A comprehensive survey of error measures for evaluating binary decision making in data science

Frank Emmert-Streib1, Salisou Moutari2, Matthias Dehmer3,4,5

1Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland
2Mathematical Sciences Research Centre, School of Mathematics and Physics, Queen's University Belfast, Belfast, UK
3Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, Steyr Campus, Austria
4Department for Biomedical Computer Science and Mechatronics, UMIT—The Health and Life Sciences University, Hall in Tyrol, Austria
5College of Computer and Control Engineering, Nankai University, Tianjin, P. R. China

Correspondence
Frank Emmert-Streib, Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, Tampere, Finland. Email: frank.emmert-streib@tut.fi

Funding information
Austrian Science Funds, Grant/Award Number: P30031

Binary decision making is a topic of great interest for many fields, including biomedical science, economics, management, politics, medicine, natural science and social science, and much effort has been spent for developing novel computational methods to address problems arising in the aforementioned fields. However, in order to evaluate the effectiveness of any prediction method for binary decision making, the choice of the most appropriate error measures is of paramount importance. Due to the variety of error measures available, the evaluation process of binary decision making can be a complex task. The main objective of this study is to provide a comprehensive survey of error measures for evaluating the outcome of binary decision making applicable to many data-driven fields.

This article is categorized under:
Fundamental Concepts of Data and Knowledge > Key Design Issues in Data Mining
Technologies > Prediction
Algorithmic Development > Statistics

KEYWORDS
classification, data science, decision making, error measures, machine learning, statistics

1 | INTRODUCTION

An important part of any scientific field is to make decisions based on available information. In its most prevalent form, this decision making is binary, that is, one selects between two options. Examples of such problems include the decision of a medical doctor for or against the surgery of a patient, the decision of a stock broker to buy or not to buy a security or the decision of an investor to provide venture capital or not to a company.

In this study, we provide a comprehensive survey of error measures for evaluating the outcome of binary decision making problems. These measures can be applied to any problem that can be cast as a classification problem and are therefore of great interest for many fields, including biomedical science, economics, information science, management, politics, medicine, natural science, social science, and web science. We will show that there is a variety of error measures available, making the evaluation process of the outcome of binary decision making a complex task.

Despite the importance of the problem, there are surprisingly few review or survey articles devoted to a comprehensive overview of this topic. Examples for such contributions can be found in Ferri, Hernández-Orallo, and Modroiu (2009), Fiedling and Bell (1997), Han, Pei, and Kamber (2011), Parker (2013) and Sokolova and Lapalme (2009). Importantly, their focus is different in various aspects including the accessibility and completeness compared to our study. For instance, in Sokolova and Lapalme (2009) only little discussion on the error measures is presented and important measures like the false discovery
rate or the normalized mutual information are omitted. There are also review articles with a particular domain of application in mind, see e.g., Baldi, Brunak, Chauvin, Andersen, & Nielsen, 2000, focusing on the secondary structure of proteins and the alignment of amino acid sequences. Unfortunately, such presentations make them difficult to follow for scientists having an interest outside such a particular application domain because the presented discussions use a problem-specific jargon that is cumbersome to convert into generic formulations. As a consequence, the benefit of such presentations is frequently tightly bound to narrow application fields. Furthermore, there are also advanced reviews available (Powers, 2011), that aim at the expert level providing many technical details.

For all of these reasons, in our survey paper, we are aiming at an accessible level of description for general purpose applications relevant for any data-driven field. This is especially important in the emerging era of data science (Emmert-Streib, Moutari, & Dehmer, 2016; Hardin et al., 2015) because this multidisciplinary field does not aim at a single application domain only but provides systematic approaches for any data generating field.

2 | UNDERLYING METHODS OF DECISION MAKING

Before we start with surveying the error measures, we want to set the context by briefly discussing the methodology used for binary decision making.

The most common method used for binary decision making is a two-class classifier. Prominent examples for such methods are support vector machines, decision trees, random forests, linear discriminant analysis, and neural networks (Breiman, 2001; Clarke, Fokoue, & Zhang, 2009; Schölkopf & Smola, 2002). Another method that makes binary decisions is a statistical hypothesis test (Lehman, 2005). In this case, one class corresponds to the “rejection of the null hypothesis” and the other class to the “acceptance of the null hypothesis.” A third type of method is more implicit or derived corresponding, for example, to a survival analysis comparing two patient groups (Kleinbaum & Klein, 2005). Such methods can be considered derived because, as in the case of the survival analysis, the decision is again made based on statistical hypothesis testing. For comparing survival curves one can use, for example, the Log-rank or Mantel–Haenszel test.

It is interesting to note that the following error measures can not only be applied to the results of computational methods but also experimental ones. For instance, a patient being sick is consulting a medical doctor and a diagnostic test is conducted. The test is performed experimentally by, for example, analyzing a blood sample in a laboratory, but the result is binary indicating for instance the presence or absence of a disease. Other examples can be found in sociology where survey questions may be assessed in a similar way.

Overall, this means there are a large variety of different methods, computational and experimental ones, that can be used for obtaining binary decisions. However, regardless of the nature of the utilized method the following error measures can be applied for the evaluation of the outcome.

3 | SUMMARIZING THE OUTCOME OF DECISION MAKING

For our following considerations we need a common ground that connects all the different methods discussed in the previous section. This can be achieved by utilizing a contingency table also called a confusion matrix.

A contingency table allows a convenient summarization of binary decision making. In the following, we assume we have two classes called +1 and −1. Here the indicators of the two classes are labels or nominal numbers (also called categorical numbers). The outcome of the decision making can be one of the following four cases:

1. The actual outcome is class +1 and we predict a +1.
2. The actual outcome is class +1 and we predict a −1.
3. The actual outcome is class −1 and we predict a +1.
4. The actual outcome is class −1 and we predict a −1.

It is convenient to give these four cases four different names. We call them as:

1. True positive: TP
2. False negative: FN
3. False positive: FP
4. True negative: TN
The purpose of a contingency table is to summarize these results, as shown in Figure 1. It is easy to see that these four quantities characterize the outcome of a binary decision making completely. For this reason we are calling them the four fundamental error measures. All error measures we will discuss in the following sections will be based on these four fundamental errors. In order to facilitate the understanding, we will utilize the contingency table whenever beneficial for explaining different error measures.

We would like to note that we use the term “fundamental error measures” to emphasize the importance of these four measures from all other error measures discussed later. Mathematically, we will see that all following measures are functions of the four fundamental error measures. Hence, for these measures the TP, FP, TN and FN are independent variables.

4 | OVERVIEW OF ERROR MEASURES

In Figure 2, we show an overview of many different error measures that can be used for evaluating binary decision making. There are three broad groups of errors. The first group focuses on correct outcome, the second on incorrect outcome and the third on both. In the following section we will discuss each of these measures.

**Fundamental error measures:**
- TP = True positive: correct positive prediction
- FP = False positive (type I error): incorrect positive prediction
- TN = True negative: correct negative prediction
- FN = False negative (type II error): incorrect negative prediction

**Summarizations:**
- P = TP + FN
- N = FP + TN
- R = TP + FP
- A = FN + TN
- T = TP + FP + FN + TN
- PR = prevalence = $\frac{TP}{T} = \frac{TP + FN}{TP + FN + TN + FP} \in [0, 1]$

**Error measures:**
- TPR = True positive rate = sensitivity = recall = $\frac{TP}{P} = \frac{TP}{TP + FN} \in [0, 1]$
- TNR = True negative rate = specificity = $\frac{TN}{N} = \frac{TN}{TN + FP} \in [0, 1]$
- PPV = positive predictive value = precision = $\frac{TP}{R} = \frac{TP}{TP + FP} \in [0, 1]$
- NPV = negative predictive value = $\frac{TN}{A} = \frac{TN}{TN + FN} \in [0, 1]$
- ACC = accuracy = $\frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FP + FN + TN} \in [0, 1]$
- $F_\beta = (1 + \beta^2) \frac{PPV \cdot \text{sensitivity}}{\beta^2 \cdot PPV + \text{sensitivity}}$
- $F_1 = 2 \cdot \frac{PPV \cdot \text{sensitivity}}{PPV + \text{sensitivity}} \in [0, 1]$
- FDR = false discovery rate = $\frac{FP}{R} = \frac{FP}{FP + TP} \in [0, 1]$
- FOR = false omission rate = $\frac{FN}{A} = \frac{FN}{FN + TN} \in [0, 1]$
- FNR = false negative rate = $\frac{FN}{P} = \frac{FN}{FN + TP} \in [0, 1]$
- FPR = false positive rate = $\frac{FP}{N} = \frac{FP}{FP + TN} \in [0, 1]$
- $E = \text{error rate} = \frac{TP + FN}{T} = \frac{TP + FN + FP + TN}{TP + FP + FN + TN} \in [0, 1]$
- MCC = Matthews correlation coefficient = $\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FN)(TN + FP)}} \in [-1, 1]$
- $\kappa = \text{Cohen’s kappa} = \frac{ACC - \text{rACC}}{1 - \text{rACC}} \in (-\infty, 1]$ with $\text{rACC} = \frac{PR + NA}{T^2}$
- NMI = normalized mutual information = $\frac{I(\text{actual, predicted})}{H(\text{actual})} = \frac{H'(\text{actual}) - H(\text{actual, predicted})}{H(\text{actual})} \in [0, 1]$

**Figure 2** Overview of different error measures for binary decision making
True positive rate and true negative rate

The true positive rate (TPR) and true negative rate (TNR) are defined by:

\[
TPR = \text{sensitivity} = \frac{TP}{P} = \frac{TP}{TP + FN} \in [0, 1],
\]

\[
TNR = \text{specificity} = \frac{TN}{N} = \frac{TN}{TN + FP} \in [0, 1].
\]

The definitions ensure that both measures are bound between 0 and 1. For an error-free classification we obtain \( FN = FP = 0 \) which implies \( TPR = TNP = 1 \). On the other hand, for \( TP = TN = 0 \) we obtain \( TPR = TNP = 0 \). In the literature, the TPR is also called sensitivity and the TNR is called specificity (Flach, 2012).

It is important to note that both quantities utilize only half of the information contained in the confusion matrix. The TPR uses only values from the first row and the TNR only values from the second row. In Figure 3 we highlight this by encircling the used fundamental errors. For simplicity, we refer to the first row as P-level and the second row as N-level. Hence, the TPR uses only values from the P-level and the TNR values from the N-level.

From Figure 3, it is clear that both measures are symmetric with respect to the utilized information; that is, just the roles of the classes are exchanged. This formulation enables to remember the measures more easily.

Positive predictive value and negative predictive value

The positive predictive value (PPV) and negative predictive value (NPV) are defined as:

\[
PPV = \text{precision} = \frac{TP}{R} = \frac{TP}{TP + FP} \in [0, 1],
\]

\[
NPV = \frac{TN}{A} = \frac{TN}{TN + FN} \in [0, 1].
\]

The definitions ensure that both measures are bound between 0 and 1. For an error-free classification, we obtain \( FN = FP = 0 \) which implies \( PPV = NPV = 1 \). On the other hand, for \( TP = TN = 0 \) we obtain \( PPV = NPV = 0 \). In the literature, the PPV is also called precision (Fawcett, 2006).

It is important to note that like TPR and TNR, the quantities PPV and NPV are estimated using only half of the information contained in the confusion matrix. The PPV uses only values from the first column and the NPV only values from the second column. In Figure 4 we highlight this by encircling the used fundamental errors. For simplicity, we refer to the first column as R-level and the second column as A-level. Hence, the PPV uses only values from the R-level and the NPV values from the A-level.

From Figure 3, it can be observed that both measures are symmetric with respect to the utilized information just the roles of the classes are exchanged. Such a formulation enables to remember the measures more easily.

Accuracy

The accuracy (A) is defined as:

\[
A = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N} = \frac{TP + TN}{R + A} \in [0, 1].
\]

The above definition ensures that the accuracy is bound between 0 and 1. For an error-free classification, we obtain \( FN = FP = 0 \) which implies \( A = 1 \). On the other hand, for \( TP = TN = 0 \) we obtain \( A = 0 \). Another terminology used to refer to the accuracy, in the context of clustering evaluation, is the Rand index. In contrast with the quantities TPR, TNR, PPV and NPV, the accuracy uses all values in the confusion matrix.
The general definition of the F-score is:

$$F_{\beta} = \left(1 + \beta^2 \right) \frac{\text{PPV} \times \text{sensitivity}}{\beta^2 \text{PPV} + \text{sensitivity}}.$$  \hspace{1cm} (6)

In this equation, the parameter $\beta$ can assume values in the interval $[0, \infty]$. The parameter $\beta$ enables to change the weighting between the PPV and the sensitivity. In Figure 5, we show an example for two different value pairs of PPV and sensitivity. It can be observed that for $\beta = 0$, the F-score corresponds to the PPV whereas for $\beta \to \infty$ it corresponds to the sensitivity. Intermediate values of $\beta$ enable to obtain "averaged" F-score values.

For $\beta = 1$, one obtains the $F_1$-score as:

$$F_1 = 2 \frac{\text{PPV} \times \text{sensitivity}}{\text{PPV} + \text{sensitivity}} \in [0, 1].$$ \hspace{1cm} (7)

The $F_1$-score is the harmonic mean of PPV and sensitivity, where the harmonic mean is defined as:

$$\frac{1}{F_1} = \frac{1}{n} \left( \frac{1}{\text{PPV}} + \frac{1}{\text{sensitivity}} \right).$$ \hspace{1cm} (8)

The $F_1$-score uses three of the four fundamental errors, namely TP, FP and FN.

### 4.5 False discovery rate and false omission rate

The false discovery rate (FDR) and false omission rate (FOR) are defined as:

$$\text{FDR} = \frac{\text{FP}}{\text{R}} = \frac{\text{FP}}{\text{FP} + \text{TP}} \in [0, 1],$$ \hspace{1cm} (9)

$$\text{FOR} = \frac{\text{FN}}{\text{A}} = \frac{\text{FN}}{\text{FN} + \text{TN}} \in [0, 1].$$ \hspace{1cm} (10)

The above definitions ensure that both measures are bound between 0 and 1. For an error-free classification, we obtain $\text{FN} = \text{FP} = 0$, which implies $\text{FDR} = \text{FOR} = 0$. On the other hand, for $\text{TP} = \text{TN} = 0$, we obtain $\text{FDR} = \text{FOR} = 1$. 
Both FDR and FOR quantities utilize only half of the information contained in the confusion matrix. The FDR uses only values from the first column and the FOR only values from the second column. In Figure 6, we highlight this by encircling the used fundamental errors. That means the FDR uses information from the R-level but in contrast to the PPV, which uses also information from this level, it focuses on failure by forming the quotient of FP and R. Similarly, the FOR uses only information from the A-level, forming the quotient of FN and A.

From Figure 6 it is clear that both measures are symmetric with respect to the utilized information and just the roles of the classes are exchanged.

Frequent application domains for the FDR and the FOR are biology, medicine, and genetics (Farcomeni, 2008; Genovese & Wasserman, 2006).

4.6 | False negative rate and false positive rate

The false negative rate (FNR) and false positive rate (FPR) are defined as:

\[
\text{FNR} = \frac{FN}{P} = \frac{FN}{FN + TP} \in [0, 1],
\]

\[
\text{FPR} = \frac{FP}{N} = \frac{FP}{FP + TN} \in [0, 1].
\]

From their definitions, both measures are bound between 0 and 1. For an error-free classification, we have FN = FP = 0, which implies FNR = FPR = 0. On the other hand, for TP = TN = 0 we obtain FNR = FPR = 1.

Furthermore, it is important to note that both quantities use only half of the information contained in the confusion matrix. The FNR uses only values from the first row whereas the FPR uses only values from the second row. In Figure 7, we highlight this by encircling the used fundamental errors.

From Figure 7, it can be again observed that both measures are symmetric with respect to the utilized information just the roles of the classes are exchanged.

4.7 | Matthews correlation coefficient

A common issue when applying machine learning (ML) techniques to a real-world problem is dealing with an imbalanced target variable. When applying ML methods to unbalanced data, the Matthews correlation coefficient (MCC) turned out to be useful (Matthews, 1975). It has been used to evaluate the performance of typical two-class classification problems (Baldi et al., 2000). The MCC metric has been introduced first by Matthews to assess the performance of protein secondary structure prediction (Matthews, 1975):

\[
\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \in [-1, 1]
\]

MCC ranges from −1 to +1. A value of −1 indicates that the prediction is completely wrong; a value of +1 indicates perfect prediction. MCC = 0 means that we consider a random classification where the model predictions have no detectable correlation to the true results. In Figure 8, we present some numerical results on the behavior of the MCC. For the shown simulations we assumed a fixed prevalence of 0.1 and a fixed sensitivity of 0.5 (left side) and 0.9 (right). The values of the specificity have been varied from 0.0 to 1.0 and all fundamental errors have been derived from this, including the shown ACC, TPR, and TNR. The chosen value for the prevalence ensures a strong imbalance between both classes.

In Figure 8a, low specificity values correspond in general to poor classification results as indicated by very low ACC values and negative values of the MCC. With an increasing specificity, the TNR and ACC increase both with a similar slope. However, due to the imbalance of the classes and the way we defined our model keeping the TPR constant, the increasing values of ACC are misleading. Indeed, the MCC increases but not linearly and much slower than the ACC. Moreover, the
The highest achievable value for MCC is only 0.68 in contrast to 0.95 for ACC. In Figure 8b, the sensitivity is much higher and, hence, the difference between ACC and MCC becomes reduced but is still clearly noticeable.

Due to the fact that an imbalance in the categories is a frequent problem encountered in analyzing data, the MCC is used throughout all application domains of data science.

### 4.8 Cohen's kappa

The next measure we will discuss is called Cohen's kappa, $\kappa$ (Cohen, 1960). It has been essentially used as a measure to assess how well the classifier performed as compared to how well it would have performed simply by chance. That means a model has a high kappa value if there is a big difference between the accuracy and the null error rate. Formally expressed, Cohen's kappa has been defined by Cohen (1960) as:

$$\kappa = \frac{ACC - rACC}{1 - rACC} \in (-\infty, 1].$$  \hfill (14)

Here, ACC is the accuracy and $rACC$ denotes the randomized accuracy defined as:

$$rACC = \frac{PR + NA}{T^2}. \hfill (15)$$

Usually, Cohen's kappa has been used as a measure of agreement between two categorical variables in various ML applications across several disciplines, including epidemiology, medicine, psychology and sociology (see e.g., Saczynski et al., 2012; Umesh, Peterson, & Sauber, 1989).

### 4.9 Normalized mutual information

The normalized mutual information is an information-theoretic measure, which has been used extensively in various applications (Principe, Xu, Zhao, & Fisher, 2000; Rost & Sander, 1993; Shridhar, Bartlett, & Seagrave, 1998). In most of these studies the normalized mutual information was applied to classification tasks and other learning problems.

When reviewing and assessing the evaluation criteria for learning problems, Baldi et al. (2000) defined the normalized mutual information as follows:
\[
\text{NMI} = \frac{TP}{T} \log \frac{TP}{T} \frac{TN}{T} \log \frac{TN}{T} \frac{FP}{T} \log \frac{FP}{T} \frac{FN}{T} \log \frac{FN}{T},
\]
\[
- \frac{TP}{T} \log \left( \frac{TP + FP}{T} \right) - \frac{FN}{T} \log \left( \frac{TP + FN}{T} \right),
\]
\[
- \frac{FP}{T} \log \left( \frac{TP + TN + FP}{T} \right) - \frac{TN}{T} \log \left( \frac{TN + FN + FP}{T} \right).
\]

In fact, this expression can be simplified (Baldi et al., 2000) to:
\[
\text{NMI} = \frac{TP}{T} \log \frac{TP}{T} \frac{TN}{T} \log \frac{TN}{T} \frac{FP}{T} \log \frac{FP}{T} \frac{FN}{T} \log \frac{FN}{T},
\]
\[
- \frac{TP}{T} \log \left( \frac{R P}{T} \right) - \frac{FN}{T} \log \left( \frac{P A}{T} \right),
\]
\[
- \frac{FP}{T} \log \left( \frac{R N}{T} \right) - \frac{TN}{T} \log \left( \frac{A N}{T} \right).
\]

Many variants of the normalized mutual information measure have been defined and applied (see e.g., Hu & Wang, 2008; Principe et al., 2000; Rost & Sander, 1993; Shridhar et al., 1998; Wallach, 2006). For instance, Hu and Wang (2008) define the normalized mutual information measures on the so-called augmented confusion matrix. This matrix has been defined by adding one column for a rejected class onto a conventional confusion matrix (Hu & Wang, 2008). Wallach (2006) considers normalized confusion matrices representing models and formulates the normalized mutual information measures on that matrix as well as other error measures such as Precision and Recall. By comparing the defined models, Wallach (2006) draws the conclusion that classical measures such as Precision and Recall may be misleading when evaluating classifier using only these measures.

4.10 Area under receiver operator characteristic curve

The final error measure we are presenting is the area under receiver operator characteristic (AUROC) curve (Bradley, 1997; Hanley & McNeil, 1982). In contrast to all previous measures discussed in this study, the AUROC curve can only be obtained via a construction process, rather than directly derived from the contingency table. The reason for this is that it does not make use of the optimal threshold of a classifier for deciding how to categorize data samples. Instead, this threshold can be derived, as we will show at the end of this section.

The first step of this construction process is to derive the ROC curve and the second is the integration of this curve to obtain the area under the curve. For constructing a ROC curve we need to obtain pairs of TPR and FPR (i.e., \(\text{TPR}_i, \text{FPR}_i\)). That means the ROC curve presents the TPR in dependence on the FPR. Due to the fact that the TPR is also the sensitivity and the FPR is also 1-specificity, an alternative representation would be the sensitivity in dependence on 1-specificity.

Let us assume that we have a dataset with \(n\) samples. Regardless of the method used, we can obtain either a score \(s_i\) or a probability \(p_i\) for every sample as an indicator of the membership for class +1 (and analogously values for class −1). In the following, we will discuss the construction based on probabilities but the discussion is similar using scores. Based on these probabilities, a decision is obtained by thresholding these values. That means for
\[
p_i > p_t,
\]
we decide to place sample \(i\) into class +1 and otherwise in class −1. By rank ordering all \(p_i\) for all samples in increasing order we obtain:
\[
p[1] \leq p[2] \leq \ldots \leq p[n].
\]

Now we apply successively all possible thresholds to obtain two groups, one group corresponding to class +1 and the other to class −1. This results in total in \(n + 1\) different thresholds and, hence, groupings:
\[
\lvert p[1] \leq p[2] \leq \ldots \leq p[n], \rvert
\]
\[
p[1] \leq \quad p[2] \leq \ldots \leq p[n],
\]
\[
\vdots
\]
For each of these groupings we can calculate the four fundamental errors and based on these, every error measure shown in Figure 2, including the TPR and FPR can also be calculated. Overall, this results in \( n + 1 \) pairs of \((\text{TPR}_i, \text{FPR}_i)\) for \( i \in \{1, \ldots, n + 1\} \) from which the ROC curve is constructed.

In Figure 9, we present some examples for ROC curves (in black) as a result from two logistic regression analyses. From Figure 9a,b, the smaller the AUROC value the closer is the ROC curve to the main diagonal (in red) corresponding to a random classification, which would result in \( \text{AUROC} = 0.5 \). On the other hand, the best possible classifier achieves \( \text{AUROC} = 1.0 \) which would correspond to a ROC curve shown in blue.

In addition to providing the AUROC values, ROC curves can also be used to determine the optimal threshold for the classifier itself since they were obtained without using this threshold, even if already determined in some way. However, in order to obtain such an optimal threshold we need to define an optimization function. In the literature, there are two frequent choices for this (Fawcett, 2006). The first is the distance from the ROC curve to the upper left corner, that is \((\text{TPR} = 1, \text{FPR} = 0)\) given by:

\[
\text{distance}_i = \sqrt{(1 - \text{TPR}_i)^2 + \text{FPR}_i^2}
\]

and the second is called the Youden-Index (Youden, 1950) given by:

\[
\text{Youden-Index}_i = \text{TPR}_i + (1 - \text{FPR}_i).
\]

The optimal threshold is then obtained by finding

\[
\text{id}_{\text{opt}} = \arg \min_i \{\text{distance}_i\}.
\]

**FIGURE 9** In the first row two examples for ROC curves are shown resulting from two logistic regression analyses. The second row shows results for Figure 9a for the distance to the ROC curve and the Youden-Index leading to optimal thresholds.
In general, these indices do not have to result in identical values, however, for the example shown in Figure 9a this is the case, as illustrated in Figure 9c. We would like to note that for reasons of visibility we scaled the displayed Youden-Index between 0 and 1.

After obtaining the optimal threshold one can derive any error measure, including the values of the TPR and FPR. For the ROC curve in Figure 9a, this leads to TPR = 0.797 and FNP = 0.266.

5 EVALUATING OUTCOME AS MULTIVARIATE PROBLEM

Now we are arriving at a point where we need to ask ourselves which measure to use when studying binary decision making problems? The general answer to this question is that there is no single measure that would be appropriate in all possible situations under all feasible conditions but, instead, one needs to realize that evaluating binary decision making should be done in a multivariate way. This means that, usually, we need to use more than one error measure to evaluate the outcome in order to avoid false interpretations of the prediction results.

In order to illustrate this aspect, we present an example. For this example, we simulate the values in the contingency table according to a model. Specifically, this means that we are defining an error model. To simplify the situation, we define this error model in the proportions of the four fundamental errors TP, FP, TN and FP, we are calling pTP, pFP, pTN and pFP. Each of these proportions can assume values between 0 and 1, and the four quantities sum up to 1, that is:

\[ p_{TP} + p_{FP} + p_{TN} + p_{FP} = 1. \]  

Furthermore, if we multiply each of these proportions with \( T \) (the total number of samples) we recover the four fundamental errors, that is:

\[ TP = T \cdot p_{TP}, \]  

Error model:

\[ p_{TP} = \{0.2, 0.21, ..., 0.6\} \]
\[ p_{FN} = \exp(-10 \cdot p_{TP}) \]
\[ p_{TN} = 1 - \exp(-8 \cdot p_{TP}) - 0.6 \]
\[ p_{FP} = 1 - (p_{TP} + p_{FN} + p_{TN}) \]

FIGURE 10 Visualization of the assumed error model that defines the proportions of the four fundamental errors
\[ TN = T \cdot pTN, \]  
\[ FP = T \cdot pFP, \]  
\[ FN = T \cdot pFN. \]  

In Figure 10, we depict a visualization of the assumed error model that defines the proportions of the four fundamental errors. It is easy to check that the sum of these four quantities equals always 1, for all possible values shown. The simulations start by assuming that \( pTP \) takes its values in the interval from 0.2 to 0.6 in step sizes of 0.01. Based on this, the values of \( pFN \) and \( pTN \) are determined by some exponential functions, as indicated in Figure 10. Finally, the value of \( pFP \) ensures just the conservation of the total probability, hence, no functional relation needs to be defined for this proportion. Overall, the model starts with a poor decision outcome corresponding to low \( pTP \) and \( pTN \) value and improves toward higher values corresponding to a better decision making. Furthermore, the model is unbalanced in the sizes of the classes as can be seen from the values of \( P \) and \( N \). It is also unbalanced in the predictions (see R and A).

In Figure 11, we present results for seven error measures that are obtained using the values from our error model. All of these pairwise dashboards show nonlinearities, at least to some extend. The most linear (but not exactly linear) correspondence is exhibited between the ACC and the \( F \)-score followed by the relationship between the FDR and the \( F \)-score. For instance, for \( pTP = 0.2 \) we obtain \( ACC = F = 0.39 \) indicating low classification performance but taking in addition the value of \( FDR = 0.7 \) into consideration we see that 70\% of all samples classified as class +1 are false. From this perspective, the classification results are even very poor. From the specificity versus sensitivity plot one can see that class +1 is much easier to recover than class −1 because there is a strong nonlinearity between these two error measures with a faster increase in the values of the sensitivity (TPR). This is reflected similarly in the PPV.

5.1 Comparing binary decision making methods

In the above examples we have shown that changes in the four fundamental errors (as defined by an error model) can lead to nonlinear effects in the dependent errors, for instance, in FDR. However, this type of issue is not the only problem in
evaluating binary decision making. Another problem is given by evaluating two instead of one binary decision making methods. In order to demonstrate this type of problem we show in Figure 12 the outcome of three binary decision making methods.

Specifically, in the first part of Figure 12 we show the proportion of the four fundamental errors for the three methods. Here we assumed that the application of a method to a dataset results in the shown errors and all three methods are applied to the same dataset. In order to demonstrate the problem we consider two scenarios. The first scenario corresponds to a comparison of the outcome of Method 1 with Method 2 and the second scenario to a comparison of the outcome of Method 1 with Method 3.

For Scenario 1, from Figure 12 we see that

\[
\begin{align*}
& p_{TP1} > p_{TP2} \\
& p_{TN1} > p_{TN2} \\
& p_{FN1} < p_{FN2} \\
& p_{FP1} < p_{FP2}
\end{align*}
\]

that means, the true predictions for Method 1 are always higher than for Method 2 and the false predictions for Method 1 are always lower than for Method 2. For this reason it seems obvious that Method 1 performs better than Method 2 regardless of what fundamental error measure is used and no matter if one considers just one of these or combinations. The result of the comparison is that Method 1 performs always better than Method 2.

However, for Scenario 2, shown in Figure 12 comparing Method 1 with Method 3, we see that

\[
\begin{align*}
& p_{TP1} > p_{TP3} \\
& p_{TN1} < p_{TN3} \\
& p_{FN1} = p_{FN3} \\
& p_{FP1} = p_{FP3}
\end{align*}
\]

First, we observe that the false predictions are identical. Second, the proportion of true positives is higher for Method 1 but the proportion of true negatives is higher for Method 2. Third, the absolute value of the distance of the true predictions is equal for the positive and negative classes, that is:

\[
\begin{align*}
\Delta p_{TP} &= p_{TP1} - p_{TP3} = 0.1 \\
\Delta p_{TN} &= p_{TN1} - p_{TN3} = -0.1
\end{align*}
\]

but the sign is different. Taking this information together one can see that without further information one cannot decide if Method 1 is better than Method 3 or vice versa.

This can be further clarified by looking at different error measures that are a function of the four fundamental errors. In Figure 12, we show results for \( T = 100 \). For Scenario 1, we see that all error measures that focus on positive outcome are

---

**FIGURE 12** Outcome of three binary decision making methods. The error measures are evaluated for \( T = 100 \) samples.
higher for Method 1 (in orange) than for Method 2 (in blue) and all error measures that focus on negative outcome are lower for Method 1 than for Method 2. In contrast, for Scenario 2, we see that there are some error measures that focus on positive outcome are higher for Method 1 (in orange) than for Method 3 (in brown) and some are lower for Method 1 (in orange) than for Method 3. For instance,

\[
\text{TPR}_1 > \text{TPR}_3, \quad (40) \\
\text{TNR}_1 < \text{TNR}_3. \quad (41)
\]

Similar results hold for error measures that focus on negative outcome. For instance,

\[
\text{FDR}_1 < \text{FDR}_3, \quad (42) \\
\text{FPR}_1 > \text{FPR}_3. \quad (43)
\]

Hence, the outcome of the comparison to decide which method is better depends on the used error measure. This example demonstrates the need to argue in favor of one or several error measures because this will influence the result of the comparison. However, in order to select one or several measures we need to add additional information beyond what is available in Figure 12 and this additional information needs to be domain-specific. However, domain-specific information results in domain-specific problems, for instance, in biology, medicine, finance or social sciences.

6 | OCCURRENCE OF ERROR MEASURES

Finally, we want to take a look into the literature to see what error measures are frequently used. In Figure 13, we show a comparison of the relative usage frequency of error measures. The underlying frequency numbers on which these relative frequencies are based on have been obtained from Google Scholar. For instance, for the PPV we find 644,000 different articles in which this term appears and for all terms combined we find 1,807,481 articles. This gives a relative usage frequency of the PPV of 35.6%. Overall, the PPV, NPV, FDR and FPR are dominating the literature.

From Figure 13, one can clearly see that not all error measures are used equally frequently but that there are tremendous differences in their usage. In order to obtain information about the application areas we repeat a similar analysis but now we evaluate the relative frequencies for application fields. For obtaining this information we use Scopus because Google Scholar does not provide information about the application fields.

In Figure 14, we show the results from this analysis for 11 error measures. For each error measure, we show the top five application fields. From these results it is also obvious that the application fields are not equally frequent but error measure specific. For instance, the FNR is most frequently used in medicine and to a much lesser extend in computer science. In contrast, for the F-score the roles are reversed. One can also see that medicine is for six out of nine error measures the field with the most frequent applications. Furthermore, medicine is among the top five application fields for all 11 error measures. A possible reason for this is that in medicine the conducted analyses need to be stringent due to the fact that patients may be harmed in case of erroneous analyses. For this reason, one cannot afford to neglect (patient) information by focusing on only a few error measures. Another reason is that patient-related analyses usually require approval by senior scientists. All these factors underline the need to consider as much information as possible captured by error measures. Furthermore, it also indicates that there is no simple answer to the question which error measure should be used because each error measure provides a different perspective about an underlying problem.
7 | CONCLUSION

In this study, we provided a comprehensive overview of many error measures that can be used for evaluating binary decision making in data science (Emmert-Streib et al., 2016). We were aiming at an accessible level of description and avoided application domain-specific formulations in order to make knowledge transfer easier for general purpose applications in fields such as biomedical science, economics, management, politics, medicine, natural science, psychology, or social science.

In summary, there are many error measures but none is adequate in isolation to capture effectively all different outcome scenarios of binary decision making. For this reason, in practical applications, combinations of these error measures need to be selected in order to obtain a dashboard overview of the complex relationships which allow realistic interpretations. Furthermore, we have seen that the contingency table provides a very useful summary of the outcome from decision making on which the calculation of many error measures is based on, with the exception of the AUROC.

Finally, we would like to note that, in practice, resampling methods (Efron, 1981; Molinaro, Simon, & Pfeiffer, 2005; Schumacher, Holländer, & Sauerbrei, 1997), for example, cross validation, will be applied for estimating variabilities and to protect against overfitting. This will allow estimating the mean value of error measures as well as the corresponding standard error. Hence, error measures are only one side of the medal when evaluating the outcome of binary decision making. Resampling methods complement these error measures by enabling estimates of the distribution of error measures.

ACKNOWLEDGMENTS

M.D. thanks the Austrian Science Funds for supporting this work (project P30031).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.
REFERENCES

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: An overview. *Bioinformatics*, 16(5), 412–424.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7), 1145–1159.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.

Clarke, B., Fokoue, E., & Zhang, H. H. (2009). Principles and theory for data mining and machine learning. Dordrecht/New York: Springer.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3), 589–599.

Emmert-Streib, F., Moutari, S., & Dehmer, M. (2016). A comprehensive survey of error measures for classification. *Pattern Recognition Letters*, 27, 861–874.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27, 861–874.

Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30(1), 27–38.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24(1), 38–49.

Flach, P. (2012). *Machine learning: The art of science and algorithms that make sense of data*. New York: Cambridge University Press.

Genovese, C. R., & Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476), 1408–1417.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Waltham, MA: Elsevier.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1), 29–36.

Hardin, J., Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., … Ward, M. D. (2015). Data science in statistics curricula: Preparing students to ‘think with data’. *The American Statistician*, 69(4), 343–353.

Hu, B. G., & Wang, Y. (2008). Evaluation criteria based on mutual information for classification including rejected class. *Acta Automatica Sinica*, 1396–1403.

Kleinbaum, D., & Klein, M. (2005). *Survival analysis: A self-learning text*. New York: Springer.

Lehman, E. (2005). *Testing statistical hypotheses*. New York: Springer.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta*, 405(2), 442–451.

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: A comparison of resampling methods. *Bioinformatics*, 21(15), 3301–3307.

Powers, D. M. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63.

Principe, J. C., Xu, D. X., Zhao, Q., & Fisher, J. W. (2000). Learning from examples with information-theoretic criteria. *Signal Processing Systems*, 26(1–2), 61–77.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2), 584–599.

Shridhar, D., Bartlett, E., & Seagrave, R. (1998). Information theoretic subset selection for neural network models. *Computers and Chemical Engineering*, 22(4–5), 613–626.

Sokolova, M., & Lapalme, G. (2009). A systematic study of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.

Wallach, H. (2006). Evaluation metrics for hard classifiers. Technical Report. Cambridge University. Retrieved from http://www.inference.phy.cam.ac.uk/hmw26/papers

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1), 32–35.

---

How to cite this article: Emmert-Streib F, Moutari S, Dehmer M. A comprehensive survey of error measures for evaluating binary decision making in data science. *WIREs Data Mining Knowl Discov*. 2019;9:e1303. [https://doi.org/10.1002/widm.1303](https://doi.org/10.1002/widm.1303)