appendix A, algorithm 4).

Experiment #4: Attribute Noise Affecting the Training Set
In the fourth experiment, we investigated the resilience to noise affecting the attribute values in the training set only. For each training set and each attribute, we selected 10% of the values and randomly permuted them. Then, we compared the performance of C_1 and C_2 in 10-fold cross-validation. We repeated this experiment 1000 times and recorded how many times C_2 was declared the better model by the respective ranking measure (see Appendix A, algorithm 5).

Table 1 shows the error rates of the ranking measures for the real-world data sets. We can make several interesting observations. First, for most data sets, the error rates of the performance measures are relatively small and not drastically different from each other. The only exception is sAUC, whose error rates are indeed remarkably high (between 64.8% and 78.3%) for the data sets Liver and Transfusion in all four experiments. The Liver and Transfusion data sets have only 6 and 4 attributes, respectively, and they are comparatively more difficult to classify than the other data sets.[4] For the data sets Liver and Transfusion, the error rates are the highest in all four experiments, whereas the error rates are virtually neglibile for the data set Credit. We speculate that if some data sets are intrinsically very easy to classify, then the injected noise has a negligible effect on the ranking measures. If a data set is easy to classify, then we can expect that our classifier produces scores close to 0 and 1, with fewer scores around 0.5. Now, we created C_2 by randomly selecting some scores from C_1 and re-assigning a random number from $[0, 1]$ to those scores. This means that we can expect that a larger number of more extreme scores (which are likely to be correct, as the classification tasks are relatively easy) are mapped to less extreme scores. Consequently, it is quite easy to identify C_1 as the better model, regardless of whichever measure is being used. By contrast, if a data set is intrinscically difficult to classify, then even tiny amounts of added noise may wreak havoc. This seems to be the case for sAUC in particular. Note, that the sAUC implicitly rewards a classifier's "boldness": the sAUC of a classifier with scores close to 0 and 1 can be larger than the sAUC of a classifier with less extreme scores, although the latter may make fewer ranking errors; Vanderlooy and Hüllermeier (2008, p.252) give an illustrative example.

Second, the error rates are, overall, higher when the noise affects the entire data sets than the error rates for the noise that affects only the training sets. This is not unexpected because in the latter case, a portion of the original data remains intact.

Third, overall, we observe a positive correlation between the measures, but the differences in error rates are remarkable for some data sets. For example, in experiment #2, the error rate of the H-measure (10.9%) is more than three times the error rate of the AUC (3.3%) for the data set Spect; on the other hand, the error rate of the AUC (1.4%) is seven times that of the H-measure (0.2%) for the data set House. Interestingly, the H-measure has, on average, slightly higher error rates than the AUC or taKS when the noise affects the class labels. This is somewhat unexpected because the H-measure performed relatively well in the corresponding experiments with the synthetic data sets (Figure 4). However, the differences of the average error rates are relatively small and might perhaps be explained by

4. We checked this by analyzing all (uncorrupted) data sets in 100 times repeated 10-fold cross-validation. The naive Bayes classifier achieved the lowest AUC for Liver and Transfusion.

		H-measure	AUC	AUCH	sAUC	KS	taKS
Experiment #1	Sonar	10.30	8.80	11.70	10.70	12.05	**7.00**
	Spect	13.90	8.90	9.60	**2.80**	11.20	8.80
	Heart	1.50	1.80	1.70	**0.10**	1.60	1.80
	Liver	18.10	**16.70**	17.50	70.20	19.30	16.80
	Ionosphere	0.40	0.70	0.80	**0.00**	0.50	0.50
	House	0.20	1.50	1.20	**0.00**	0.10	1.00
	Cylinder	**0.40**	1.00	1.20	7.90	0.80	0.80
	Credit	**0.00**	**0.00**	**0.00**	0.10	**0.00**	**0.00**
	Transfusion	5.50	3.40	4.60	78.30	7.30	**3.20**
	Pima	0.40	**0.00**	0.10	0.10	1.00	**0.00**
Experiment #2	Sonar	5.30	4.80	6.70	6.30	7.15	**3.40**
	Spect	10.90	3.30	4.50	**0.70**	6.95	2.80
	Heart	0.30	0.20	0.30	**0.00**	0.70	0.20
	Liver	16.50	**13.20**	14.00	66.50	17.10	**13.20**
	Ionosphere	**0.00**	**0.00**	**0.00**	**0.00**	0.10	**0.00**
	House	0.20	1.40	1.30	**0.00**	0.20	1.10
	Cylinder	**0.00**	0.10	0.10	2.90	0.10	**0.00**
	Credit	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Transfusion	3.40	0.90	1.30	69.80	2.90	**0.70**
	Pima	0.10	**0.00**	**0.00**	0.10	0.30	**0.00**
Experiment #3	Sonar	8.00	6.30	8.10	6.00	9.40	**4.30**
	Spect	8.80	3.50	3.40	**1.40**	6.90	2.90
	Heart	0.50	0.90	0.70	**0.10**	1.00	0.90
	Liver	18.40	**15.60**	16.80	64.80	21.00	**15.60**
	Ionosphere	0.10	0.30	0.30	**0.00**	0.10	0.30
	House	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Cylinder	0.30	0.20	0.30	2.50	0.70	**0.00**
	Credit	**0.00**	**0.00**	**0.00**	0.10	**0.00**	**0.00**
	Transfusion	3.90	2.10	3.00	74.60	3.80	**1.70**
	Pima	0.10	**0.00**	**0.00**	**0.00**	0.20	**0.00**
Experiment #4	Sonar	5.50	4.50	5.80	5.40	5.85	**3.20**
	Spect	12.10	4.00	5.20	**1.20**	7.10	3.40
	Heart	0.50	0.90	0.70	**0.00**	0.80	0.90
	Liver	14.90	**13.40**	13.70	65.90	17.20	13.50
	Ionosphere	0.10	**0.00**	**0.00**	**0.00**	0.20	**0.00**
	House	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Cylinder	**0.00**	**0.00**	0.10	5.10	0.30	**0.00**
	Credit	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
	Transfusion	2.50	**1.50**	2.00	70.80	2.80	**1.50**
	Pima	**0.10**	**0.10**	**0.10**	0.20	**0.10**	**0.10**

Table 1: Error rates [%] of the ranking measures. Experiment #1: 10% of the class labels in each *data set* were randomly assigned; Experiment #2: 10% of the class labels in each *training set* were randomly assigned; Experiment #3: 10% of the values of each attribute were randomly permuted in each *data set*; and Experiment #4: 10% of the values of each attribute were randomly permuted in each *training set*. Each data set was analyzed 1000 times in 10-fold cross-validation. Lowest error rates are shown in boldface.

statistical fluctuations. An alternative explanation is that in the experiments with synthetic data, all ranking scores from $[0, 1]$ were equally likely. In the experiments with real-world data sets, however, that was not the case. These data sets are relatively easy to classify. Therefore, we can expect to see more scores concentrated towards 1 and 0 and fewer scores around 0.5, which might have a negative effect on the H-measure; however, this is only speculation.

5. Discussion and Conclusions

Ranking measures play an important role in model evaluation and selection. Using both synthetic and real-world data sets, we compared the robustness of various ranking measures to different types and levels of noise. The AUC has recently been criticized as an incoherent measure (Hand, 2009; Hand & Anagnostopoulos, 2013; Parker, 2013); nonetheless, it was arguably the most robust among the conventional measures in our experiments. This is an important finding, as it lends further empirical credibility to the AUC and complements its recently published vindications (Flach, Hernández-Orallo, & Ferri, 2011; Hernández-Orallo et al., 2012; Bradley, 2013). The AUC was also more robust than the sAUC, which confirms the observations by Vanderlooy and Hüllermeier (2008) that the sAUC is not an efficient alternative to the AUC.

In our experiments with the synthetic data sets, KS and the H-measure performed best under misclassification noise. Under probability noise, however, they performed worse than the AUC and AUCH. All metrics except the sAUC performed more or less similarly under class proportion noise. Overall, the differences between the metrics with respect to their resilience to noise were rather small for relatively low noise levels in the synthetic data. Also, in most of the investigated real-world data sets, the magnitude of the difference was arguably not that dramatic. The sAUC should be used with caution, though, because it performed poorly in the experiments with synthetic data (notably class proportion noise, Figure 6) and in the experiments with the more difficult real-world data sets (Liver, Transfusion). These observations confirm earlier results, which showed that the sAUC is not robust to noise (Ferri et al., 2009).

Our experiments do not allow the conclusion that the H-measure is preferable to the AUC with respect to robustness. Also, we believe that the H-measure is arguably more intricate than the other measures, and its geometrical interpretation is not as straightforward as that of the AUC. This does of course not mean that the H-measure is not useful or that the AUC can always be trusted. Hilden (1991) describes an interesting example where the AUC is in fact misleading. Also, note that Parker (2013) comes to a conclusion that is different from ours: he recommends the H-measure, both on empirical and theoretical grounds. However, Parker evaluated a measure based on its (dis-)agreement with other measures, not based on its robustness to noise.

We also proposed a novel ranking measure, called taKS. A key characteristic of this measure is its simplicity. The taKS is easily derived, and it has a simple geometrical interpretation as the average distance between two curves: the true positive and the false positive rate curve, each plotted as a function of the classification threshold. In our study, taKS was remarkably robust to noise. However, we caution that the arguments against the AUC (Hilden, 1991; Hand, 2009; Hand & Anagnostopoulos, 2013) should not be dismissed light-

heartedly. Particularly, Parker (2013) has recently extended Hand's analysis, showing that related metrics (the area under Cohen's κ curve and average precision) are similarly incoherent. According to Parker's theorem 1, the problem is that these measures result from the integration over all possible classification thresholds. As taKS is measured via a normalized summation, it could be similarly incoherent. Our experimental results are promising, but more research is needed to elucidate the usefulness of taKS. Many open questions remain, for example, what is the precise relation between taKS and other measures, for example, the partial AUC (McClish, 1989)? When is taKS (in-)coherent? And particularly, what is the role of data set idiosyncrasies for the selection of a ranking measure? We also remember that the results of empirical studies should not be viewed in isolation but against the backdrop of previous research. The AUC was remarkably robust in our experiments, *and* it has been successfully used in numerous studies; in addition, it has recently been vindicated theoretically (Hernández-Orallo et al., 2012). Taken together, we therefore conclude that the AUC might still be a good choice for practical applications.

Finally, we note that all investigated metrics share an important caveat: as scalars, they cannot paint the full picture of a classifier's performance. By condensing the performance into a single number, we are bound to lose important information about the behavior of a model over a range of operating conditions, which is generally better described by two-dimensional plots such as ROC curves. One should always be wary of reading too much into a single number. A single number can be misleading. On the other hand, scalars have the obvious advantage that they allow us to tabulate the results of various classifiers easily. This is desirable when we compare a very large number of models, as it is generally the case in data mining competitions, for example.

Acknowledgments

We thank the three anonymous reviewers very much for their detailed and constructive comments that have greatly helped improve the manuscript.

Appendix A. Pseudocodes

Algorithm 1 Pseudocode for taKS.

Require: A matrix X with k rows (one for each test case) and 2 columns (first column: real class label; second column: predicted score for positive class, s_+). X is ordered based on decreasing values of s_+; at least two scores must be different. *# if all scores are identical then taKS is not defined.*

TPR, FPR $\leftarrow <0>$ *# lists, each containing one element: 0*
tp, fp $\leftarrow 0$
np $\leftarrow$ number of positive cases in X; nn $\leftarrow$ number of negative cases in X
i $\leftarrow 1$
while (i $\leq$ number of rows of X) **do**
  threshold $\leftarrow$ i
  ii $\leftarrow$ i
  score$_i$ $\leftarrow s_+$ of the ith case
    while (score$_{ii+1}$ == score$_i$) **and** (ii +1 $\leq$ number of cases in X) **do**
     threshold $\leftarrow$ threshold + 1
     ii $\leftarrow$ ii + 1
    end while
  tp $\leftarrow$ number of positive cases at or above threshold
  fp $\leftarrow$ number of negative cases at or above threshold
  push tp/np onto TPR; push fp/nn onto FPR
  i $\leftarrow$ threshold + 1
end while
taKS $\leftarrow$ mean(TPR\{first, last} $-$ FPR\{first, last})
return taKS

Algorithm 2 Real-world data set, experiment #1. Corrupt 10% of the class labels in the data set

Require: data set D
for $i = 1$ to 1000 **do**
  Randomly select 10% of the cases from D and randomly assign class label.
  for $k = 1$ to 10 **do**
    Sample k-th training and k-th test set from corrupted D.
    Build naive Bayes classifier from k-th training set.
    Apply classifier to k-th test set and obtain output C_{1k}.
    Derive $X_k(C_{1k})$.
    Randomly select 10% of the prediction scores of C_{1k}.
    Replace each selected score by a random number from $[0, 1]$ to obtain C_{2k}.
    Derive $X_k(C_{2k})$.
  end for
  $X(C_1) \Leftarrow$ average of $X_k(C_{1k})$.
  $X(C_2) \Leftarrow$ average of $X_k(C_{2k})$.
  if $X(C_2) > X(C_1)$ **then**
    $\epsilon(X) \Leftarrow \epsilon(X) + 1$
  else
    if $X(C_2) == X(C_1)$ **then**
      $\epsilon(X) \Leftarrow \epsilon(X) + 0.5$
    else
      $\epsilon(X) \Leftarrow \epsilon(X) + 0$
    end if
  end if
end for
return $\epsilon(X)$

Algorithm 3 Real-world data set, experiment #2. Corrupt 10% of the class labels per training set

Require: data set D

for $i = 1$ to 1000 **do**
  for $k = 1$ to 10 **do**
    Sample k-th training and k-th test set from D.
    Randomly select 10% of the training cases.
    Randomly assign a class label to the selected cases.
    Build naive Bayes classifier from k-th corrupted training set.
    Apply classifier to k-th test set and obtain output C_{1k}.
    Derive $X_k(C_{1k})$.
    Randomly select 10% of the prediction scores of C_{1k}.
    Replace each selected score by a random number from $[0, 1]$ to obtain C_{2k}.
    Derive $X_k(C_{2k})$.
  end for
  $X(C_1) \Leftarrow$ average of $X_k(C_{1k})$.
  $X(C_2) \Leftarrow$ average of $X_k(C_{2k})$.
  if $X(C_2) > X(C_1)$ **then**
    $\epsilon(X) \Leftarrow \epsilon(X) + 1$
  else
    if $X(C_2) == X(C_1)$ **then**
      $\epsilon(X) \Leftarrow \epsilon(X) + 0.5$
    else
      $\epsilon(X) \Leftarrow \epsilon(X) + 0$
    end if
  end if
end for
return $\epsilon(X)$

Algorithm 4 Real-world data set, experiment #3. Corrupt 10% of the attribute values in the data set

Require: data set D
for $i = 1$ to 1000 **do**
  Randomly select 10% of the values of each attribute of D.
  Randomly permute the selected values per attribute.
  for $k = 1$ to 10 **do**
    Sample k-th training and k-th test set from corrupted D.
    Build naive Bayes classifier from k-th corrupted training set.
    Apply classifier to k-th test set and obtain output C_{1k}.
    Derive $X_k(C_{1k})$.
    Randomly select 10% of the prediction scores of C_{1k}.
    Replace each selected score by a random number from $[0, 1]$ to obtain C_{2k}.
    Derive $X_k(C_{2k})$.
  end for
  $X(C_1) \Leftarrow$ average of $X_k(C_{1k})$.
  $X(C_2) \Leftarrow$ average of $X_k(C_{2k})$.
  if $X(C_2) > X(C_1)$ **then**
    $\epsilon(X) \Leftarrow \epsilon(X) + 1$
  else
    if $X(C_2) == X(C_1)$ **then**
      $\epsilon(X) \Leftarrow \epsilon(X) + 0.5$
    else
      $\epsilon(X) \Leftarrow \epsilon(X) + 0$
    end if
  end if
end for
return $\epsilon(X)$

Algorithm 5 Real-world data set, experiment #4. Corrupt 10% of the attribute values per training set

Require: data set D

for $i = 1$ to 1000 **do**
  for $k = 1$ to 10 **do**
    Sample k-th training and k-th test set from D.
    For the training set only: select 10% of the values of each attribute.
    Randomly permute the selected values per attribute.
    Build naive Bayes classifier from k-th corrupted training set.
    Apply classifier to k-th test set and obtain output C_{1k}.
    Derive $X_k(C_{1k})$.
    Randomly select 10% of the prediction scores of C_{1k}.
    Replace each selected score by a random number from $[0, 1]$ to obtain C_{2k}.
    Derive $X_k(C_{2k})$.
  end for
  $X(C_1) \Leftarrow$ average of $X_k(C_{1k})$.
  $X(C_2) \Leftarrow$ average of $X_k(C_{2k})$.
  if $X(C_2) > X(C_1)$ **then**
    $\epsilon(X) \Leftarrow \epsilon(X) + 1$
  else
    if $X(C_2) == X(C_1)$ **then**
      $\epsilon(X) \Leftarrow \epsilon(X) + 0.5$
    else
      $\epsilon(X) \Leftarrow \epsilon(X) + 0$
    end if
  end if
end for
return $\epsilon(X)$

References

Adams, N., & Hand, D. (1999). Comparing classifiers when the misallocation costs are uncertain. *Pattern Recognition*, *32*(7), 1139–1147.

Bache, K., & Lichman, M. (2013). UCI machine learning repository. *[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science*.

Bamber, D. (1975). The area under the ordinal dominance graph and the area below the receiver operating characteristic curve. *Journal of Mathematical Psychology*, *12*, 387–415.

Bengio, S., Mariéthoz, J., & Keller, M. (2005). The expected performance curve. *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, 9–16.

Berrar, D., & Flach, P. (2012). Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *Briefings in Bioinformatics*, *13*(1), 83–97.

Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(3), 1145–1159.

Bradley, A. (2013). ROC curve equivalence using the Kolmogorov-Smirnov test. *Pattern Recognition Letters*, *34*(5), 470–475.

Calders, T., & Jaroszewicz, S. (2007). Efficient AUC optimization for classification. In Kok, J., Koronacki, J., de Mántaras, R., Matwin, S., Mladenič, D., & Skowron, A. (Eds.), *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 42–53. Springer.

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: An empirical analysis of supervised learning performance criteria. In *Proceedings of the* 10^{th} *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 69–78. ACM Press.

Drummond, C., & Holte, R. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, *65*, 95–130.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Kluwer Academic Publishers*, 1–38.

Ferri, C., Flach, P., Hernández-Orallo, J., & Senad, A. (2005). Modifying ROC curves to incorporate predicted probabilities. In *Proceedings of the 2nd Workshop on ROC Analysis in Machine Learning*. Bonn, Germany.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, *30*, 27–38.

Flach, P. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Proceedings of the* 20^{th} *International Conference on Machine Learning*, pp. 194–201. AAAI Press.

Flach, P. (2010). ROC analysis. In Sammut, C., & Webb, G. (Eds.), *Encyclopedia of Machine Learning*, pp. 869–874. Springer.

Flach, P., Hernández-Orallo, J., & Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the* 28^{th} *International Conference on Machine Learning*, pp. 69–78.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M., & Dougherty, E. (2010). Small-sample precision of ROC-related estimates. *Bioinformatics*, *26*, 822–830.

Hand, D. (2006). Classifier technology and the illusion of progress. *Statistical Science*, *21*(1), 1–14.

Hand, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*, 103–123.

Hand, D., & Anagnostopoulos, C. (2013). When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?. *Pattern Recognition Letters*, *34*(5), 492–495.

Hanley, J., & McNeil, B. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, *148*(3), 839–843.

Hernández-Orallo, J., Flach, P., & Ferri, C. (2012). A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, *13*, 2813–2869.

Hilden, J. (1991). The area under the ROC curve and its competitors. *Medical Decision Making*, *11*(2), 95–101.

Korb, K., Hope, L., & Hughes, M. (2001). The evaluation of predictive learners: Some theoretical and empirical results. In DeRaedt, L., & Flach, P. (Eds.), *Lecture Notes in Artificial Intelligence*, pp. 276–287. Springer.

Lobo, J., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*, 145–151.

McClish, D. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making*, *9*(3), 190–195.

Parker, C. (2013). On measuring the performance of binary classifiers. *Knowledge and Information Systems*, *35*, 131–152.

Prati, R., Batista, G., & Monard, M. (2011). A survey on graphical methods for classification predictive performance evaluation. *IEEE Transactions on Knowledge and Data Engineering*, *23*(11), 1601–1618.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*(3), 203–231.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Sheskin, D. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures*. 4th Edition, Chapman and Hall, London/New York.

Vanderlooy, S., & Hüllermeier, E. (2008). A critical analysis of variants of the AUC. *Machine Learning*, *72*, 247–262.

Webb, G., & Ting, K. (2005). On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, *58*(1), 25–32.

Wu, S., & Flach, P. (2005). A scored AUC metric for classifier evaluation and selection. *Proceedings of the Second Workshop on ROC Analysis in Machine Learning*.

Wu, S., Flach, P., & Ferri, C. (2007). An improved model selection heuristic for AUC. In Kok, J., Koronacki, J., de Mántaras, R., Matwin, S., Mladenič, D., & Skowron, A. (Eds.), *Proceedings of the 18th European Conference on Machine Learning (ECML 2007)*, pp. 478–489. Springer.