Pattern Recognition Letters 30 (2009) 27–38

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

An experimental comparison of performance measures for classification

C. Ferri *, J. Hernández-Orallo, R. Modroiu

Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, València 46022, Spain

ARTICLE INFO

Article history:
Received 18 December 2006
Received in revised form 4 December 2007
Available online 2 September 2008

Communicated by T.K. Ho

Keywords:
Classification
Performance measures
Ranking
Calibration

ABSTRACT

Performance metrics in classification are fundamental in assessing the quality of learning methods and learned models. However, many different measures have been defined in the literature with the aim of making better choices in general or for a specific application area. Choices made by one metric are claimed to be different from choices made by other metrics. In this work, we analyse experimentally the behaviour of 18 different performance metrics in several scenarios, identifying clusters and relationships between measures. We also perform a sensitivity analysis for all of them in terms of several traits: class threshold choice, separability/ranking quality, calibration performance and sensitivity to changes in prior class distribution. From the definitions and experiments, we make a comprehensive analysis of the relationships between metrics, and a taxonomy and arrangement of them according to the previous traits. This can be useful for choosing the most adequate measure (or set of measures) for a specific application. Additionally, the study also highlights some niches in which new measures might be defined and also shows that some supposedly innovative measures make the same choices (or almost) as existing ones. Finally, this work can also be used as a reference for comparing experimental results in pattern recognition and machine learning literature, when using different measures.

1. Introduction

The correct evaluation of learned models is one of the most important issues in pattern recognition. One side of this evaluation can be based on statistical significance and confidence intervals, when we want to claim that one model is better than another or that one method is better than another. The other side of evaluation relies on which *metric* is used to evaluate a learned model. Evaluating a regression model with absolute error is certainly not the same as doing so with squared error. In fact, for regression, the relation and appropriateness of several evaluation measures have been analysed both theoretically and experimentally (Harrell, 2001; Ripley, 1996; Bishop, 1995), and the difference between the existing measures is sufficiently clear. The continuous character of the output (and measures) makes the task easier, especially on the theoretical level. However, for classification, there is a very extensive number of measures, some of them without a clearly justified theoretical basis, some of them recently introduced, and there is no comprehensive analysis of whether some of them bring a really new point of view when evaluating classifiers.

In this work, we concentrate on metrics for evaluating classifiers, such as Accuracy, *F*-measure, Rank Rate, Area Under the ROC Curve (AUC), Squared Error (Brier Score), LogLoss/Entropy, etc. Some of these metrics have been introduced for very different applications and, supposedly, measure quite different things. More specifically, we will use 18 different metrics, which we classify in three families as follows:

- Metrics based on a threshold and a *qualitative* understanding of error: accuracy, macro-averaged accuracy (arithmetic and geometric), mean *F*-measure (*F*-score) and Kappa statistic. These measures are used when we want a model to minimise the number of errors. Hence, these metrics are usual in many direct applications of classifiers. Inside this family, some of these measures are more appropriate for balanced or imbalanced datasets, for signal or fault detection, or for information retrieval tasks.
- Metrics based on a *probabilistic* understanding of error, i.e. measuring the deviation from the true probability: mean absolute error, mean squared error (Brier score), LogLoss (cross-entropy), two versions of the probability (rank) rate and two measures for calibration. These measures are especially useful when we want an assessment of the reliability of the classifiers, not only measuring when they fail but whether they have selected the wrong class with a high or low probability. This is also crucial for committee models (machine ensembles) to properly perform a weighted fusion of the models.
- Metrics based on how well the model *ranks* the examples: AUC (Flach et al., 2003), which for two classes is equivalent to the Mann–Whitney–Wilcoxon statistic, and is closely related to the concept of separability. These are important for many

* Corresponding author. Fax: +34 96 387 73 59.
E-mail addresses: cferri@dsic.upv.es (C. Ferri), jorallo@dsic.upv.es (J. Hernández-Orallo), emodroiu@dsic.upv.es (R. Modroiu).

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.08.010

applications, such as mailing campaign design, Customer Relationship Management (CRM), recommender systems, fraud detection, spam filtering, etc., where classifiers are used to select the best n instances of a set of data or when good class separation is crucial.

We also include two other measures, *SAUC* and *PAUC* which are *supposedly* in the middle between the last two groups.

In this paper, we analyse how these 18 metrics correlate to each other, in order to ascertain to what extent and in which situations the results obtained and the model choices performed with one metric are extensible to the other metrics. The results show that most of these metrics really measure different things and in many situations the choice made with one metric can be different from the choice made with another. These differences become larger for multiclass problems, problems with very imbalanced class distribution and problems with small datasets. But it also shows that other measures are highly correlated (or they make the same choice) and results obtained on one measure could be extrapolated to other measures. For instance, we will see that probabilistic rank rate always makes the same decision as absolute error.

The analysis is completed with a set of experiments to quantify sensitivity to four important traits which are present in some measures but not present in others. These traits are class threshold choice optimality, separability/ranking quality, calibration performance and sensitivity (or conversely robustness) to changes in prior class distribution. From this analysis, we can quantify the relations on these 'dimensions', which is a very useful complement to the results of the correlation analysis.

To our knowledge, this is the first experimental work which thoroughly compares the most generally used classifier evaluation metrics for binary and multiclass problems arriving at conclusions regarding interdependence and sensitivity of these measures. It is also the first work which gives a comprehensive taxonomy of these measures.

The paper is organised as follows: In the following Section, we explore some related work. In Section 3, we introduce the measures employed in this work and we present a first taxonomy, based on their definition. Next, in Section 4, we explain the methodology used in the experiments. The results of these experiments are detailed in Section 5. Section 6 presents four different experiments to analyse the sensitivity of the 18 measures to the four above-mentioned traits: class threshold choice, separability/ranking quality, calibration performance and sensitivity to changes in prior class distribution. Finally, Section 7 includes the conclusions, and gives some ideas for future work.

2. Related works

Several works have shown the fact that usually, given a dataset, the learning method that obtains the best model according to a given measure, is not the best method if we employ a different measure. For instance, it is said in (Huang et al., 2003) that Naive Bayes and pruned decision trees are very similar in predictive accuracy. However, using exactly the same algorithms, in (Huang and Ling, 2005) the authors show that Naive Bayes is significantly better than pruned decision trees in AUC. The different results cannot be explained here by slightly different implementations or variants of machine learning algorithms, but on the fact that the two measures (accuracy and AUC) evaluate different things.

The relationship between AUC and accuracy has been specially studied. For instance, Cortes and Mohri (2003) makes a detailed statistical analysis of the relationship between the AUC and the error rate. The results show that "the average AUC is monotonically increasing as a function of the classification accuracy, but that the standard deviation for uneven distributions and higher error rates is noticeable. Thus, algorithms designed to minimize the error rate may not lead to the best possible AUC values". On the other hand, Rosset (2004) is a surprising work, since it shows that if we use AUC for selecting models using a validation dataset, we obtain better results in accuracy (in a different test dataset) than when employing accuracy for selecting the models. Following this idea, (Wu et al., 2007) shows that an AUC-inspired measure (SAUC) is better for selecting models when we want to improve the AUC of the models. It has also been shown Domingos and Provost (2003), Ferri et al. (2003) that although pruning usually improves accuracy in decision trees, it normally decreases the AUC of the decision trees. Specifically, most of the studies on the effect of pruning on decision trees have been performed taking accuracy into account (see e.g. Esposito et al. (1997)). Not many works can be found which compare other measures in terms similar to those in which accuracy and AUC have been studied recently. An exception is Davis and Goadrich (2006) where the authors study the relationship between ROC curves and Precision-Recall curves.

In Multi-Classifier Systems (Kuncheva, 2004) many works have been devoted to study the resulting accuracy of an ensemble of combined classifiers given the original accuracies and some other conditions (Melnik et al., 2004; Narasimhamurthy, 2005; Kuncheva et al., 2003) or how to combine them in order to increase accuracy (Young Sohn, 1999). However, little is known when the performance measure is different from accuracy (Freund et al., 2003; Cortes and Mohri, 2003; Lebanon and Lafferty, 2002).

Finally, in other works Zadrozny and Elkan (2001), different probabilistic measures (MSE, log-loss and profit) are used to evaluate different methods for calibrating classifiers.

All of these works are difficult to compare and understand together especially because there is no comprehensive study of the several metrics they are using to evaluate performance.

There are some previous works that compare some performance measures for classification theoretically. Flach (2003), Fuernkranz and Flach (2005) analyse several metrics (AUC, accuracy, Fmeasure) using the ROC space. Buja et al. (2005) studies several metrics (LogLoss, squared error, and others) checking whether these are proper scoring rules, defining proper score rules as, "functions that score probability estimates in view of data in a Fisher-consistent manner". Huang and Ling (2007) is also a theoretical work on the features a metric should have and proposing new ones.

However, empirical studies have been scarce and limited in literature. The only exception to this is Caruana and Niculescu-Mizil (2004), independent and simultaneous to a preliminary work of ours (Ferri et al., 2004). Caruana and Niculescu-Mizil's work analyses the behaviour of several performance measures against a great number of machine learning techniques, the relationship between the measures using multi-dimensional scaling and correlations and, finally, derives a new central measure based on other measures. The main goal of the paper is to analyse which family of algorithms behaves best with which family of measures. We disagree on this point because each machine learning family of algorithms includes hundreds or even thousand of variants. Some of them are tuned to optimise accuracy, others to optimise MSE, others to optimise AUC, ... so a clear result on whether neural networks, or support-vector machines are better for this or other measures is, in our opinion, very difficult to state.

Additionally, the work in (Caruana and Niculescu-Mizil, 2004) only used two-class measures and relatively large datasets. Small datasets are essential, because of the size of the training and the test set is an especially important issue when comparing measures: measures based on a probabilistic view of error integrate more information than qualitative ones and, consequently, they are supposedly better for smaller datasets.

The methodology used in (Caruana and Niculescu-Mizil, 2004)'s analysis is different. Some experiments use Euclidean distances

between measures, which requires normalisation and, even with this, it is very sensitive to outliers, non-linear behaviours (e.g. MSE) or unboundness (e.g. LogLoss). Then they use correlations. They combine multi-dimensional scaling (on a 2D projection) with the results from the correlation matrix, but they do not use any clustering techniques to arrange the measures. As we will see below, there are more than two basic dimensions, so we decided not to use projections (as multi-dimensional scaling) as a result.

Finally, our work is much more exhaustive in many ways. We include many more measures, some of them very important, such as macro-averaged accuracy, the AUC variants and probability rate, we use a much larger number of datasets, we analyse the results in a segmented way for two-class vs. multiclass, small vs. large datasets, balanced vs. imbalanced, etc.

After these differences in goals and methodologies, the results obtained by Caruana and Niculescu-Mizil (2004) and the rest of works mentioned in this section, as we will see below, are more complementary than overlapping with this work.

3. Measures

In this section, we present the definition of the measures we will analyse. The selection of measures is based both on the properties of each measure (we want to cover a broad range of different measures) and their use (we want to cover the most popular ones in machine learning and pattern recognition literature).

Taking into account a broad range of different measures, we are particularly interested in three types of measures, as mentioned in the introduction: measures that are sensitive to a good choice of threshold, measures that quantify the quality of rankings (separability), and measures which quantify the deviation of the estimated probability wrt. to the actual probability. For instance, a very good classifier in terms of separability (rankings) can yield very bad accuracy if we choose a bad threshold to separate the classes. On the other hand, a classifier can have very good results for a threshold, but perform very badly for other thresholds (when costs or context changes, as ROC analysis deals with).

The difference which is sometimes most difficult to grasp is the difference between good rankings and good probabilities. A classifier can produce very good rankings, but probabilities might differ from the actual probabilities. In this case, we say that the classifier is not well calibrated. More precisely, calibration is defined as the degree of approximation of the predicted probabilities to the actual probabilities. It is usually a measure of the reliability of the prediction (DeGroot and Fienberg, 1982). If we predict that we are 99% sure, we should expect to be right 99% of the times. More formally, a classifier is perfectly calibrated if for a sample of examples with predicted probability p, the expected proportion of positives is close to p. The problem of measuring calibration is that the test set must be split into several segments or bins. If too few bins are defined, the real probabilities are not properly detailed to give an accurate evaluation. If too many bins are defined, the real probabilities are not properly estimated. A partial solution to this problem is to make the bins overlap. These different approaches have produced several measures to estimate calibration.

In fact, the relation between good class separability and calibration has been analysed in literature. The most remarkable approach is based on the so-called "decompositions of the Brier score" (Sanders, 1963; Murphy, 1972), which separate the Brier score (Mean Squared Error) measure into a reliability, a resolution, and an uncertainty components, or, alternatively, into a calibration term and refinement term. This calibration term requires binning.

Below we will introduce the definitions of 18 measures. The first 5 are qualitative and the other 13 are probabilistic.

3.1. Definition of measures

We use the following notation. Given a (test) dataset, m denotes the number of examples, and c the number of classes. $f(i,j)$ represents the actual probability of example i to be of class j. We assume that $f(i,j)$ always takes values in $\{0,1\}$ and is strictly not a probability but an indicator function. With $m_j = \sum_{i=1}^{m} f(i,j)$, we denote the number of examples of class j. $p(j)$ denotes the prior probability of class j, i.e., $p(j) = m_j/m$.

Given a classifier, $p(i,j)$ represents the estimated probability of example i to be of class j taking values in $[0,1]$. $C_\theta(i,j)$ is 1 iff j is the predicted class for i obtained from $p(i,j)$ using a given threshold θ (or decision rule, especially in multiclass problems). Otherwise, $C_\theta(i,j)$ is 0. We will omit θ below.

- ***Accuracy***: (*Acc*). This is the most common and simplest measure to evaluate a classifier. It is just defined as the degree of right predictions of a model (or conversely, the percentage of misclassification errors)

$$\mathrm{Acc} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{c} f(i,j)C(i,j)}{m}.$$

- ***Kappa statistic***: (*KapS*). This is originally a measure of agreement between two classifiers (Cohen, 1960), although it can also be employed as a classifier performance measure (Witten and Frank, 2005) or for estimating the similarity between the members of an ensemble in Multi-classifiers Systems (Kuncheva, 2004)

$$\mathrm{KapS} = \frac{P(A) - P(E)}{1 - P(E)},$$

where $P(A)$ is the relative observed agreement among classifiers, and $P(E)$ is the probability that agreement is due to chance. In this case, $P(A)$ is just the accuracy of the classifier, i.e. $P(A) = \mathrm{Acc}$ as defined above, and $P(B)$ is defined as follows:

$$P(E) = \frac{\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{c}\sum_{i=1}^{m} f(i,k)C(i,j)\right]\cdot\left[\sum_{j=1}^{c}\sum_{i=1}^{m} f(i,j)C(i,k)\right]\right)}{m^2}.$$

- ***Mean F-measure***: (*MFM*). This measure has been widely employed in information retrieval (Baeza-Yates and Ribeiro-Neto, 1999)

$$F-\text{measure }(j) = \frac{2\cdot\text{recall }(j)\cdot\text{precision }(j)}{\text{recall }(j)+\text{precision }(j)},$$

where

$$\text{recall }(j) = \frac{\text{correctly classified positives}}{\text{total positives}} = \sum_{i=1}^{m}\frac{f(i,j)C(i,j)}{m_j},$$

$$\text{precision }(j) = \frac{\text{correctly classified positives}}{\text{total predicted as positives}} = \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{\sum_{i=1}^{m_j} C(i,j)},$$

where j is the index of the class considered as "positive". Finally, mean F-measure is defined as follows:

$$\mathrm{MFM} = \frac{\sum_{j=1}^{c} F-\text{Measure }(j)}{c}.$$

- ***Macro average arithmetic***: (*MAvA*). This is defined as the arithmetic average of the partial accuracies of each class. This is usually referred as macro average (Mitchell, 1997).

$$\mathrm{MAvA} = \frac{\sum_{j=1}^{c}\frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{m_j}}{c}.$$

- ***Macro average geometric***: (*MAvG*). This is defined as the geometric average of the partial accuracies of each class.

$$\text{MAvG} = \sqrt[c]{\prod_{j=1}^{c} \frac{\sum_{i=1}^{m} f(i,j)C(i,j)}{m_j}}.$$

- ***AUC of each class against the rest, using the uniform class distribution***: (*AUNU*). The AUC (Area Under the ROC Curve) (Fawcett, 2006) of a binary classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Mann–Whitney–Wilcoxon statistic interpretation).

$$\text{AUC}\ (j,k) = \frac{\sum_{i=1}^{m} f(i,j) \sum_{t=1}^{m} f(t,k) I(p(i,j), p(t,j))}{m_j \cdot m_k}.$$

  $I(\cdot)$ is a comparison function satisfying $I(a,b) = 1$ iff $a > b, I(a,b) = 0$ iff $a < b$ and $I(a,b) = 0.5$ iff $a = b$.
  This measure has been extended for multi-class problems in more than one way. We present four variants below.
  AUNU computes the area under the ROC curve treating a c-dimensional classifier as c two-dimensional classifiers, where classes are assumed to have uniform distribution, in order to have a measure which is independent of class distribution change. Formally

$$\text{AUNU} = \frac{\sum_{j=1}^{c} \text{AUC}\ (j, \text{rest}_j)}{c},$$

  where rest_j gathers together all classes different from class j.
  Here, the area under the ROC curve is computed in the one against all approach, i.e. we compute this measure as the average of c combinations.
- ***AUC of each class against the rest, using the a priori class distribution***: (*AUNP*). This measure (Fawcett, 2001) computes the area under the ROC curve treating a c-dimensional classifier as c two-dimensional classifiers, taking into account the prior probability of each class ($p(j)$)

$$\text{AUNP} = \sum_{j=1}^{c} p(j) \text{AUC}\ (j, \text{rest}_j).$$

- ***AUC of each class against each other, using the uniform class distribution***: (*AU1U*). This metric also represents the approximation of AUC in the case of multi-dimensional classifiers, computing the AUC of $c(c-1)$ binary classifiers (all possible pairwise combinations) and considering uniform distribution of the classes (Hand and Till, 2001)

$$\text{AU1U} = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k \neq j}^{c} \text{AUC}\ (j,k).$$

- ***AUC of each class against each other, using the a priori class distribution***: (*AU1P*). In order to complete all the reasonable extensions of AUC for more than two classes, we define a final AUC-based measure. This measure represents the approximation of AUC in the case of multi-dimensional classifiers, computing AUC of $c(c-1)$ binary classifiers and considering the a priori distribution of the classes

$$\text{AU1P} = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k \neq j}^{c} p(j) \text{AUC}\ (j,k).$$

- ***Scored AUC***: (*SAUC*). is a variant of the AUC, which includes probabilities in the definition (Wu et al., 2007). The idea is to introduce a rank measure which is robust to rank changes due to small probability variations.
  First, Scored AUC for two classes is defined as

$$\text{Scored AUC}(j,k) = \frac{\sum_{i=1}^{m} f(i,j) \sum_{t=1}^{m} f(t,k) I(p(i,j), p(t,j)) \cdot (p(i,j) - p(t,k))}{m_j \cdot m_k},$$

  which is equal to AUC (i,j) except from an additional factor $(p(i,j) - p(t,k))$, which is added to quantify the deviation in probability estimation whenever the rank is incorrect.
  From the binary AUC, the multiclass SAUC measure is defined as

$$\text{SAUC} = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k \neq j}^{c} \text{Scored AUC}\ (j,k).$$

- ***Probabilistic AUC***: (*PAUC*). is also a variant of the AUC which includes probabilities in the definition (Ferri et al., 2004) in more or less the same line as SAUC, although the part with the indicator function is no longer used. This means that it is not a proper rank measure.First, Probabilistic AUC for two classes is defined as

$$\text{Prob AUC}\ (j,k) = \frac{\sum_{i=1}^{m} \frac{f(i,j)p(i,j)}{m_j} - \sum_{i=1}^{m} \frac{f(i,k)p(i,j)}{m_k} + 1}{2}.$$

  From the binary AUC, the multiclass PAUC measure is defined as

$$\text{PAUC} = \frac{1}{c(c-1)} \sum_{j=1}^{c} \sum_{k \neq j}^{c} \text{Prob AUC}\ (j,k).$$

- ***Macro Average Mean Probability Rate***: (*MAPR*). It is computed as an arithmetic average of the mean predictions for each class (see Mitchell, 1997)

$$\text{MAPR} = \frac{\sum_{j=1}^{c} \frac{\sum_{i=1}^{m} f(i,j)p(i,j)}{m_j}}{c}.$$

- ***Mean Probability Rate***: (*MPR*). This measure is also a measure which analyses the deviation from the true probability. It is a non-stratified version of the previous one, the arithmetic average of the predicted probabilities of the actual class (Lebanon and Lafferty, 2002)

$$\text{MPR} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} f(i,j)p(i,j)}{m}.$$

- ***Mean Absolute Error***: (*MAE*). This metric shows how much the predictions deviate from the true probability and it only differs from the previous one in that the product is changed by the absolute value of the difference

$$\text{MAE} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} |f(i,j) - p(i,j)|}{m \cdot c}.$$

- ***Mean Squared Error***: (*MSE*). This is just a quadratic version of MAE, which penalises strong deviations from the true probability. This metric is also known as Brier score (Brier, 1950) and integrates calibration and other components, usually grouped under the term 'refinement'

$$\text{MSE} = \frac{\sum_{j=1}^{c} \sum_{i=1}^{m} (f(i,j) - p(i,j))^2}{m \cdot c}.$$

- ***LogLoss***: (*LogL*). This is also a measure of how good probability estimates are (also known as cross entropy) and it has been used when calibration is important (Good, 1952, 1968; Dowe et al., 1996)

$$\text{LogL} = \frac{-\sum_{j=1}^{c} \sum_{i=1}^{m} (f(i,j) \log_2 p(i,j))}{m}.$$

  To avoid the case of $\log_2(0), \log_2 p(i,j)$ is computed as $\log_2(\max(p(i,j), \epsilon))$, where ϵ will be set to 0.00001 for the experiments.
- ***Calibration Loss***: (*CalL*). In (Fawcett and Niculescu-Mizil, 2007) and, independently, in (Flach and Takashi Matsubara, 2007), the relationship between the AUC-based measures, and ROC analysis in general, with calibration has been clarified. A perfectly calibrated classifier always gives a convex ROC curve. A method

for calibrating a classifier is to compute the convex hull or, equivalently, to use isotonic regression. In (Flach and Takashi Matsubara, 2007), they derive a decomposition of the Brier Score into calibration loss and refinement loss. Calibration loss is defined as the mean squared deviation from empirical probabilities derived from slope of ROC segments

$$\text{CalLoss}\ (j) = \sum_{b=1}^{r_j} \sum_{i \in s_{j,b}} \left(p(i,j) - \sum_{i \in s_{j,b}} \frac{f(i,j)}{|s_{j,b}|} \right)^2,$$

where r_j is the number of segments in the ROC curve for class j, i.e. the number of different estimated probabilities for class $j : |\{p(i,j)\}|$. Each ROC segment is denoted by $s_{j,b}$, with $b \in 1..r_j$, and formally defined as

$$s_{j,b} = \{i \in 1,\ldots,m | \forall k \in 1,\ldots,m : p(i,j) \geqslant p(k,j) \wedge i \notin s_{j,d}, \forall d < b\}.$$

From the previous binary CalLoss, the general multiclass Calibration Loss measure is defined as

$$\text{CalL} = \frac{1}{c} \sum_{j=1}^{c} \text{CalLoss}\ (j).$$

- ***Calibration by Bins***: (*CalB*). A calibration measure based on overlapping binning is CAL (Caruana and Niculescu-Mizil, 2004). This is defined as follows. For each class, we must order all cases by predicted positive class $p(i,j)$, giving new indices i^*. Take the 100 first elements (i^* from 1 to 100) as the first bin. Calculate the percentage of positives (class j) in this bin as the actual probability, $\hat{f}_j$. The error for this bin is $\sum_{i^* \in 1,\ldots,100} |p(i,j) - \hat{f}_j|$. Take the second bin with elements from 2 to 101 and compute the error in the same way. At the end, average the errors. The problem of using 100 as (Caruana and Niculescu-Mizil, 2004) suggest is that it might be a much too large bin for small datasets. Instead of 100 we set a different bin length, $s = m/10$, to make it more size-independent. Formally:

$$\text{CAL}(j) = \frac{1}{m-s} \sum_{b=1}^{m-s} \sum_{i^*=b}^{b+s-1} \left| p(i^*,j) - \frac{\sum_{i^*=b}^{b+s-1} f(i^*,j)}{s} \right|.$$

We indicate with i^* that indices are ordered by $p(i,j)$. For more than two classes, the measure is the average for all classes, i.e.

$$\text{CalB} = \frac{1}{c} \sum_{j=1}^{c} \text{CAL}(j).$$

Some of the previous measures (MFM, MAvA, MAvG, AUC variants, SAUC, PAUC, MAPR, MPR, CalL and CalB) have to be carefully implemented to exclude any class for which the test set has no instances.

3.2. Taxonomy of measures according to their properties

Previously, we mentioned that from the 18 measures, the first 5 measures are qualitative and the remaining 13 are probabilistic. This is now clear from the definitions if we just check that the first 5 use the term $C(i,j)$ in their definition (which is compared to the actual $f(i,j)$). So, the first 5 measures are sensitive to the class threshold. The other 13 measures do not use the term $C(i,j)$ in their definition but use the term $p(i,j)$. This is the estimated probability which is compared to the actual probability $f(i,j)$.

We can enrich the previous analysis if we also consider whether the measure takes into account the ranking (this corresponds to the $I(p(i,j),p(t,j))$ term) but not the direct value of the probability estimation. Additionally, we can also analyse whether the measures are sensitive to class frequency changes or not.

In Table 1, we indicate whether each of the 18 measures is influenced or not by changes in these four traits: changes in class thresholds, changes in calibration which preserve the ranking, changes in ranking which do not cross the class thresholds (but usually affect calibration), and changes in class frequency.

Table 1
Characterisation of measures according to different traits

Measure	Class threshold	Calibration	Ranking	Class frequencies
Acc	Yes	No	No	Yes
KapS	Yes	No	No	Yes
FME	Yes	No	No	Partially
MAVA	Yes	No	No	No
MAVG	Yes	No	No	No
AU1u	No	No	Yes	No
AU1p	No	No	Yes	Yes
AUnu	No	No	Yes	No
AU1p	No	No	Yes	Yes
SAUC	No	Yes	Yes	No
PAUC	No	Yes	Yes	No
MAPR	No	Yes	Yes	No
MPR	No	Yes	Yes	Yes
MAE	No	Yes	Yes	Yes
MSE	No	Yes	Yes	Yes
LogL	No	Yes	Yes	Yes
CalL	No	Yes	Yes	No
CalB	No	Yes	Yes	Yes

As can be seen in the table, according to the first three traits, threshold, calibration and ranking, the measures can be grouped as those focused on error (yes, no, no), those focused on ranking (no, no, yes) and those focused on probabilities (no, yes, yes). The fourth trait, sensitivity to class frequency change is present in some of them.

Some interesting things can be observed from the table. The most surprising issue is that there is no measure which has a 'Yes' in both the class threshold column and either the calibration or ranking columns. This means that, to date, and as far as we know, there is no measure which simultaneously combines the threshold and the estimated probability. In fact, in the definition of the 18 measures, none of them use $C(i,j)$ and $p(i,j)$ at the same time. This would be a good niche to study in the future, especially because in some applications the deviation from the actual probability is only relevant when the classifier fails. For instance, measures with a term like $f(i,j) \cdot C(i,k) \cdot p(i,k)$ might be analysed. Another interesting observation from the table is that many measures have exactly the same characterisation for the four traits, so the differences can only be shown from a quantitative analysis. For instance, it seems that LogL is more sensitive to calibration than MSE, and MSE is more sensitive than MAE, but we cannot quantify these difference from the previous table. The experimental analyses in Sections 5 and 6 will confirm one of these statements and refute the other.

4. Methodology

The experiments were performed using Witten and Frank (2005), which we extended with several new metrics, not included in the current distribution. We used six well-known machine learning algorithms: J48, Naive Bayes, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbour, AdaBoost with ten J48 trees and we performed the experiments with 30 small and medium-size datasets included in the machine learning repository (Blake and Merz, 1998), 15 of them being two-class (binary) problems and 15 of them being multiclass. Also half of them are considered to be balanced, and the rest, imbalanced. Table 2 includes further details of the datasets (size, number of classes, number of nominal attributes, number of numerical attributes, percentage of the majority class). The datasets which we consider balanced are highlighted in bold.

Table 2
Datasets used in the experiments

#	Datasets	Size	Classes	Nom.	Num.	%Maj–%Min.
1	Autos5c	202	5	10	15	33.16–10.89
2	Balance Scale	625	3	0	4	46.08–7.84
3	Breast Cancer	286	2	0	9	70.27–29.72
4	Chess	3196	2	36	0	52.22–47.48
5	Cmc	1473	3	7	2	42.70–22.61
6	Credit rating	690	2	9	6	55.50–44.50
7	Dermatology	366	6	33	1	30.60–5.46
8	German–credit	1000	2	13	7	70.00–30
9	Glass	214	6	9	0	35.51–4.2
10	Heart–statlog	270	2	13	0	55.55–44.45
11	Hepatitis	155	2	14	5	79.35–20.65
12	House voting	435	2	16	0	54.25–45.75
13	Ionosphere	351	2	0	34	64.10–35.9
14	Iris plan	158	3	0	4	33.33–33.33
15	Monks1	556	2	6	0	50–50
16	Monks2	601	2	6	0	65.72–34.27
17	Monks3	554	2	6	0	51.99–48.01
18	New thyroid	215	3	0	5	69.97–13.95
19	Pima	768	2	0	8	65.10–34.90
20	Sick	3772	2	22	7	93.87–6.12
21	Soybean	683	19	31	0	13.46–1.17
22	Segmentation	2310	7	0	19	14.28–14.28
23	Spect	80	2	0	44	50–50
24	Tae	151	3	2	3	34.43–32.45
25	Tic–tac	958	2	8	0	65.34–34.65
26	Vehicle3c	846	3	0	18	51.41–23.52
27	Waveform	5000	3	0	21	33.92–33.06
28	Wine	178	3	0	13	39.88–26.97
29	Wovel	990	11	3	11	9.09–9.09
30	Zoo	101	7	16	1	40.6–3.96

The above mentioned models were evaluated using 20×5 fold cross-validation, each of the 6 models being applied to each of the 30 datasets, getting 600 results for each dataset, making 18,000 results in total. We set up seven types of analysis: an overall analysis for all datasets, an analysis for binary and multiclass problems, for balanced and imbalanced problems, and for short datasets and large datasets. In each case we calculated the Pearson (standard) linear correlation and Spearman rank correlation between all eighteen metrics. Apart from the global view for all the datasets which will be presented using linear and rank correlations, the other six analyses will only be shown for rank correlations.

It is important to remark that we compute the correlation for each dataset, i.e., we analyse the results of the 6 models for one dataset and the corresponding 100 combinations of the cross-validation. Not merging results from different datasets is crucial, since measure values are influenced in very different ways depending on the dataset, e.g. number of classes, imbalance, problem difficulty, etc. Consequently, we construct one correlation matrix *per dataset*. Finally, we average (arithmetically) the 30 correlation matrices. The results when we use the Pearson correlation indicate the strength and direction of a linear relationship between two measures, while the rank correlation assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. A high rank correlation between two measures means that for the same problem these two measures have ranked the 6 models similarly. In other words, both measures would usually select the same model. Unlike the standard correlation coefficient, the assumption that the relationship between the variables is linear is not required. In order to avoid negative values for correlation we worked with 1-MAE, 1-MSE, 1-LogL, 1-CalL and 1-CalB.

Since there are hence eight correlation matrices, and these are difficult to understand at a glance, we will use dendrograms for representation; where the linkage distance is defined as $(1 - \mathit{correlation})$. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by a clustering algorithm. This kind of diagram has several advantages: we can easily visualise the clusters formed by the measures, as well as the linkage distance among clusters. It is also quite easy to find out the number of clusters and their components once we have selected a linkage distance. There are several methods for constructing a dendrogram using a linkage distance. We will use the "average group distance" method, which joins an existing group to the element (or group) whose average distance to the group is minimum.

5. Analysis of results

In this section, we discuss some of the interesting outcomes we found from the analysis of the correlation between metrics. First we analyse the correlation matrix (both linear and rank) for all datasets, as shown in Table 3.

In Fig. 1, we show dendrograms built from the obtained linear and rank correlations using all the available results. This figure represents the relations between the measures in an abridged and more comprehensible way.

The correlations shown on the matrix for both kinds of correlations, as well as both dendrograms are very similar (rank correlations are slightly higher than linear correlations, as expected). Consequently, there is no point in replicating the analysis. Hence, we will focus on the results and dendrogram for rank correlation.

A general observation is that all correlations are positive, and usually strong (greater than 0.5). The only clear exceptions are some correlations between LogL and some probabilistic measures, which we might consider in the same family a priori. This is mostly due to the fact that LogL is an unbounded measure, i.e, a 0 wrong probability has an infinity penalty (or very high if $\log(0)$ is avoided in implementations).

A first specific observation is the close relationship between all the qualitative measures: MAvA and MAvG, MFM, Acc and KapS, as expected. Although not exactly equal, their choices are almost the same. A second specific observation can be made with the ranking measures (AU*). The 4 variants of AUC behave quite similarly, so they can even be used interchangeably. This means that previous studies in literature using these different variants for evaluating rankers can be contrasted safely, independently of which variant they have used. Additionally, it is interesting to note that no other measure correlates to AUC more than 0.82, justifying the use of the AUC as a genuinely different and compact measure/family. Finally, on probabilistic measures, there is a clear equivalence between MPR and MAE, which is not surprising if we take a look at their definitions. The same happens for PAUC and MAPR. In fact, it is shown that PAUC has no relation whatsoever with AUC. These four measures, jointly with SAUC collapse at a linkage distance of 0.1, which means that all of them are very similar. MSE behaves differently, as does the fore-mentioned LogL, which seem to be out of this group.

In fact, if we take a look to dendrograms, and using linkage distance 0.1, we discover 7 clusters: AUC measures, qualitative measures, 'plain' probabilistic measures (MPR, MAE, PAUC, MAPR) with SAUC, and then 4 isolated measures: MSE, LogL, CalL and CalB. MSE and LogL use a quadratic or logarithmic function on the probabilities, which might explain their distance to the other probabilistic measures. The two calibration measures are outsiders because they are not proper performance measures and they try to recognise the degree in which probabilities are calibrated. Their closest measure is MSE (correlations about 0.7), since MSE can be decomposed into a calibration term and other terms. However, it is interesting to see that neither calibration measures correlate well (0.42), which suggest that both measures of calibration are significantly different (partly because bins for CalL are smaller).

Table 3
Linear (bottom-left) and rank (top-right) correlation results for all datasets

–	Acc	KapS	MFM	MavA	MavG	A1U	A1P	ANU	ANP	Sauc	Pauc	Mapr	MPR	MAE	MSE	LogL	CalL	CalB
Acc		0.98	0.95	0.92	0.83	0.70	0.72	0.70	0.72	0.68	0.76	0.76	0.79	0.79	0.88	0.40	0.55	0.58
KapS	0.98		0.97	0.95	0.88	0.72	0.74	0.72	0.73	0.71	0.79	0.79	0.78	0.78	0.87	0.40	0.55	0.56
MFM	0.95	0.97		0.98	0.93	0.71	0.71	0.70	0.69	0.74	0.81	0.81	0.78	0.78	0.85	0.38	0.54	0.57
MAvA	0.90	0.94	0.97		0.95	0.73	0.71	0.71	0.70	0.75	0.81	0.81	0.74	0.74	0.82	0.38	0.51	0.52
MAvG	0.85	0.90	0.95	0.98		0.67	0.66	0.66	0.64	0.75	0.80	0.80	0.72	0.72	0.75	0.32	0.47	0.52
AU1U	0.69	0.71	0.70	0.72	0.69		0.99	1.00	0.97	0.47	0.58	0.58	0.52	0.52	0.81	0.67	0.45	0.48
AU1P	0.71	0.73	0.70	0.71	0.68	0.98		0.99	1.00	0.46	0.57	0.57	0.52	0.52	0.82	0.67	0.46	0.48
AUNU	0.69	0.71	0.70	0.71	0.68	1.00	0.99		0.98	0.45	0.56	0.56	0.50	0.50	0.80	0.67	0.46	0.46
AUNP	0.71	0.73	0.69	0.70	0.67	0.97	0.99	0.98		0.45	0.55	0.55	0.51	0.51	0.81	0.67	0.46	0.46
SAUC	0.67	0.71	0.73	0.74	0.75	0.45	0.45	0.43	0.43		0.97	0.97	0.92	0.92	0.61	0.03	0.32	0.60
PAUC	0.74	0.77	0.79	0.80	0.80	0.55	0.55	0.54	0.53	0.97		1.00	0.95	0.95	0.72	0.14	0.40	0.65
MAPR	0.74	0.77	0.79	0.80	0.80	0.55	0.55	0.54	0.53	0.97	1.00		0.95	0.95	0.72	0.14	0.40	0.65
MPR	0.78	0.77	0.77	0.72	0.70	0.48	0.49	0.47	0.48	0.90	0.93	0.93		1.00	0.73	0.11	0.42	0.69
MAE	0.78	0.77	0.77	0.72	0.70	0.48	0.49	0.47	0.48	0.90	0.93	0.93	1.00		0.73	0.11	0.42	0.69
MSE	0.88	0.87	0.85	0.81	0.76	0.80	0.81	0.79	0.81	0.58	0.70	0.70	0.71	0.71		0.63	0.67	0.66
LogL	0.47	0.47	0.45	0.45	0.42	0.73	0.74	0.73	0.73	0.08	0.20	0.20	0.17	0.17	0.67		0.55	0.24
CalL	0.61	0.61	0.59	0.55	0.53	0.47	0.48	0.47	0.48	0.38	0.46	0.46	0.50	0.50	0.70	0.50		0.29
CalB	0.61	0.59	0.59	0.55	0.53	0.49	0.50	0.48	0.48	0.57	0.64	0.64	0.67	0.67	0.69	0.31	0.42	

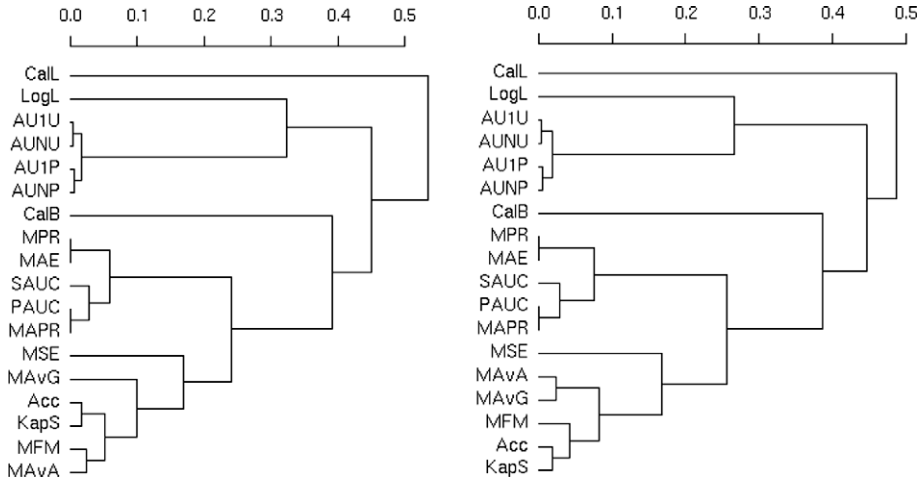

Fig. 1. Dendrograms of standard correlations (left) and rank correlations (right) between the metrics for all datasets.

If we use linkage distance 0.3, we discover 4 clusters. One with the AUC measures and LogL, a second cluster with all the probabilistic and qualitative measures (including MSE), and a third and fourth isolated measures: CalL and CalB. It is remarkable that MSE finds its highest correlation with the qualitative measures (Acc in particular) and LogL with the AUC measures. Other surprising results are the low correlation between the AUC measures and some of the derived measures (SAUC and PAUC), especially SAUC, which has an AUC-based definition, but only shows correlations of about 0.45. LogL shows the worst correlation of all with the group of 'plain' probabilistic measures (MPR, MAE, PAUC, MAPR, SAUC), mainly due to its logarithmic behaviour.

Despite the methodology being different, these results are consistent with (Caruana and Niculescu-Mizil, 2004), our previous preliminary results (Ferri et al., 2004) and other works we have referred to in Section 2, the only difference being that we do not find a strong correlation between MSE and LogL, which was found on these two works, and might be found in the implementation of LogL, which might avoid $log(0)$ in different ways and also because (Caruana and Niculescu-Mizil, 2004) only analyses two class problems.

In fact, if we compare the correlation results of 2-class problems with multiclass problems (see Fig. 2), we get some expected results. All the AUC variants collapse for 2 classes, since they are all extensions of multiclass problems but equivalent to 2-class problems. The rest of the correlations are similar in both cases although a little bit stronger for multiclass problems. The only big difference is that MSE is joined to the AUC measures in the 2-class datasets and not to the qualitative measures. This suggests that MSE behaves differently for 2-class problems and multiclass problems.

If we compare the correlations for the datasets with balanced class distribution against the correlations for the datasets with imbalanced class distribution (see Fig. 3), the results show more variations. Correlations are much lower for imbalanced datasets, and the way in which the definition of each measure mixes the partial functions for each class is very relevant. For instance, qualitative measures are very close for balanced datasets. There is virtually no difference between Acc, KapS, MAvG, MFM and MAvA for balanced datasets, while it is amplified for imbalanced dataset. In fact, CalL is associated with qualitative measures for balanced datasets, but not for imbalanced datasets. The same thing occurs for LogL, which is associated with AUC measures for balanced datasets, but not for imbalanced datasets. Finally, it happens conversely with CalB. It is associated with 'plain' probabilistic measures for imbalanced datasets, but not for balanced datasets. This highlights the relevance of metric choice depending on the class balance of the dataset.

Finally, the results for relatively small vs. large datasets (see Fig. 4), is significant around linkage distance of 0.2. While qualitative measures and AUC measures are joined for small datasets, and it is the probabilistic measures which gather more information

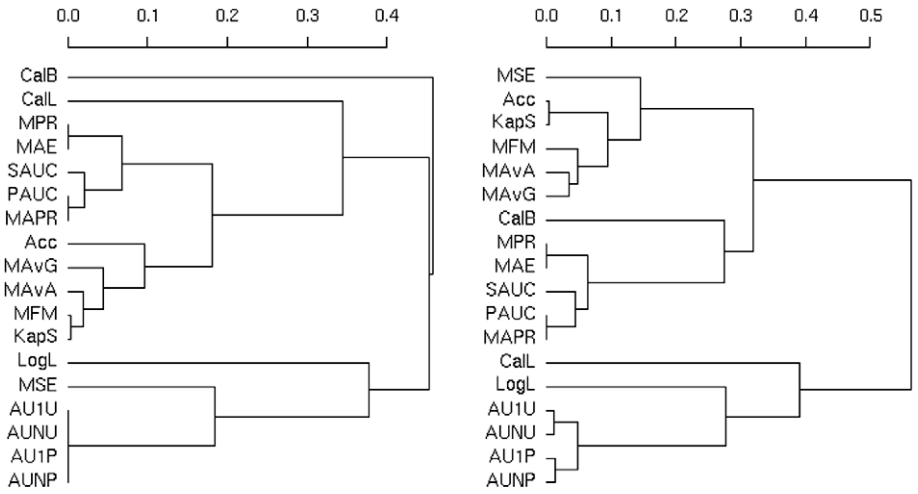

Fig. 2. Dendrograms of rank correlations between the metrics for two-class datasets (left) and multi-class datasets (right).

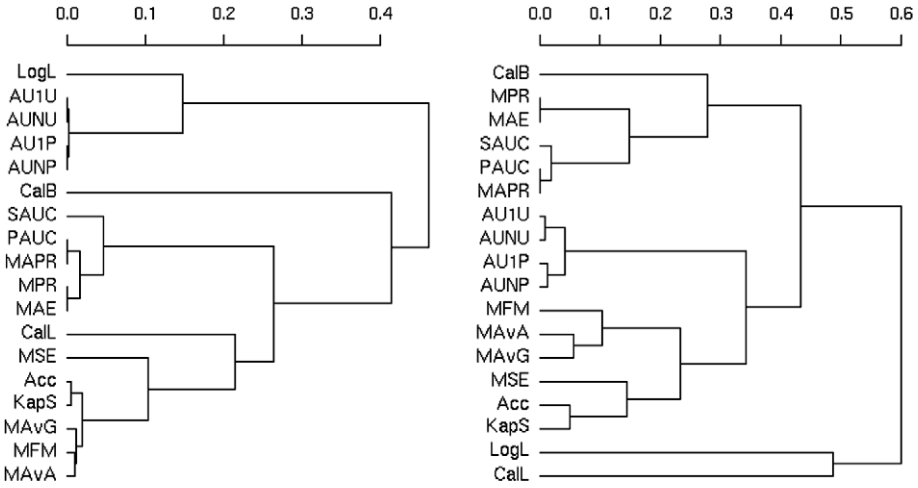

Fig. 3. Dendrogram of rank correlations between the metrics for balanced datasets (left) and imbalanced datasets (right).

from less data, when it comes to large datasets, AUC and LogL are very clearly separated from the rest of the measures.

6. Sensitivity analysis

The previous analysis disentangles the relationships of the performance measures and clusters them in groups, according to their correlation. From the definitions, the correlations and the clusters, we have been able to give an interpretation of these groups and their relation to the three families of measures: those based on error, those based on ranking and those based on probabilities. Additionally, we have also analysed the influence of imbalance to these measures. However, for those cases where we have a "Yes" on one trait we do not know the degree of sensitivity to that trait. The following experiments try to shed more light on this. Furthermore, some measures integrate features from more than one of the previous families and behave differently to changes on prior probability distribution (especially, class frequencies). To gain more insight into the relationship of these measures and their ability to capture misclassifications, bad rankings (separability), bad probabilities, and class proportion drifts, we have devised some experiments to directly assess these issues, to complement the theoretical arrangement performed in Table 1.

We present four experiments on model selection over four synthetic scenarios. The idea is that, given two models M_1 and M_2, with M_1 being better than M_2, we progressively introduce some noise to both models to check whether the performance measures are able to choose M_1. In order to analyse the four traits mentioned above, noise is applied to both models in four different ways.

- *Misclassification noise:* Noise is applied to actual classes. In this scenario we measure how sensitive the measure is to changes in the actual class produced.
- *Probability noise:* Noise is applied to the probabilities the models produce for each prediction. In this scenario, we measure how sensitive the measure is to changes on model probabilities. This can be interpreted as a situation where we analyse the reliability of the estimated probabilities, good calibration, etc.
- *Ranking noise:* Noise is applied to the ranking of the model prediction. In this scenario we measure how sensitive the measure is to model ranking change. This can be interpreted as situations where we analyse the reliability of the order of scores, i.e. good or bad class separability, etc.
- *Class proportion noise:* Noise is applied to the frequency of the dataset classes, i.e. we vary the proportion of classes. In this scenario we measure how sensitive the measure is to class

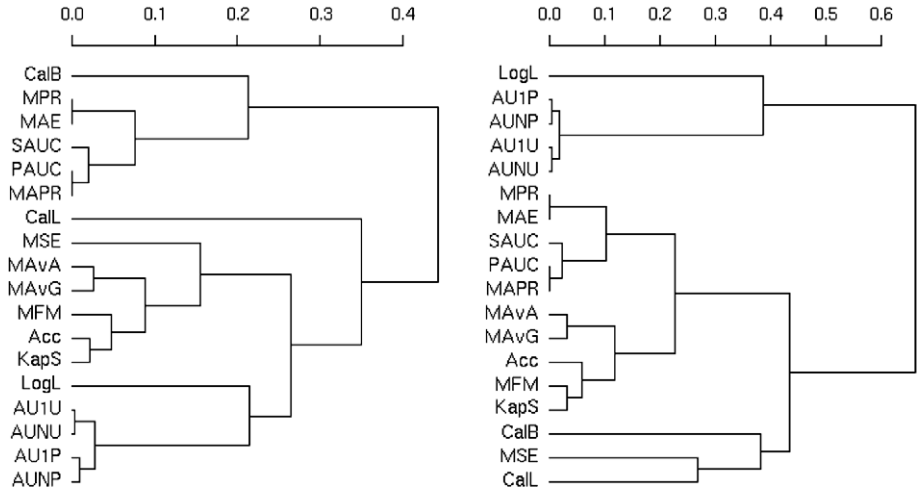

Fig. 4. Dendrogram of rank correlations between the metrics for small datasets (left) and large datasets (right).

proportion drifts. This can be interpreted as situations where we analyse the robustness (or conversely, sensitivity) to changes in prior class distribution.

The first part for the four experiments is the same. Two binary (two-class) models are created. These artificial classifiers are randomly created in the following way. First, we generate 100 real numbers from a uniform distribution on [0,1]. These numbers represent probabilities of the positive class. We order these probabilities decreasingly. Secondly, we assign a positive class to the elements where the probability is greater than 0.5 and negative class to the rest. Finally, we randomly modify 10 probabilities using the same uniform distribution on [0,1]. As a result, we have a classifier, denoted by M_1, with a good separability/ranking and a performance level of about 95%. This is just a way to generate such an artificial classifier, and does not have a significant influence on the following experiments if done differently (provided that the classifier has good separability and classification rate).

The second classifier, M_2, is obtained from the first one by randomly modifying 10 additional probabilities (different from the ones which were modified on the first classifier and copied to the second). Consequently, on average, the second classifier has worse separability and a worse classification rate (about 90%) than the first one.

A remarkable issue of this setting is that, as a result, M_2 is better calibrated initially than M_1. We will see this through the *CalB* measure, which will be ignored during the analysis, due to its 'strange' behaviour, since it is not a performance metric useful for model selection, but is purely a measure of calibration.

We generate two classifiers in the previous way 10.000 times (so giving 10.000 experiments) for each level of noise. For each experiment we record the selection made by the measure. If it selects M_1, we score 1, if it select M_2, we score 0. For ties, we score 0.5. When no noise is introduced, since M_1 is generally better than M_2, all performance measures should select M_1 on average. Things change, though, if we introduce some type of noise to M_1 and M_2.

The following experiments gradually introduce different types of noise, in order to check the degree to which the selection made by each measure is affected.

6.1. Misclassification noise

In this first scenario, we apply noise to actual class in order to measure the sensitivity (or conversely, robustness) to changes (or noise) in the actual class. Noise ranges from 0% (where no class label is modified on the dataset) to 100% (where all class labels are randomly generated). At 100% noise, all labels are new and both models should behave similarly. In the rest, since M_1 must be better than M_2, we record the estimated probability (i.e. frequency) of making a wrong guess (choosing M_2 instead).

As we can see in Fig. 5, measures go from 0.0 (no wrong guesses) to 0.5 (half the chance of a wrong guess). The point at 0% and 100% noise shows a high coincidence for all measures. The interesting part of the plot is precisely the evolution from 0% noise to 100%. We can see four lines where the measures cluster. Setting CalB appart, the first cluster on the bottom (with an average around 0.29 mistakes) is more robust to this misclassification noise and is logically composed on measures based on misclassification or error: Acc, MFM, MAvA, MAvG and KapS. The rest of the measures are at significant distance (with an average around 0.31): AUC measures, SAUC, PAUC, MSE, MPR, MAPR, MAE, CalL. The least robust is LogL with an average of 0.33. This means that should the dataset have noise on the class labels (i.e. noise in the test set), these latter measures will not behave well when choosing the best

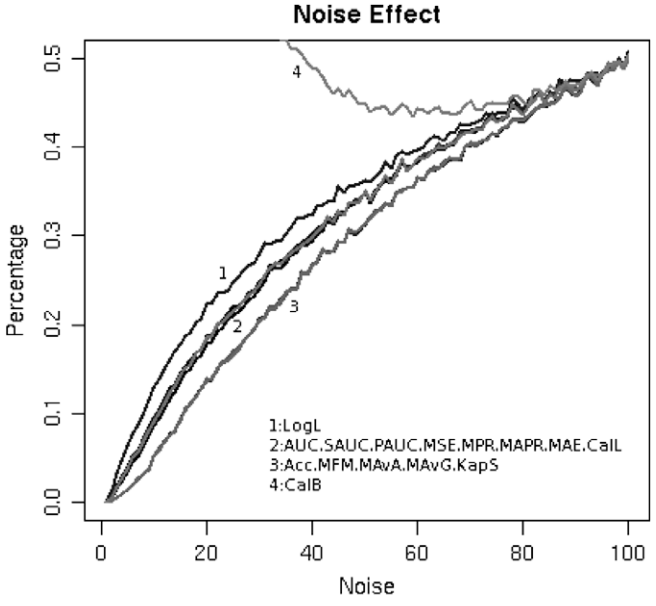

Fig. 5. Measure sensitivity to misclassification noise.

model. The two main clusters are clearly consistent with the first column ("Class Threshold") in Table 1.

6.2. Probability noise

In this second scenario, we apply noise to the probabilities of the models in order to measure the sensitivity (or conversely, robustness) to bad probability estimation. Class labels on the datasets are left unaltered, but classification thresholds might vary.

Here, *all* the probabilities are modified at all the degrees of noise. The level of noise determines the degree to which probabilities are modified. A value α (α obtained randomly from a uniform distribution $[-\beta..\beta)$ is added to each probability. β goes from 0 (noise = 0) to 0.5 (noise = 100).

As we can see in Fig. 6, the results are very different from those in the previous scenario. Here, qualitative measures such as Acc, MAvA, MAvG, KapS and MFM make bad choices in general. The groups begin at the bottom with the four AUC measures, which along with MSE present an average wrong choice ratio between 0.087 and 0.088. The explanation of why AUC measures behave well is simple. The ranking is only affected significantly at great degrees of probability noise. Consequently, the good choice is preserved. The behaviour of MSE is more difficult to explain, since MSE behaves quadratically, but it can be understood if we see that at the right of the picture, probabilities are modified +/−0.25 on average, which means that great changes, for which the quadratic stress will make a big impact, are not common.

At quite a distance (more or less in the middle band of the graph) we find MPR, MAPR, MAE and PAUC (with averages about 0.13), which are all probabilistic measures equally sensitive to small or large probability changes. SAUC is, surprisingly, found next with 0.15. SAUC might be found here for different reasons, because the rankings in the middle (which have more weight due to the inclusion of probabilities in this measure) have more relevance in this measure. At the upper band, we find the qualitative measures (with an average value or 0.18): Acc, MAvA, MAvG, KapS and MFM. The explanation for this bad behaviour is that for qualitative measures, if all probabilities are modified, as is the case here, the border between classes is highly affected, and this makes these measures worse for selecting the good model. Finally, CalL and LogL present inverse behaviours and are more erratic then the rest. CalL starts badly but it is the best measure in the end. This means that when distortion on probabilities is high, CalL is still able to tell between models. LogL on the contrary, behaves reasonably well for small probability distortions but its logarithmic character makes it the worst measure for high distortions (since many probabilities will be cut to 0 or 1, yielding $-\infty$ and 0 logarithms).

With this experiment, we have a view which gives more information than that seen in column ("Calibration") in Table 1, especially for probabilistic measures.

6.3. Ranking noise

The third scenario introduces noise to the ranking. In particular, given a model where the probabilities are sorted, we introduce random swappings of two consecutive elements. The degree of noise goes from no swapping (noise = 0) to 100.000 swappings (noise = 80).

Class labels on the datasets remain unaltered, and classification thresholds are constant at 0.5. However, calibration, misclassification and separability are affected by this type of noise.

The results shown on Fig. 7 indicate three types of measures. On the one hand (excluding CalB), at the top, we have the measures based on misclassification or error: Acc, MFM, MAvA, MAvG and KapS. These are more affected by these swappings than the rest since chained swappings are more frequent in probabilities around 0.5, which implies a change of class. On the other hand, we find the probabilistic and ranking measures such as SAUC, PAUC, MSE, MPR, MAPR, MAE, CalL and the AUC measures. These clusters are clearly consistent with the third column ("Ranking") in Table 1.

6.4. Class frequency variation

The last experiment evaluates what happens if one of the classes has few examples and how this affects the robustness of a comparison between models. From the original classes 0 and 1, we progressively eliminate examples from class 1. At a noise level of 0 we eliminate no elements from class 1. At a noise level of 50 we eliminate all the elements (50 elements) from class 1.

The results in Fig. 8 show that the AUC-based measures (the four AUCs, and SAUC, PPAUC/MAPR) are the ones which behave worst, because of their 1-vs-1 or 1-vs-*n* definition. We find that

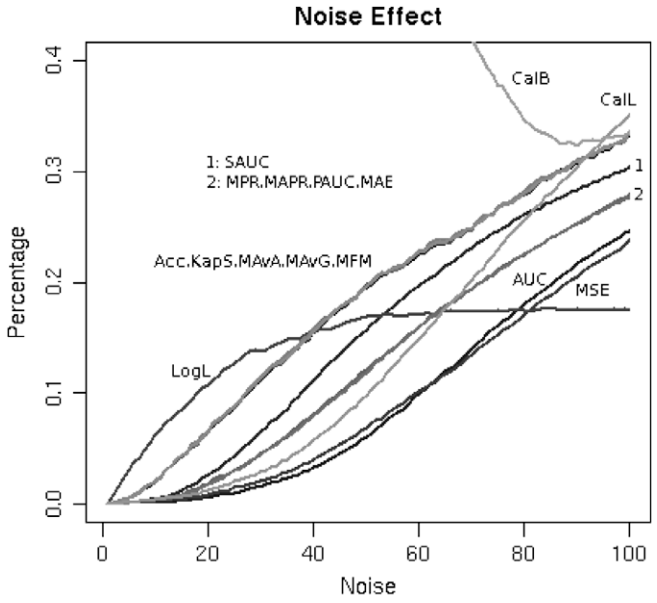

Fig. 6. Measure sensitivity to probability noise.

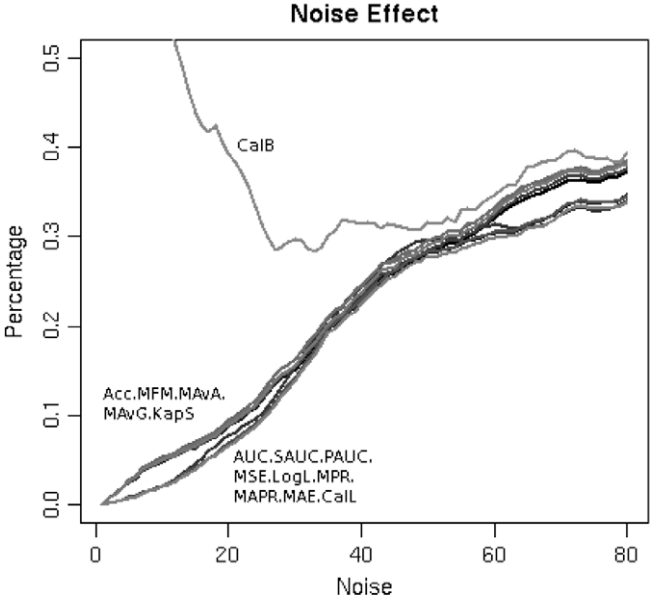

Fig. 7. Measure sensitivity to ranking noise.

SAUC and PAUC/MAPR are especially bad. At a distance, we find MAvA and MAvG which have a non monotonic behaviour. At a short distance MPR/MAE. LogL, CalL and MSE are quite robust. Then, MFM and KapS, which beahve very well until value 30. Acc is very robust and remains almost constant. Even with noise close to 50 where about only one element is of class 1, the selection error is just 0.034.

The interpretation is straightforward in this case. If we want to measure the quality of models that might be affected by classes with a very low percentage of elements, the AUC-based measures and the MPR, MAE and macro-averages are not a good idea, because the global measure is heavily influenced by a poor assessment of an infrequent class. This is precisely because these measures give equal value to all classes independently of their frequency. On the other hand, Acc, MFM, KapS, LogL, CalL and MSE give a relevance to each class which is proportional to its frequency. In this sense a badly assessed class is not a problem. This is consistent with the fourth column ("Class Frequency") on Table 1, and the dendrograms for imbalanced datasets shown in the previous section.

6.5. Discussion

From the previous four scenarios, we can say that accuracy and other qualitative measures are the best when noise is present on the dataset (the first experiment). Consequently, models evaluated with qualitative measures will be more robust when concept drift or other strong changes appear in the evidence. Probability-based measures are not good here and AUC measures behave relatively well. However, qualitative measures are very bad when distortion is produced during learning because a bad algorithm is used or small training datasets (models are distorted, which are reproduced by the second and the third experiment). In these cases, AUC measures are the best. According to this, if we have a learning scenario where distortion might happen on the datasets or on the learning process, the AUC measure is preferable, as has been shown in many previous studies (e.g. Rosset, 2004). Finally, if we have very few examples from a class, measures which are based on macro-averages or 1-vs-1 or 1-vs-*n* combinations are a bad option, because they will be highly influenced by a bad estimation of the error for the minority class.

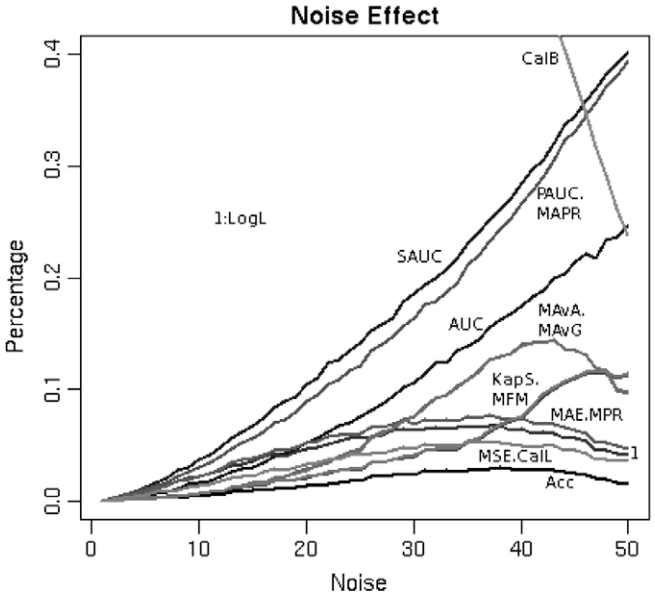

Fig. 8. Measure sensitivity to class frequency noise.

7. Conclusions

We have studied the relationships between the most common performance measures for classifiers. In this study, we have started from the definitions, then we have designed a set of experiments to analyse the correlations between measures and their sensitivity to several identified traits. The results uncover the existence of important similarities between measures but also significant differences between others.

The previous analysis shows that most of the measures used in machine learning and pattern recognition for evaluating classifiers really measure different things, especially for multiclass problems and problems with imbalanced class distribution, where correlations are worse. One of the most surprising results from the study is that the correlations between metrics inside the same family are not very high, showing that with a probabilistic understanding of error, it is very different to use MSE, LogL or MPR. It is even more different for the calibration measures. With a qualitative understanding of error, it is still different to use Acc or MAvG, although correlations in this group are higher. The only compact group happens when we want to rank predictions, and it is not significantly different to use different variants of AUC. Consequently, the previous analyses in pattern recognition or machine learning (stating, e.g., that one method is better than other) using different metrics (even inside the same family, except AUC measures) could not be comparable and extensible to the other metrics, since, the differences in performance between modern machine learning methods are usually tight.

As future work, one interesting issue would be to analyse the relationship between measures when one is used with a small sample and the other is used with a large sample from the same distribution. This would complete our sensitivity study on which measure captures more information and is more robust for small datasets. Another line of future research would be the development of new measures (not as an average of measures as other works have done (Caruana and Niculescu-Mizil, 2004; Huang and Ling, 2007), but in the way suggested at the end of section 3.1), or the inclusion of more measures in the study, such as the chi-square statistic (Palocsay et al., 2001), or the Critical Success Index (CSI) and Heidke's Skill Statistic (HSS) (Marzban, 1998; Marzban and Haupt, 2005).

Summing up, apart from the clarification and the many observations found in the relationship between metrics and their sensitivity to several characteristics, this work can be used as a reference when comparing two different experimental works in literature which use different metrics, in order to see whether the results are comparable or not.

Acknowledgements

The authors would like to thank David L. Dowe for some comments and discussions on LogLoss and other probabilistic measures. We also thank the anonymous reviewers for their corrections and helpful comments. This work has been partially supported by the EU (FEDER) and the Spanish MEC under Grant TIN 2007-68093-C02-02, Generalitat Valenciana under Grant GV06/301, UPV under Grant TAMAT and the Spanish project "Agreement Technologies" (Consolider Ingenio CSD2007-00022).

References

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison Wesley.
Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford University Press.
Blake, C.L., Merz, C.J., 1998. UCI Repository of Machine Learning Datbases.
Brier, G.W., 1950. Verification of forecasts expressed in terms of probabilities. Mon. Weather Rev. 78, 1–3.

Buja, A., Stuetzle, W., Shen, Y., 2005. Loss functions for binary class probability estimation: Structure and applications. <http://www-stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf>.

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

Cohen, J., 1960. A coefficient of agreement for nominal scales. J. Educ. Psychol. Meas. 20, 37–46.

Cortes, C., Mohri, M., 2003. AUC optimization vs. error rate minimization. In: Advances in Neural Information Processing Systems 16, NIPS 2003.

Davis, J., Goadrich, M., 2006. The relationship between precision-recall and ROC curves. In: ICML'06: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240.

DeGroot, M., Fienberg, S., 1982. The comparison and evaluation of forecasters. Statistician 31 (1), 12–22.

Domingos, P., Provost, F., 2003. Tree induction for probability-based ranking. Mach. Learn. 52, 199–216.

Dowe, D.L., Farr, G.E., Hurst, A.J., Lentin, K.L., 1996. Information-theoretic football tipping. In: Third Conference on Maths and Computers in Sport, vol. 14, pp. 233–241.

Esposito, F., Malerba, D., Semeraro, G., 1997. A comparative analysis of methods for pruning decision trees. IEEE Trans. Pattern Anal. Machine Intell. (TPAMI) 19 (5), 476–491.

Fawcett, T., 2001. Using rule sets to maximize ROC performance. In: 2001 IEEE International Conference on Data Mining (ICDM-01).

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recogn. Lett. 27 (8), 861–874.

Fawcett, T., Niculescu-Mizil, A., 2007. PAV and the ROC convex hull. Mach. Learn. 68 (1), 97–106.

Ferri, C., Flach, P.A., Hernández-Orallo, J., 2003. Improving the AUC of probabilistic estimation trees. In: Machine Learning: ECML 2003, 14th European Conference on Machine Learning, Proceedings, Lecture Notes in Computer Science, Springer, pp. 121–132.

Ferri, C., Flach, P.A., Hernández-Orallo, J., Senad, A. 2004. Modifying ROC curves to incorporate predicted probabilities. In: Second Workshop on ROC Analysis in ML.

Ferri, C., Hernández, J., Modroiu, R., 2004. An experimental comparison of classification performance metrics. In: Proceedings of The Learning 2004 Conference, Universidad Miguel Hernández, Elche, Spain, pp. 39–44.

Flach, P.A., 2003. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In: Machine Learning, Proceedings of the 20th International Conference (ICML 2003), pp. 194–201.

Flach, P.A., Takashi Matsubara, E., 2007. A simple lexicographic ranker and probability estimator. In: The 18th European Conference on Machine Learning, Springer, pp. 575–582.

Flach, P.A., Blockeel, H., Ferri, C., Hernandez-Orallo, J., Struyf, J., 2003. Decision Support for Data Mining: Introduction to ROC Analysis and its Application. Kluwer Academic Publishers. pp. 81–90.

Freund, Y., Iyer, R.D., Schapire, R.E., Singer, Y., 2003. An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. 4, 933–969.

Fuernkranz, J., Flach, P.A., 2005. ROC'n' rule learning – towards a better understanding of covering algorithms. Mach. Learn. 58 (1), 39–77.

Good, I.J., 1952. Rational decisions. J. Roy. Statist. Soc. Ser. B 14, 107–114.

Good, I.J., 1968. Corroboration, explanation, evolving probability, simplicity, and a sharpened razor. British J. Philos. Sci. 19, 123–143.

Hand, D.J., Till, R.J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. 45 (2), 171–186.

Harrell Jr., F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models. Logistics Regression, and Survival Analysis. Springer.

Huang, J., Ling, C.X., 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans. Knowl. Data Eng. (TKDE) 17 (3), 299–310.

Huang, J., Ling, C.X., 2007. Constructing new and better evaluation measures for machine learning. In: Manuela M., Veloso, (Eds.), IJCAI, pp 859–864.

Huang, J., Lu, J., Ling, C.X., 2003. Comparing naive Bayes, decision trees, and svm with auc and accuracy. In: ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society, p. 553.

Kuncheva, L.I., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., 2003. Limits on the majority vote accuracy in classifier fusion. Pattern Anal. Appl. 6 (1), 22–31.

Lebanon, G., Lafferty, J.D., 2002. Cranking: Combining rankings using conditional probability models on permutations. In: Machine Learning, Proceedings of the 19th International Conference (ICML 2002), pp. 363–370.

Marzban, C., 1998. Bayesian probability and scalar performance measures in gaussian models. J. Appl. Meteorol. 5, 72–82.

Marzban, C., Haupt, S.E., 2005. On genetic algorithms and discrete performance measures. In: Proceedings of the Fourth Conference on Artificial Intelligence Applications to Environmental Science.

Melnik, O., Vardi, Y., Zhang, C., 2004. Mixed group ranks: Preference and confidence in classifier combination. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 26 (8), 973–981.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill.

Murphy, A.H., 1972. Scalar and vector partitions of the probability score: Part II. *n*-state situation. J. Appl. Meteorol. 11, 1182–1192.

Narasimhamurthy, A., 2005. Theoretical bounds of majority voting performance for a binary classification problem. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 27 (12), 1988–1995.

Palocsay, S.W., Stevens, S.P., Brookshire, R.G., 2001. An empirical evaluation of probability estimation with neural networks. Neural Comput. Appl. 10 (1), 48–55.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Rosset, S., 2004. Model selection via the AUC. In: Machine Learning, Proceedings of the 21st International Conference (ICML 2004).

Sanders, F., 1963. On subjective probability forecasting. J. Appl. Meteorol. 2 (191–201).

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Elsevier.

Wu, S., Flach, P.A., Ferri, C., 2007. An improved model selection heuristic for auc. In: 18th European Conference on Machine Learning, pp. 478–489.

Young Sohn, S., 1999. Meta analysis of classification algorithms for pattern recognition. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) 21 (11), 1137–1144.

Zadrozny, B., Elkan, C., 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In: Proceedings of the 18th International Conference on Machine Learning (ICML 2001), pp. 609–616.