Classification assessment methods

Alaa Tharwat
Faculty of Computer Science and Engineering,
Frankfurt University of Applied Sciences, Frankfurt, Germany

Received 4 May 2018
Revised 7 August 2018
Accepted 17 August 2018

Abstract

Classification techniques have been applied to many applications in various fields of sciences. There are several ways of evaluating classification algorithms. The analysis of such metrics and its significance must be interpreted correctly for evaluating different learning algorithms. Most of these measures are scalar metrics and some of them are graphical methods. This paper introduces a detailed overview of the classification assessment measures with the aim of providing the basics of these measures and to show how it works to serve as a comprehensive source for researchers who are interested in this field. This overview starts by highlighting the definition of the confusion matrix in binary and multi-class classification problems. Many classification measures are also explained in details, and the influence of balanced and imbalanced data on each metric is presented. An illustrative example is introduced to show (1) how to calculate these measures in binary and multi-class classification problems, and (2) the robustness of some measures against balanced and imbalanced data. Moreover, some graphical measures such as Receiver operating characteristics (ROC), Precision-Recall, and Detection error trade-off (DET) curves are presented with details. Additionally, in a step-by-step approach, different numerical examples are demonstrated to explain the preprocessing steps of plotting ROC, PR, and DET curves.

Keywords Receiver operating characteristics (ROC), Confusion matrix, Precision-Recall (PR) curve, Classification, Assessment methods

Paper type Original Article

1. Introduction

Classification techniques have been applied to many applications in various fields of sciences. In classification models, the training data are used for building a classification model to predict the class label for a new sample. The outputs of classification models can be discrete as in the decision tree classifier or continuous as the Naive Bayes classifier [7]. However, the outputs of learning algorithms need to be assessed and analyzed carefully and this analysis must be interpreted correctly, so as to evaluate different learning algorithms.

The classification performance is represented by scalar values as in different metrics such as accuracy, sensitivity, and specificity. Comparing different classifiers using these measures is easy, but it has many problems such as the sensitivity to imbalanced data and ignoring the performance of some classes. Graphical assessment methods such as Receiver operating characteristics (ROC) and Precision-Recall curves give different interpretations of the classification performance.

Applied Computing and Informatics
Vol. 17 No. 1, 2021
pp. 168-192
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1016/j.aci.2018.08.003

Some of the measures which are derived from the confusion matrix for evaluating a diagnostic test are reported in [19]. In that paper, only eight measures were introduced. Powers introduced an excellent discussion of the precision, Recall, F-score, ROC, Informedness, Markedness and Correlation assessment methods with details explanations [16]. Sokolova et al. reported some metrics which are used in medical diagnosis [20]. Moreover, a good investigation of some measures and the robustness of these measures against different changes in the confusion matrix are introduced in [21]. Tom Fawcett presented a detailed introduction to the ROC curve including (1) good explanations of the basics of the ROC curve, (2) clear example for generating the ROC curve, (3) comprehensive discussions, and (4) good explanations of the Area under curve (AUC) metric [8]. Jesse Davis and Mark Goadrich reported the relationship between the ROC and Precision-Recall curves [5]. Our paper introduces a detailed overview of the classification assessment methods with the goal of providing the basic principles of these measures and to show how it works to serve as a comprehensive source for researchers who are interested in this field. This paper has details of most of the well-known classification assessment methods. Moreover, this paper introduces (1) the relations between different assessment methods, (2) numerical examples to show how to calculate these assessment methods, (3) the robustness of each method against imbalanced data which is one of the most important problems in real-time applications, and (4) explanations of different curves in a step-by-step approach.

This paper is divided into eight sections. Section 2 gives an overview of the classification assessment methods. This section begins by explaining the confusion matrix for binary and multi-class classification problems. Based on the data that can be extracted from the confusion matrix, many classification metrics can be calculated. Moreover, the influence of balanced and imbalanced data on each assessment method is introduced. Additionally, an illustrative numerical example is presented to show (1) how to calculate these measures in both binary and multi-class classification problems, and (2) the robustness of some measures against balanced and imbalanced data. Section 3 introduces the basics of the ROC curve, which are required for understanding how to plot and interpret it. This section also presents visualized steps with an illustrative example for plotting the ROC curve. The AUC measure is presented in Section 4. In this section, the AUC algorithm with detailed steps is explained. Section 5 presents the basics of the Precision-Recall curve and how to interpret it. Further, in a step-by-step approach, different numerical examples are demonstrated to explain the preprocessing steps of plotting ROC and PR curves in Sections 3 and 5. Classification assessment methods for biometric models including steps of plotting the DET curve are presented in Section 6. In Section 7, results in terms of different assessment methods of a simple experiment are presented. Finally, concluding remarks will be given in Section 8.

2. Classification performance

The assessment method is a key factor in evaluating the classification performance and guiding the classifier modeling. There are three main phases of the classification process, namely, *training* phase, *validation* phase, and *testing* phase. The model is trained using input patterns and this phase is called the training phase. These input patterns are called training data which are used for training the model. During this phase, the parameters of a classification model are adjusted. The training error measures how well the trained model fits the training data. However, the training error always smaller than the testing error and the validation error because the trained model fits the same data which are used in the training phase. The goal of a learning algorithm is to learn from the training data to predict class labels for unseen data; this is in the testing phase. However, the testing error or out-of-sample error cannot be estimated because the class labels or outputs of testing samples are unknown. This is the reason why the validation phase is used for evaluating the performance of the trained

model. In the validation phase, the validation data provide an unbiased evaluation of the trained model while tuning the model's hyperparameters.

According to the number of classes, there are two types of classification problems, namely, binary classification where there are only two classes, and multi-class classification where the number of classes is higher than two. Assume we have two classes, i.e., binary classification, P for *positive* class and N for *negative* class. An unknown sample is classified to P or N. The classification model that was trained in the training phase is used to predict the true classes of unknown samples. This classification model produces continuous or discrete outputs. The discrete output that is generated from a classification model represents the predicted discrete class label of the unknown/test sample, while continuous output represents the estimation of the sample's class membership probability.

Figure 1 shows that there are four possible outputs which represent the elements of a 2×2 *confusion matrix* or a *contingency table*. The green diagonal represents correct predictions and the pink diagonal indicates the incorrect predictions. If the sample is positive and it is classified as positive, i.e., correctly classified positive sample, it is counted as a *true positive* (*TP*); if it is classified as negative, it is considered as a *false negative* (*FN*) or *Type II error*. If the sample is negative and it is classified as negative it is considered as *true negative* (*TN*); if it is classified as positive, it is counted as *false positive* (*FP*), *false alarm* or *Type I error*. As we will present in the next sections, the confusion matrix is used to calculate many common classification metrics.

Figure 2 shows the confusion matrix for a multi-class classification problem with three classes (A, B, and C). As shown, TP_A is the number of true positive samples in class A, i.e., the number of samples that are correctly classified from class A, and E_{AB} is the samples from class A that were incorrectly classified as class B, i.e., misclassified samples. Thus, the false negative in the A class (FN_A) is the sum of E_{AB} and E_{AC} ($FN_A = E_{AB} + E_{AC}$) which indicates the sum of all class A samples that were incorrectly classified as class B or C. Simply, FN of any class which is located in a column can be calculated by adding the errors in that class/column. Whereas the false positive for any predicted class which is located in a row represents the sum of all errors in that row. For example, the false positive in class A (FP_A) is calculated as follows, $FP_A = E_{BA} + E_{CA}$. With $m \times m$ confusion matrix there are m correct classifications and $m^2 - m$ possible errors [22].

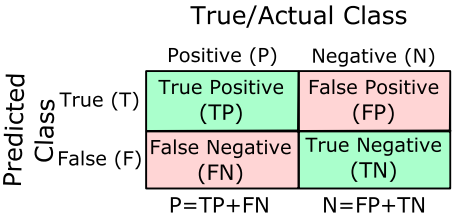

Figure 1.
An illustrative example of the 2×2 confusion matrix. There are two true classes P and N. The output of the predicted class is true or false.

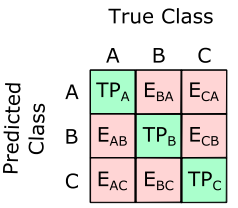

Figure 2.
An illustrative example of the confusion matrix for a multi-class classification test.

2.1 Classification metrics with imbalanced data

Different assessment methods are sensitive to the imbalanced data when the samples of one class in a dataset outnumber the samples of the other class(es) [25]. To explain this is so, consider the confusion matrix in Figure 1. The class distribution is the ratio between the positive and negative samples ($\frac{P}{N}$) represents the relationship between the left column to the right column. Any assessment metric that uses values from both columns will be sensitive to the imbalanced data as reported in [8]. For example, some metrics such as accuracy and precision[1] use values from both columns of the confusion matrix; thus, as data distributions change, these metrics will change as well, even if the classifier performance does not. Therefore, such these metrics cannot distinguish between the numbers of corrected labels from different classes [11]. This fact is partially true because there are some metrics such as Geometric Mean (*GM*) and Youden's index (*YI*)[2] use values from both columns and these metrics can be used with balanced and imbalanced data. This can be interpreted as that the metrics which use values from one column cancel the changes in the class distribution. However, some metrics which use values from both columns are not sensitive to the imbalanced data because the changes in the class distribution cancel each other. For example, the accuracy is defined as follows, $Acc = \frac{TP+TN}{TP+TN+FP+FN}$ and the *GM* is defined as follows, $GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$; thus, both metrics use values from both columns of the confusion matrix. Changing the class distribution can be obtained by increasing/decreasing the number of samples of negative/positive class. With the same classification performance, assume that the negative class samples are increased by α times; thus, the *TN* and *FP* values will be αTN and αFP, respectively; thus, the accuracy will be, $Acc = \frac{TP+\alpha TN}{TP+\alpha TN+\alpha FP+FN} \neq \frac{TP+TN}{TP+TN+FP+FN}$. This means that the accuracy is affected by the changes in the class distribution. On the other hand, the *GM* metric will be, $GM = \sqrt{\frac{TP}{TP+FN} \times \frac{\alpha TN}{\alpha TN+\alpha FP}} = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}}$ and hence the changes in the negative class cancel each other. This is the reason why the *GM* metric is suitable for the imbalanced data. Similarly, any metric can be checked to know if it is sensitive to the imbalanced data or not.

2.2 Accuracy and error rate

Accuracy (*Acc*) is one of the most commonly used measures for the classification performance, and it is defined as a ratio between the correctly classified samples to the total number of samples as follows [20]:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where *P* and *N* indicate the number of positive and negative samples, respectively.

The complement of the accuracy metric is the *Error rate* (*ERR*) or *misclassification rate*. This metric represents the number of misclassified samples from both positive and negative classes, and it is calculated as follows, $EER = 1 - Acc = (FP + FN)/(TP + TN + FP + FN)$ [4]. Both accuracy and error rate metrics are sensitive to the imbalanced data. Another problem with the accuracy is that two classifiers can yield the same accuracy but perform differently with respect to the types of correct and incorrect decisions they provide [9]. However, Takaya Saito and Marc Rehmsmeier reported that the accuracy is suitable with imbalanced data because they found that the accuracy values of the balanced and imbalanced data in their example were identical [17]. The reason why the accuracy values were identical in their example is that the sum of *TP* and *TN* in the balanced and imbalanced data was the same.

2.3 Sensitivity and specificity

Sensitivity, *True positive rate* (*TPR*), *hit rate*, or *recall*, of a classifier represents the positive correctly classified samples to the total number of positive samples, and it is estimated according to Eq. (2) [20]. Whereas *specificity*, *True negative rate* (*TNR*), or inverse recall is expressed as the ratio of the correctly classified negative samples to the total number of negative samples as in Eq. (2) [20]. Thus, the specificity represents the proportion of the negative samples that were correctly classified, and the sensitivity is the proportion of the positive samples that were correctly classified. Generally, we can consider sensitivity and specificity as two kinds of accuracy, where the first for actual positive samples and the second for actual negative samples. Sensitivity depends on *TP* and *FN* which are in the same column of the confusion matrix, and similarly, the specificity metric depends on *TN* and *FP* which are in the same column; hence, both sensitivity and specificity can be used for evaluating the classification performance with imbalanced data [9].

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}, \; TNR = \frac{TN}{FP + TN} = \frac{TN}{N} \tag{2}$$

The accuracy can also be defined in terms of sensitivity and specificity as follows [20]:

$$\begin{aligned} Acc &= \frac{TP + TN}{TP + TN + FP + FN} \\ &= TPR \times \frac{P}{P + N} + TNR \times \frac{N}{P + N} \\ &= \frac{TP}{TP + FN}\frac{P}{P + N} + \frac{TN}{TN + FP}\frac{N}{P + N} \\ &= \frac{TP}{P + N} + \frac{TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \tag{3}$$

2.4 False positive and false negative rates

False positive rate (*FPR*) is also called *false alarm rate* (*FAR*), or Fallout, and it represents the ratio between the incorrectly classified negative samples to the total number of negative samples [16]. In other words, it is the proportion of the negative samples that were incorrectly classified. Hence, it complements the specificity as in Eq. (4) [21]. The *False negative rate* (*FNR*) or *miss rate* is the proportion of positive samples that were incorrectly classified. Thus, it complements the sensitivity measure and it is defined in Eq. (5). Both FPR and FNR are not sensitive to changes in data distributions and hence both metrics can be used with imbalanced data [9].

$$FPR = 1 - TNR = \frac{FP}{FP + TN} = \frac{FP}{N} \tag{4}$$

$$FNR = 1 - TPR = \frac{FN}{FN + TP} = \frac{FN}{P} \tag{5}$$

2.5 Predictive values

Predictive values (positive and negative) reflect the performance of the prediction. *Positive prediction value* (*PPV*) or *precision* represents the proportion of positive samples that were correctly classified to the total number of positive predicted samples as indicated in Eq. (6) [20]. On the contrary, *Negative predictive value* (*NPV*), inverse precision, or true negative accuracy (*TNA*) measures the proportion of negative samples that were correctly classified to the total

number of negative predicted samples as indicated in Eq. (7) [16]. These two measures are sensitive to the imbalanced data [21,9]. *False discovery rate* (FDR) and *False omission rate* (FOR) measures complements the PPV and NPV, respectively (see Eq. (6) and (7)).

$$PPV = \text{Precision} = \frac{TP}{FP + TP} = 1 - FDR \tag{6}$$

$$NPV = \frac{TN}{FN + TN} = 1 - FOR \tag{7}$$

The accuracy can also be defined in terms of precision and inverse precision as follows [16]:

$$\begin{aligned} Acc &= \frac{TP + FP}{P + N} \times PPV + \frac{TN + FN}{P + N} \times NPV \\ &= \frac{TP + FP}{P + N} \times \frac{TP}{TP + FP} + \frac{TN + FN}{P + N} \times \frac{TN}{TN + FN} \\ &= \frac{TP + TN}{TP + TN + FP + FN} \end{aligned} \tag{8}$$

2.6 Likelihood ratio

The likelihood ratio combines both sensitivity and specificity, and it is used in diagnostic tests. In that tests, not all positive results are true positives and also the same for negative results; hence, the positive and negative results change the probability/likelihood of diseases. Likelihood ratio measures the influence of a result on the probability. *Positive likelihood* ($LR+$) measures how much the odds of the disease increases when a diagnostic test is positive, and it is calculated as in Eq. (9) [20]. Similarly, *Negative likelihood* ($LR-$) measures how much the odds of the disease decreases when a diagnostic test is negative, and it is calculated as in Eq. (9). Both measures depend on the sensitivity and specificity measures; thus, they are suitable for balanced and imbalanced data [6].

$$LR+ = \frac{TPR}{1 - TNR} = \frac{TPR}{FPR}, \; LR- = \frac{1 - TPR}{TNR} \tag{9}$$

Both $LR+$ and $LR-$ are combined into one measure which summarizes the performance of the test, this measure is called *Diagnostic odds ratio* (*DOR*). The *DOR* metric represents the ratio between the positive likelihood ratio to the negative likelihood ratio as in Eq. (10). This measure is utilized for estimating the discriminative ability of the test and also for comparing between two diagnostic tests. From Eq. (10) it can be remarked that the value of *DOR* increases when (1) the *TP* and *TN* are high and (2) the *FP* and *FN* are low [18].

$$DOR = \frac{LR+}{LR-} = \frac{TPR}{1 - TNR} \times \frac{TNR}{1 - TPR} = \frac{TP \times TN}{FP \times FN} \tag{10}$$

2.7 Youden's index

Youden's index (*YI*) or *Bookmaker Informedness* (BM) metric is one of the well-known diagnostic tests. It evaluates the discriminative power of the test. The formula of Youden's index combines the sensitivity and specificity as in the DOR metric, and it is defined as follows, $YI = TPR + TNR - 1$ [20]. The *YI* metric is ranged from zero when the test is poor to one which represents a perfect diagnostic test. It is also suitable with imbalanced data. One of the major disadvantages of this test is that it does not change concerning the differences between the sensitivity and specificity of the test. For example, given two tests, the sensitivity

values for the first and second tests are 0.7 and 0.9, respectively, and the specificity values for the first and second tests are 0.8 and 0.6, respectively; the *YI* value for both tests is 0.5.

2.8 Another metrics

There are many different metrics that can be calculated from the previous metrics. Some details about each measure are as follow:

- *Matthews correlation coefficient* (*MCC*): this metric was introduced by Brian W. Matthews in 1975 [14], and it represents the correlation between the observed and predicted classifications, and it is calculated directly from the confusion matrix as in Eq. (11). A coefficient of +1 indicates a perfect prediction, −1 represents total disagreement between prediction and true values and zero means that no better than random prediction [16,3]. This metric is sensitive to imbalanced data.

$$\begin{aligned} MCC &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ &= \frac{\frac{TP}{N} - TPR \times PPV}{\sqrt{PPV \times TPR(1-TPR)(1-PPV)}} \end{aligned} \tag{11}$$

- *Discriminant power* (*DP*): this measure depends on the sensitivity and specificity and it is defined as follows, $DP = \frac{\sqrt{3}}{\pi}\left(log\left(\frac{TPR}{1-TNR}\right) + log\left(\frac{TNR}{1-TPR}\right)\right)$ [20]. This metric evaluates how well the classification model distinguishes between positive and negative samples. Since this metric depends on the sensitivity and specificity metrics; it can be used with imbalanced data.
- *F-measure:* this is also called F_1-score, and it represents the harmonic mean of precision and recall as in Eq. (12) [20]. The value of *F*-measure is ranged from zero to one, and high values of *F*-measure indicate high classification performance. This measure has another variant which is called F_β-measure. This variant represents the weighted harmonic mean between precision and recall as in Eq. (13). This metric is sensitive to changes in data distributions. Assume that the negative class samples are increased by α times; thus, the $F-$measure is calculated as follows, $F-\text{measure} = \frac{2TP}{2TP+\alpha FP+\alpha FN}$ and hence this metric is affected by the changes in the class distribution.

$$\begin{aligned} F-\text{measure} &= \frac{2PPV \times TPR}{PPV + TPR} \\ &= \frac{2TP}{2TP+FP+FN} \end{aligned} \tag{12}$$

$$\begin{aligned} F_\beta-\text{measure} &= \left(1+\beta^2\right)\frac{PPV.TPR}{\beta^2 PPV + TPR} \\ &= \frac{(1+\beta^2)TP}{(1+\beta^2)TP+\beta^2 FN+FP} \end{aligned} \tag{13}$$

Adjusted *F*-measure (*AGF*) was introduced in [13]. The *F*-measures used only three of the four elements of the confusion matrix and hence two classifiers with different *TNR* values may have the same *F*-score. Therefore, the *AGF* metric is introduced to use all elements of the

confusion matrix and provide more weights to samples which are correctly classified in the minority class. This metric is defined as follows:

$$AGF = \sqrt{F_2.InvF_{0.5}} \tag{14}$$

where F_2 is the F-measure where $\beta = 2$ and $InvF_{0.5}$ is calculated by building a new confusion matrix where the class label of each sample is switched (i.e. positive samples become negative and vice versa).

- *Markedness* (*MK*): this is defined based on PPV and NPV metrics as follows, $MK = PPV + NPV - 1$ [16]. This metric sensitive to data changes and hence it is not suitable for imbalanced data. This is because the Markedness metric depends on PPV and NPV metrics and both PPV and NPV are sensitive to changes in data distributions.
- *Balanced classification rate or balanced accuracy* (*BCR*): this metric combines the sensitivity and specificity metrics and it is calculated as follows, $BCR = \frac{1}{2}(TPR + TNR) = \frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$. Also, *Balance error rate* (*BER*) or *Half total error rate* (*HTER*) represents $1 - BCR$. Both *BCR* and *BER* metrics can be used with imbalanced datasets.
- *Geometric Mean* (*GM*): The main goal of all classifiers is to improve the sensitivity, without sacrificing the specificity. However, the aims of sensitivity and specificity are often conflicting, which may not work well, especially when the dataset is imbalanced. Hence, the *Geometric Mean* (*GM*) metric aggregates both sensitivity and specificity measures according to Eq. (15) [3]. *Adjusted Geometric Mean* (*AGM*) is proposed to obtain as much information as possible about each class [11]. The *AGM* metric is defined according to Eq. (16).

$$GM = \sqrt{TPR \times TNR} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}} \tag{15}$$

$$AGM = \begin{cases} \dfrac{GM + TNR(FP + TN)}{1 + FP + TN} & \text{if } TPR > 0 \\ 0 & \text{if } TPR = 0 \end{cases} \tag{16}$$

GM metric can be used with imbalanced datasets. Lopez et al. reported that the *AGM* metric is suitable with the imbalanced data [12]. However, changing the distribution of negative class has a small influence on the *AGM* metric and hence it is not suitable with the imbalanced data. This is can be proved simply by assuming that the negative class samples are increased by α times. Thus, the *AGM* metric is calculated as follows, $AGM = \frac{GM + TNR(\alpha FP + \alpha TN)}{1 + \alpha FP + \alpha TN}$; as a consequence, the *AGM* metric is slightly affected by the changes in the class distribution.

- *Optimization precision* (*OP*): This metric is defined as follows:

$$OP = Acc - \frac{|TPR - TNR|}{TPR + TNR} \tag{17}$$

where the second term $\frac{|TPR - TNR|}{TPR + TNR}$ computes how balanced both class accuracies are and this metric represents the difference between the global accuracy and that term [9]. High *OP* value indicates high accuracy and well-balanced class accuracies. Since the *OP* metric depends on the accuracy metric, it is not suitable for imbalanced data.

- *Jaccard*: This metric is also called Tanimoto similarity coefficient. Jaccard metric explicitly ignores the correct classification of negative samples as follows, $Jaccard = \frac{TP}{TP+FP+FN}$. Jaccard metric is sensitive to changes in data distributions.

Figure 4 shows the relations between different classification assessment methods. As shown, all assessment methods can be calculated from the confusion matrix. As shown, there are two classes; red class and blue class. After applying a classifier, the classifier is represented by a black circle and the samples which are inside the circle are classified as red class samples and the samples outside the circle are classified as blue class samples. Additionally, from the figure, it is clear that many assessment methods depend on the *TPR* and *TNR* metrics, and all assessment methods can be estimated from the confusion matrix.

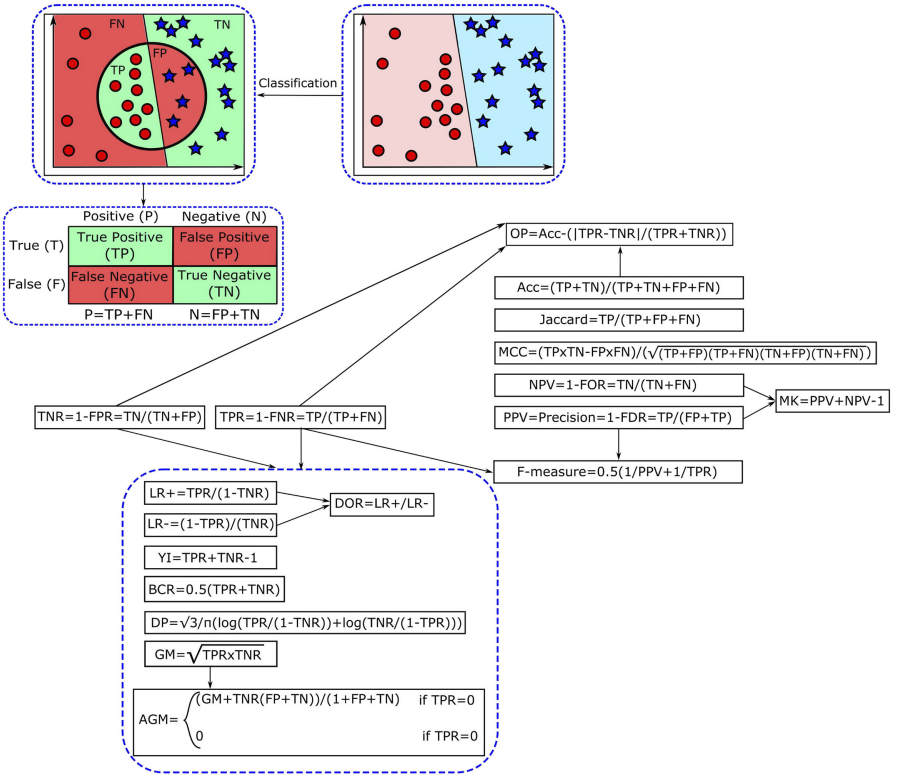

Figure 3.
Visualization of different metrics and the relations between these metrics. Given two classes, red class and blue class. The black circle represents a classifier that classifies the sample inside the circle as red samples (belong to the red class) and the samples outside the circle as blue samples (belong to the blue class). Green regions indicate the correctly classified regions and the red regions indicate the misclassified regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

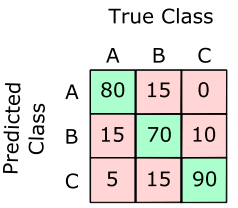

Figure 4.
Results of a multi-class classification test (our example).

2.9 Illustrative example

In this section, two examples are introduced. These examples explain how to calculate classification metrics using two classes or multiple classes.

2.9.1 Binary classification example. In this example, assume we have two classes (A and B), i.e., binary classification, and each class has 100 samples. The A class represents the positive class while the B class represents the negative class. The number of correctly classified samples in class A and B are 70 and 80, respectively. Hence, the values of TP, TN, FP, and FN are 70, 80, 20, and 30, respectively. The values of different classification metrics are as follows, $Acc = \frac{70+80}{70+80+20+30} = 0.75$, $TPR = \frac{70}{70+30} = 0.7$, $TNR = \frac{80}{80+20} = 0.8$, $PPV = \frac{70}{70+20} \approx 0.78$, $NPV = \frac{80}{80+30} \approx 0.73$, $Err = 1 - Acc = 0.25, BCR = \frac{1}{2}(0.7 + 0.8) = 0.75, FPR = 1 - 0.8 = 0.2, FNR = 1 - 0.7 = 0.3, F-\text{measure} = \frac{2\times70}{(2\times70+20+30)} = 0.74, OP = Acc - \frac{|TPR-TNR|}{TPR+TNR} = 0.75 - \frac{|0.7-0.8|}{0.7+0.8} \approx 0.683$, $LR+ = \frac{0.7}{1-0.8} = 3.5$, $LR- = \frac{1-0.7}{0.8} = 0.375$, $DOR = \frac{3.5}{0.375} \approx 9.33$, $YI = 0.7 + 0.8 - 1 = 0.5$, and $Jaccard = \frac{70}{70+20+30} \approx 0.583$.

We increased the number of samples of the B class to 1000 to show how the classification metrics are changed when using imbalanced data, and there are 800 samples from class B were correctly classified. As a consequence, the values of TP, TN, FP, and FN are 70, 800, 200, and 30, respectively. Consequently, only the values of accuracy, precision/PPV, NPV, error rate, Optimization precision, F-measure, and Jaccard are changed as follows, $Acc = \frac{70+800}{70+800+200+30} \approx 0.79$, $PPV = \frac{70}{70+200} \approx 0.26, NPV = \frac{800}{800+30} \approx 0.96$, $Err = 1 - Acc = 0.21, OP = 0.79 - \frac{|0.7-0.8|}{0.7+0.8} \approx 0.723, F-\text{measure} = \frac{2\times70}{(2\times70+200+30)} = 0.378$, and $Jaccard = \frac{70}{70+200+30} \approx 0.233$. This example reflects that the accuracy, precision, NPV, F-measure, and Jaccard metrics are sensitive to imbalanced data.

2.9.2 Multi-classification example. In this example, there are three classes A, B, and C, the results of a classification test are shown in Figure 4. From the figure, the values of TP_A, TP_B, and TB_C are 80, 70, and 90, respectively, which represent the diagonal in Figure 4. The values of false negative for each class (true class) are calculated as mentioned before by adding all errors in the column of that class. For example, $FN_A = E_{AB} + E_{AC} = 15 + 5 = 20$, and similarly $FN_B = E_{BA} + E_{BC} = 15 + 15 = 30$ and $FN_C = E_{CA} + E_{CB} = 0 + 10 = 10$. The values of false positive for each class (predicted class) are calculated as mentioned before by adding all errors in the row of that class. For example, $FP_A = E_{BA} + E_{CA} = 15 + 0 = 15$, and similarly $FP_B = E_{AB} + E_{CB} = 15 + 10 = 25$ and $FP_C = E_{AC} + E_{BC} = 5 + 15 = 20$. The value of true negative for the class A (TN_A) can be calculated by adding all columns and rows excluding the row and column of class A; this is similar to the TN in the 2×2 confusion matrix. Hence, the value of TN_A is calculated as follows, $TN_A = 70 + 90 + 10 + 15 = 185$, and similarly $TN_B = 80 + 0 + 5 + 90 = 175$ and $TN_C = 80 + 70 + 15 + 15 = 180$. Using TP, TN, FP, and FN we can calculate all classification measures. For example, the accuracy is $\frac{80+70+90}{100+100+100} = 0.8$. The sensitivity and specificity are calculated for each class. For example, the sensitivity of A is $\frac{TP_A}{TP_A+FN_A} = \frac{80}{80+15+5} = 0.8$, and similarly the sensitivity of B and C classes are $\frac{70}{70+15+15} = 0.7$ and $\frac{90}{90+0+10} = 0.9$, respectively, and the specificity values of A, B, and C are $\frac{185}{185+15} \approx 0.93$, $\frac{175}{(175+25)} = 0.875$, and $\frac{180}{(180+20)} = 0.9$, respectively.

3. Receiver operating characteristics (ROC)

The *receiver operating characteristics* (ROC) curve is a two-dimensional graph in which the TPR represents the y-axis and FPR is the x-axis. The ROC curve has been used to evaluate many systems such as diagnostic systems, medical decision-making systems, and machine learning systems [26]. It is used to make a balance between the benefits, i.e., true positives, and costs, i.e., false positives. Any classifier that has discrete outputs such as decision trees is

designed to produce only a class decision, i.e., a decision for each testing sample, and hence it generates only one confusion matrix which in turn corresponds to one point into the ROC space. However, there are many methods that were introduced for generating full ROC curve from a classifier instead of only a single point such as using class proportions [26] or using some combinations of scoring and voting [8]. On the other hand, in continuous output classifiers such as the Naive Bayes classifier, the output is represented by a numeric value, i.e., score, which represents the degree to which a sample belongs to a specific class. The ROC curve is generated by changing the threshold on the confidence score; hence, each threshold generates only one point in the ROC curve [8].

Figure 5 shows an example of the ROC curve. As shown, there are four important points in the ROC curve. The point A, in the lower left corner $(0, 0)$ represents a classifier where there is no positive classification, while all negative samples are correctly classified and hence $TPR = 0$ and $FPR = 0$. The point C, in the top right corner (1,1), represents a classifier where all positive samples are correctly classified, while the negative samples are misclassified. The point D in the lower right corner $(1, 0)$ represents a classifier where all positive and negative samples are misclassified. The point B in the upper left corner $(0, 1)$ represents a classifier where all positive and negative samples are correctly classified; thus, this point represents the perfect classification or the *Ideal operating point*. Figure 5 shows the perfect classification performance. It is the green curve which rises vertically from (0,0) to (0,1) and then horizontally to (1,1). This curve reflects that the classifier perfectly ranked the positive samples relative to the negative samples.

A point in the ROC space is better than all other points that are in the southeast, i.e., the points that have lower *TPR*, higher *FPR*, or both (see Figure 5). Therefore, any classifier appears in the lower right triangle performs worse than the classifier appears in the upper left triangle.

Figure 6 shows an example of the ROC curve. In this example, a test set consists of 20 samples from two classes; each class has ten samples, i.e., ten positive and ten negative samples. As shown in the table in Figure 6, the initial step to plot the ROC curve is to sort the samples according to their scores. Next, the threshold value is changed from maximum to minimum to plot the ROC curve. To scan all samples, the threshold is ranged from ∞ to $-\infty$. The samples are classified into the positive class if their scores are higher than or equal the

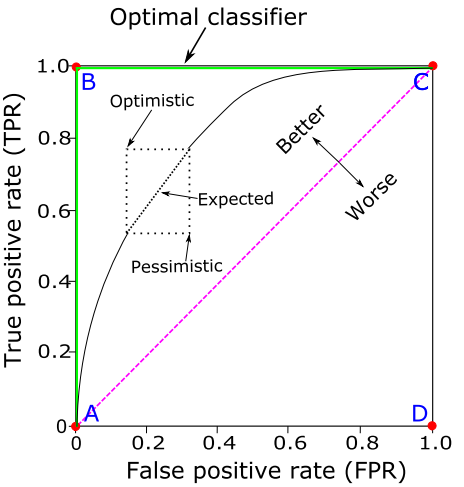

Figure 5.
A basic ROC curve showing important points, and the optimistic, pessimistic and expected ROC segments for equally scored samples.

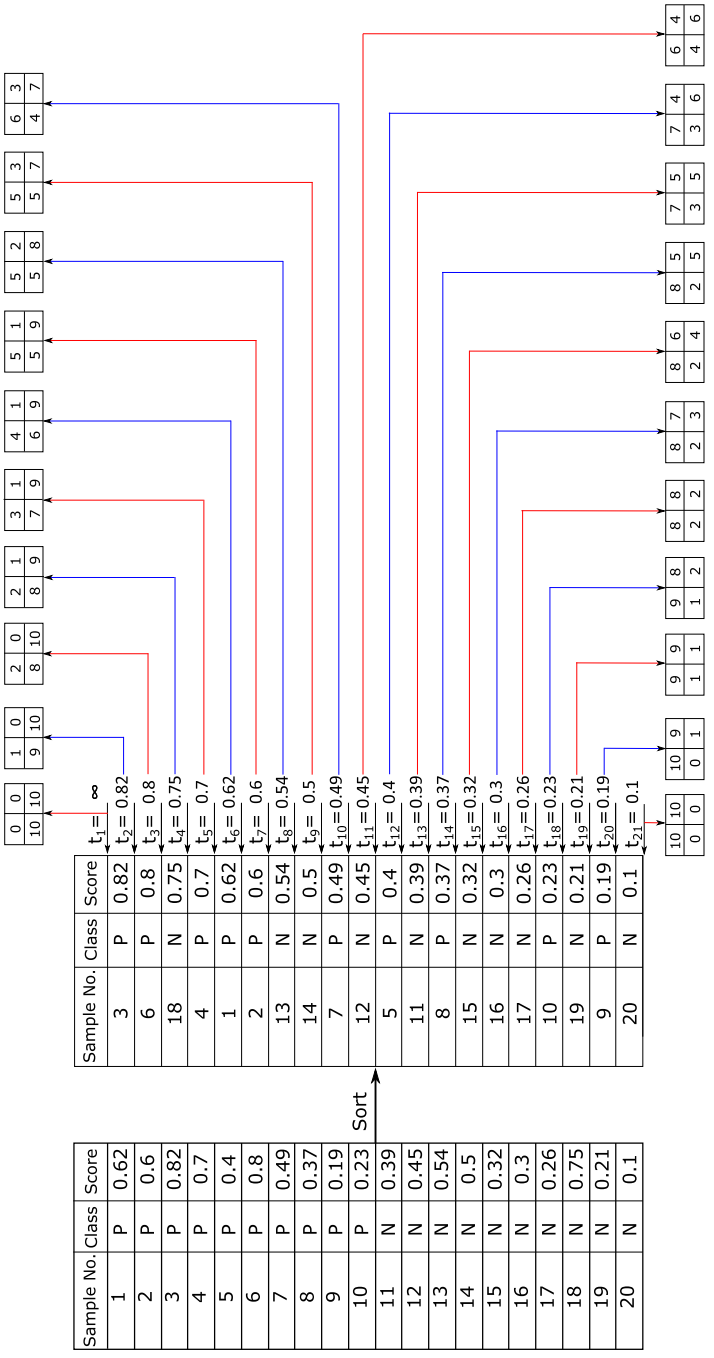

Figure 6.
An illustrative example to calculate the *TPR* and *FPR* when the threshold value is changed.

threshold; otherwise, it is estimated as negative [8]. Figures 7 and 8 shows how changing the threshold value changes the *TPR* and *FPR*. As shown in Figure 6, the threshold value is set at maximum ($t_1 = \infty$); hence, all samples are classified as negative samples and the values of *FPR* and *TPR* are zeros and the position of t_1 is in the lower left corner (the point (0,0)). The threshold value is decreased to 0.82, and the first sample is classified correctly as a positive sample (see Figures 6–8(a)). The *TPR* increased to 0.1, while the *FPR* remains zero. As the threshold is further reduced to be 0.8, the *TPR* is increased to 0.2 and the *FPR* remains zero. As shown in Figure 7, increasing the *TPR* moves the ROC curve up while increasing the *FPR* moves the ROC curve to the right as in t_4. The ROC curve must pass through the point (0,0) where the threshold value is ∞ (in which all samples are classified as negative samples) and the point (1,1) where the threshold is $-\infty$ (in which all samples are classified as positive samples).

Figure 8 shows graphically the performance of the classification model with different threshold values. From this figure, the following remarks can be drawn.

- t_1: The value of this threshold was ∞ as shown in Figure 8a) and hence all samples are classified as negative samples. This means that (1) all positive samples are incorrectly classified; hence, the value of *TP* is zero, (2) all negative samples are correctly classified and hence there is no *FP* (see also Figure 6).
- t_3: The threshold value decreased as shown in Figure 8b) and as shown there are two positive samples are correctly classified. Therefore, according to the positive class, only the positive samples which have scores more than or equal this threshold (t_3) will be correctly classified, i.e., *TP*, while the other positive samples are incorrectly classified, i.e., *FN*. In this threshold, also all negative samples are correctly classified; thus, the value of *FP* is still zero.
- t_8: As the threshold further decreased to be 0.54, the threshold line moves to the left. This means that more positive samples have the chance to be correctly classified; on the other hand, some negative samples are misclassified. As a consequence, the values of *TP* and *FP* are increased as shown in Figure 8(c), and the values of *TN* and *FN* decreased.

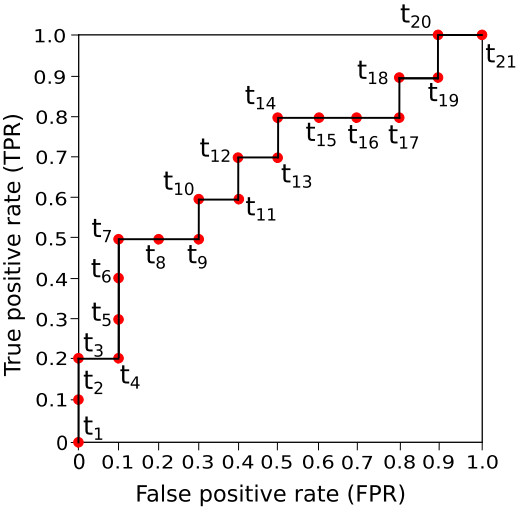

Figure 7.
An illustrative example of the ROC curve. The values of *TPR* and *FPR* of each point/threshold are calculated in Table 1.

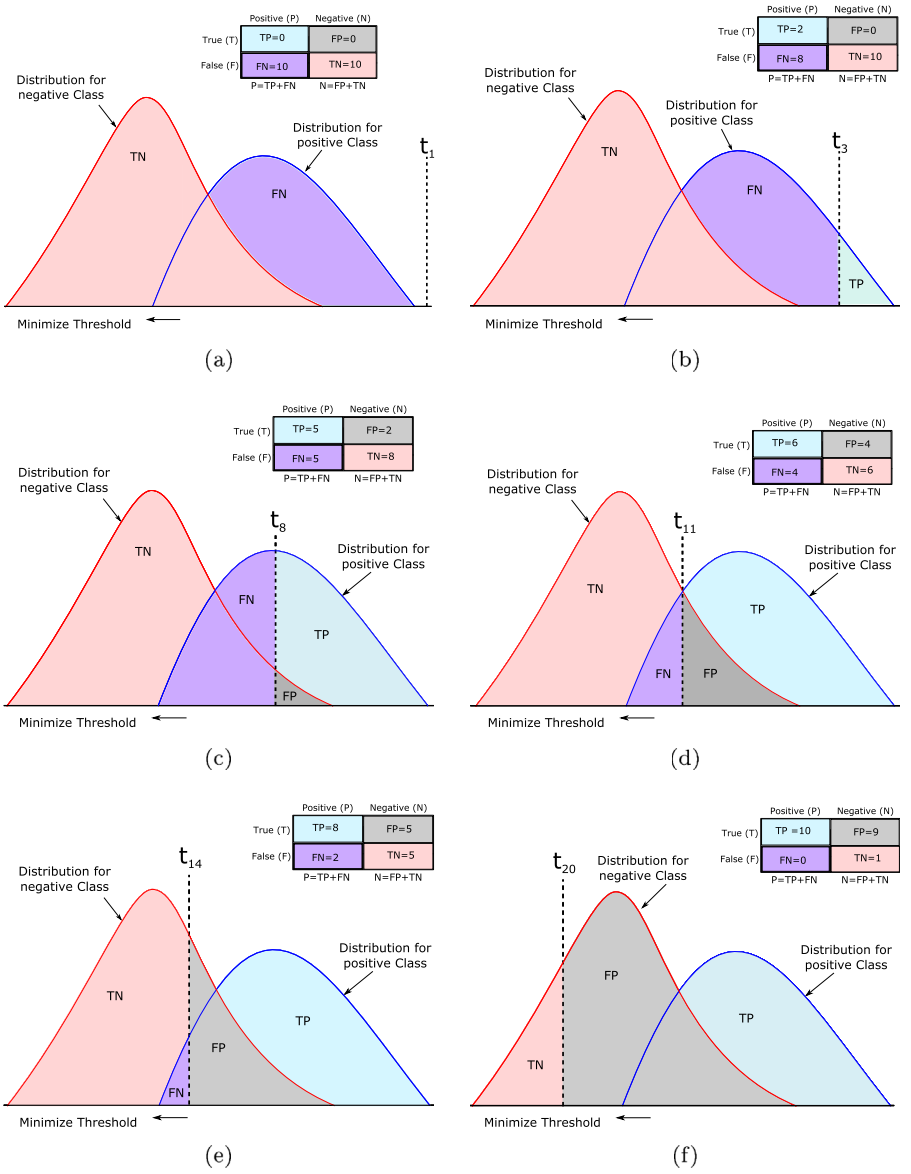

Figure 8.
A visualization of how changing the threshold changes the TP, TN, FP, and FN values.

- t_{11}: This is an important threshold value where the numbers of errors from both positive and negative classes are equal (see Figure 8(d)) $TP = TN = 6$ and $FP = FN = 4$).
- t_{14}: Reducing the value of the threshold to 0.37 results more correctly classified positive samples and this increases TP and reduces FN as shown in Figure 8(e). On the

contrary, more negative samples are misclassified and this increases *FP* and reduces *TN*.

- t_{20}: As shown in Figure 8(f), decreasing the threshold value hides the *FN* area. This is because all positive samples are correctly classified. Also, from the figure, it is clear that the *FP* area is much larger than the area of *TN*. This is because 90% of the negative samples are incorrectly classified, and only 10% of negative samples are correctly classified.

From Figure 7 it is clear that the ROC curve is a step function. This is because we only used 20 samples (a finite set of samples) in our example and a true curve can be obtained when the number of samples increased. The figure also shows that the best accuracy (70%) (see Table 1) is obtained at (0.1,0.5) when the threshold value was ≥ 0.6, rather than at ≥ 0.5 as we might expect with a balanced data. This means that the given learning model identifies positive samples better than negative samples. Since the ROC curve depends mainly on changing the threshold value, comparing classifiers with different score ranges will be meaningless. For example, assume we have two classifiers, the first generates scores in the range [0,1] and the other generates scores in the range $[-1,+1]$ and hence we cannot compare these classifiers using the ROC curve.

The steps of generating ROC curve are summarized in Algorithm 1. The algorithm requires $O(\text{nlogn})$ for sorting samples, and $O(n)$ for scanning them; resulting in $O(\text{nlogn})$ total complexity, where n is the number of samples. As shown, the two main steps to generate ROC points are (1) sorting samples according to their scores and (2) changing the threshold value from maximum to minimum to process one sample at a time and update the values of *TP* and *FP* in each time. The algorithm shows that the *TP* and the *FP* start at zero. The algorithm scans all samples and the value of *TP* is increased for each positive sample while the value of *FP* is increased for each negative sample. Next, the values of *TPR* and *FPR* are calculated and pushed into the ROC stack (see step 6). When the threshold becomes very low

Table 1.
Values of *TP*, *FN*, *TN*, *FP*, *TPR*, *FPR*, *FNR*, precision (*PPV*), and accuracy (*Acc* in %) of our ROC example when changes the threshold value.

Threshold	*TP*	*FN*	*TN*	*FP*	*TPR*	*FPR*	*FNR*	*PPV*	*Acc*
$t_1 = \infty$	0	10	10	0	0	0	1	–	50
$t_2 = 0.82$	1	9	10	0	0.1	0	0.9	1.0	55
$t_3 = 0.80$	2	8	10	0	0.2	0	0.8	1.0	60
$t_4 = 0.75$	2	8	9	1	0.2	0.1	0.8	0.67	55
$t_5 = 0.70$	3	7	9	1	0.3	0.1	0.7	0.75	60
$t_6 = 0.62$	4	6	9	1	0.4	0.1	0.6	0.80	65
$t_7 = 0.60$	5	5	9	1	0.5	0.1	0.5	0.83	70
$t_8 = 0.54$	5	5	8	2	0.5	0.2	0.5	0.71	65
$t_9 = 0.50$	5	5	7	3	0.5	0.3	0.5	0.63	60
$t_{10} = 0.49$	6	4	7	3	0.6	0.3	0.4	0.67	65
$t_{11} = 0.45$	6	4	6	4	0.6	0.4	0.4	0.60	60
$t_{12} = 0.40$	7	3	6	4	0.7	0.4	0.3	0.64	65
$t_{13} = 0.39$	7	3	5	5	0.7	0.5	0.3	0.58	60
$t_{14} = 0.37$	8	2	5	5	0.8	0.5	0.2	0.62	65
$t_{15} = 0.32$	8	2	4	6	0.8	0.6	0.2	0.57	60
$t_{16} = 0.30$	8	2	3	7	0.8	0.7	0.2	0.53	55
$t_{17} = 0.26$	8	2	2	8	0.8	0.8	0.2	0.50	50
$t_{18} = 0.23$	9	1	2	8	0.9	0.8	0.1	0.53	55
$t_{19} = 0.21$	9	1	1	9	0.9	0.9	0.1	0.50	50
$t_{20} = 0.19$	10	0	1	9	1.0	0.9	0	0.53	55
$t_{21} = 0.10$	10	0	0	10	1.0	1.0	0	0.50	50

(threshold $\rightarrow -\infty$), all samples are classified as positive samples and hence the values of both *TPR* and *FPR* are one.

Steps 5–8 handle sequences of equally scored samples. Assume we have a test set which consists of P positive samples and N negative samples. In this test set, assume we have p positive samples and n negative samples with the same score value. There are two extreme cases. In the first case which is the optimistic case, all positive samples end up at the beginning of the sequence, and this case represents the upper L segment of the rectangle in Figure 5. In the second case, i.e., pessimistic case, all the negative samples end up at the beginning of the sequence, and this case represents the lower L segment of the rectangle in Figure 5. The ROC curve represents the expected performance which is the average of the two cases, and it represents the diagonal of the rectangle in Figure 5. The size of this rectangle is $\frac{pn}{PN}$, and the number of errors in both optimistic and pessimistic cases can be calculated as follows, $\frac{pn}{2PN}$.

Algorithm 1: Generating ROC Curve.

1: Given a set of test samples ($S_{test} = \{s_1, s_2, \ldots, s_N\}$), where N is the total number of test samples, $f(i)$ is the classifier that classify the ith sample to positive or negative classes, P and N represent the total number of positive and negative samples, respectively.
2: Sort the samples corresponding to their scores, where S_{sorted} is the sorted samples.
3: $FP \leftarrow 0$, $TP \leftarrow 0$, $f_{prev} \leftarrow -\infty$, and $ROC = [\,]$. 4: **for** $i = 1$ to $|S_{sorted}|$ **do**
5: **if** $f(i) \neq f_{prev}$ **then**
6: $ROC(i) \leftarrow (\frac{FP}{N}, \frac{TP}{P})$, $f_{prev} \leftarrow f(i)$
7: **end if**
8: **if** $S_{sorted}(i)$ is a positive sample **then**
9: $TP \leftarrow TP + 1$.
10: **else**
11: $FP \leftarrow FP + 1$.
12: **end if**
13: **end for**
14: $ROC(i) \leftarrow (\frac{FP}{N}, \frac{TP}{P})$.

In multi-class classification problems, plotting ROC becomes much more complex than in binary classification problems. One of the well-known methods to handle this problem is to produce one ROC curve for each class. For plotting ROC of the class i (c_i), the samples from c_i represent positive samples and all the other samples are negative samples.

ROC curves are robust against any changes to class distributions. Hence, if the ratio of positive to negative samples changes in a test set, the ROC curve will not change. In other words, ROC curves are insensitive with the imbalanced data. This is because ROC depends on *TPR* and *FPR*, and each of them is a columnar ratio[3].

The following example compares between the ROC using balanced and imbalanced data. Assume the data is balanced and it consists of two classes each has 1000 samples. The point (0.2,0.5) on the ROC curve means that the classifier obtained 50% sensitivity (500 positive samples are correctly classified from 1000 positive samples) and 80% specificity (800 negative samples are correctly classified from 1000 negative samples). If the class distribution changed to be imbalanced and the first and second classes have 1000 and

10,000 samples, respectively. Hence, the same point (0.2, 0.5) means that the classifier obtained 50% sensitivity (500 positive samples are correctly classified from 1000 positive samples) and 80% specificity (8000 negative samples are correctly classified from 1000 negative samples). The AUC[4] score for both cases are the same while the other metrics which are sensitive to the imbalanced data will be changed. For example, the accuracy rates of the classifier using the balanced and imbalanced data are 65 and 77.3%, respectively, and the precision values will be 0.71 and 0.20, respectively. These results reflect how the precision and accuracy metrics are sensitive to the imbalanced data as mentioned in Section 2.1.

It is worth mentioning that the comparison between different classifiers using ROC is valid only when (1) there is only single dataset, (2) there are multiple datasets with the same data size and the same positive:negative ratio.

4. Area under the ROC curve (AUC)

Comparing different classifiers in the ROC curve is not easy. This is because there is no scalar value represents the expected performance. Therefore, the Area under the ROC curve (AUC) metric is used to calculate the area under the ROC curve. The AUC score is always bounded between zero and one, and there is no realistic classifier has an AUC lower than 0.5 [4,15].

Figure 9 shows the AUC value of two classifiers, A and B. As shown, the AUC of B classifier is greater than A; hence, it achieves better performance. Moreover, the gray shaded area is common in both classifiers, while the red shaded area represents the area where the B classifier outperforms the A classifier. It is possible for a lower AUC classifier to outperform a higher AUC classifier in a specific region. For example, in Figure 9, the classifier B outperforms A except at $FPR > 0.6$ where A has a slight difference (blue shaded area). However, two classifiers with two different ROC curves may have the same AUC score.

The AUC value is calculated as in Algorithm 2. As shown, the steps in Algorithm 2 represent a slight modification from Algorithm 1. In other words, instead of generating ROC points in Algorithm 1, Algorithm 2 adds areas of trapezoids[5] of the ROC curve [4]. As shown in Algorithm 2 the AUC score can be calculated by adding the areas of trapezoids of the AUC

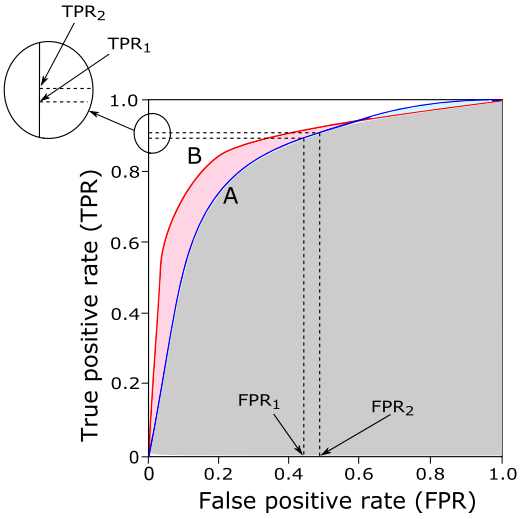

Figure 9.
An illustrative example of the AUC metric.

measure. Figure 9 shows an example of one trapezoid; the base of this trapezoid is $(FPR_2 - FPR_1)$, and the height of the trapezoid is $(TPR_1 + TPR_2)/2$; hence, the total area of this trapezoid is calculated as follows, $A = \text{Base} \times \text{Height} = (FPR_2 - FPR_1) \times (TPR_1 + TPR_2)/2$.

Algorithm 2: Calculating the AUC measure.

1: The same first two steps in Algorithm 1.
2: $FP \leftarrow 0$, $TP \leftarrow 0$, $f_{prev} \leftarrow -\infty$, $FP_{prev} \leftarrow 0$, $TP_{prev} \leftarrow 0$, and $A \leftarrow 0$, where A is the area under the ROC curve, i.e., *AUC* score. 3: for $i = 1$ to $|S_{sorted}|$ **do**
4: **if** $f(i) \neq f_{prev}$ **then**
5: $A \leftarrow A + \text{Trapezoid\_Area}(FP, FP_{prev}, TP, TP_{prev})$.
6: $f_{prev} \leftarrow f(i)$, $FP_{prev} \leftarrow FP$, $TP_{prev} \leftarrow TP$
7: **end if**
8: **if** $S_{sorted}(i)$ is a positive sample **then**
9: $TP \leftarrow TP + 1$.
10: **else**
11: $FP \leftarrow FP + 1$.
12: **end if**
13: **end for**
14: $A \leftarrow (A + \text{Trapezoid\_Area}(FP, FP_{prev}, TP, TP_{prev}))/(P \times N)$.
15: **function** $\text{Trapezoid\_Area}(X_1, X_2, Y_1, Y_2)$
16: $Base \rightarrow |X_1 - X_2|$, $Height \rightarrow (Y_1 + Y_2)/2$
17: **return** $Base \times Height$.

The AUC can be also calculated under the PR curve using the trapezoidal rule as in the ROC curve, and the AUC score of the perfect classifier in PR curves is one as in ROC curves.

In multi-class classification problems, Provost and Domingos calculated the total AUC of all classes by generating a ROC curve for each class and calculate the AUC value for each ROC curve [10]. The total AUC (AUC_{total}) is the summation of all AUC scores weighted by the prior probability of each class as follows, $AUC_{total} = \sum_{c_i \in C} AUC(c_i).p(c_i)$, where $AUC(c_i)$ is the AUC under the ROC curve of the class c_i, C is a set of classes, and $p(c_i)$ is the prior probability of c_i [10]. This method of calculating the AUC score is simple and fast but it is sensitive to class distributions and error costs.

5. Precision-Recall (PR) curve

Precision and recall metrics are widely used for evaluating the classification performance. The Precision-Recall (PR) curve has the same concept of the ROC curve, and it can be generated by changing the threshold as in ROC. However, the ROC curve shows the relation between sensitivity/recall (*TPR*) and 1-specificity (*FPR*) while the PR curve shows the relationship between recall and precision. Thus, in the PR curve, the *x*-axis is the recall and the *y*-axis is the precision, i.e., the *x*-axis of ROC curve is the *y*-axis of PR curve [8]. Hence, in the PR curve, there is no need for the *TN* value.

In the PR curve, the precision value for the first point is undefined because the number of positive predictions is zero, i.e., $TP = 0$ and $FP = 0$. This problem can be solved by estimating the first point in the PR curve from the second point. There are two cases for estimating the first point depending on the value of *TP* of the second point.

1. The number of true positives of the second point is zero: In this case, since the second point is (0,0), the first point is also (0,0).

2. The number of true positives of the second point is not zero: this is similar to our example where the second point is (0.1, 1.0). The first point can be estimated by drawing a horizontal line from the second point to the *y*-axis. Thus, the first point is estimated as (0.0, 1.0).

As shown in Figure 10, the PR curve is often zigzag curve; hence, PR curves tend to cross each other much more frequently than ROC curves. In the PR curve, a curve above the other has a better classification performance. The perfect classification performance in the PR curve is represented in Figure 10 by a green curve. As shown, this curve starts from the (0,1) horizontally to (1,1) and then vertically to (1,0), where (0,1) represents a classifier that achieves 100% precision and 0% recall, (1,1) represents a classifier that obtains 100% precision and sensitivity and this is the ideal point in the PR curve, and (1,0) indicates the classifier obtains 100% sensitivity and 0% precision. Hence, we can say that the closer the PR curve is to the upper right corner, the better the classification performance is. Since the PR curve depends only on the precision and recall measures, it ignores the performance of correctly handling negative examples (*TN*) [16].

Eq. (18) indicates the nonlinear interpolation of the PR curve that was introduced by Davis and Goadrich [5].

$$y = \frac{TP_A + x}{TP_A + x + FP_A + \frac{FP_B - FP_A}{TP_B - TP_A}.x} \tag{18}$$

where TP_A and TP_B represent the true positives of the first and second points, respectively, FP_A and FP_B represent the false positives of the first and second points, respectively, y is the precision of the new point, and x is the recall of the new point. The value of x can be any value between zero and $|TP_B - TP_A|$. A smooth curve can be obtained by calculating many intermediate points between two points A and B. In our example in Figure 10, assume the first point is the fifth point and the second point is the sixth point (see Table 1). From Table 1, the point A is (0.3,0.75) and the point B is (0.4,0.8). The value of $|TP_B - TP_A| = |4 - 3| = 1$ and hence the value of x can be any value between zero and one. Let $x = 0.5$, which is the middle

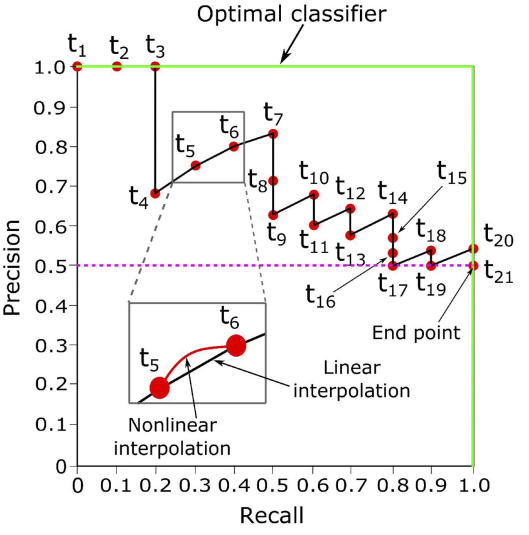

Figure 10.
An illustrative example of the PR curve. The values of precision and recall of each point/ threshold are calculated in Table 1.

point between A and B and hence the recall for the new point is $\frac{0.3+0.4}{2} = 0.35$. The precision of the new point is calculated as follows, $y = \frac{3+x}{3+x+1+\frac{1-1}{4-3}\times x} = \frac{3+0.5}{3+0.5+1+0} \approx 0.778$, where the new point using the linear interpolation is $(\frac{0.3+0.4}{2}, \frac{0.75+0.8}{2}) = (0.35, 0.775)$. In our example, for simplicity, we used the linear interpolation.

The end point in the PR curve is calculated as follows, $\left(1, \frac{P}{P+N}\right)$. This is because (1) the recall increases by increasing the threshold value and at the end point the recall reaches to the maximum recall, (2) increasing the threshold value increases both *TP* and *FP*. Therefore, if the data are balanced, the precision of the end point is $\frac{P}{P+N} = \frac{1}{2}$. The horizontal line which passes through $\frac{P}{P+N}$ represents a classifier with the random performance level. This line separates the area of the PR curve into (1) the area above the line and this is the area of good performance and (2) the area below the line and this is the area of poor performance (see Figure 10). Thus, the ratio of positives and negatives defines the baseline. Hence, changing the ratio between the positive and negative classes changes that line and hence changes the classification performance.

As indicated in Eq. (6), according to the precision metric, lowering the threshold value increases the *TP* or *FP*. Increasing *TP* increases the precision while increasing the *FP* decreases the precision. Hence, lowering the threshold value fluctuates the precision. On the other hand, as indicated in Eq. (2), lowering the threshold may leave the recall value unchanged or increase it. Due to the precision axis in the PR curve; hence, the PR curve is sensitive to the imbalanced data. In other words, the PR curves and their AUC values are different between balanced and imbalanced data.

6. Biometrics measures

Biometrics matching is slightly different than the other classification problems and hence it is sometimes called two-instance problem. In this problem, instead of classifying one sample into one of *c* groups or classes, biometric determines if the two samples are in the same group. This can be achieved by identifying an unknown sample by matching it with all the other known samples. This step generates a score or similarity distance between the unknown sample and the other samples. The model assigns the unknown sample to the person which has the most similar score. If this level of similarity is not reached, the sample is rejected. In other words, if the similarity score exceeds a pre-defined threshold; hence, the corresponding sample is said to be matched; otherwise, the sample is not matched. Theoretically, scores of clients (persons known by the biometric system) should always be higher than the scores of imposters (persons who are not known by the system). In biometric systems, a single threshold separates the two groups of scores; thus, it can be utilized for differentiating between clients and imposters. In real applications, for many reasons sometimes imposter samples generate scores higher than the scores of some client samples. Accordingly, it is a fact that however the classification threshold is perfectly chosen, some classification errors occur. For example, given a high threshold; hence, the imposters' scores will not exceed this limit. As a result, no imposters are incorrectly accepted by the model. On the contrary, some clients are falsely rejected (see Figure 11 (top panel)). In opposition to this, lowering the threshold value accepts all clients and also some imposters are falsely accepted.

Two of the most commonly used measures in biometrics are the *False acceptance rate* (*FAR*) and *False rejection/recognition rate* (*FRR*). The *FAR* is also called *false match rate* (*FMR*) and it is the ratio between the number of false acceptance to the total number of imposters attempts. Hence, it measures the likelihood that the biometric model will incorrectly accept an access by an imposter or an unauthorized user. Hence, to prevent imposter samples from being easily correctly identified by the model, the similarity score has to exceed a certain level (see Figure 11) [2]. The *FRR* or *false non-match rate* (*F NMR*)

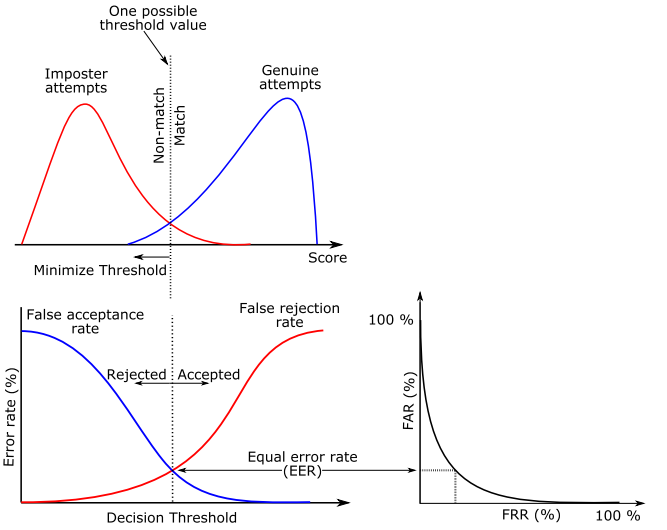

Figure 11.
Illustrative example to test the influence of changing the threshold value on the values of *FAR*, *FRR*, and EER.

measures the likelihood that the biometric model will incorrectly reject a client, and it represents the ratio between the number of false recognitions to the total number of clients' attempts [2]. For example, if FAR = 10% this means that for one hundred attempts to access the system by imposters, only ten will be succeeded and hence increasing *FAR* decreases the accuracy of the model. On the other hand, with FRR = 10%, ten authorized persons will be rejected from 100 attempts and hence reducing *FRR* will help to avoid a high number of trails of authorized clients. As a consequence, *FAR* and *FRR* in biometrics are similar to false positive rate (*FPR*) and false negative rate (*FNR*), respectively (see Section 2.4). Equal error rate (*EER*) measure solves the problem of selecting a threshold value partially, and it

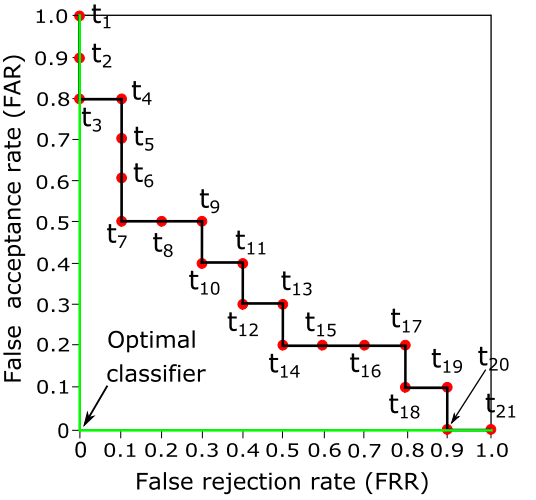

Figure 12.
An illustrative example of the DET curve. The values of *FRR* and *FAR* of each point/threshold are calculated in Table 1.

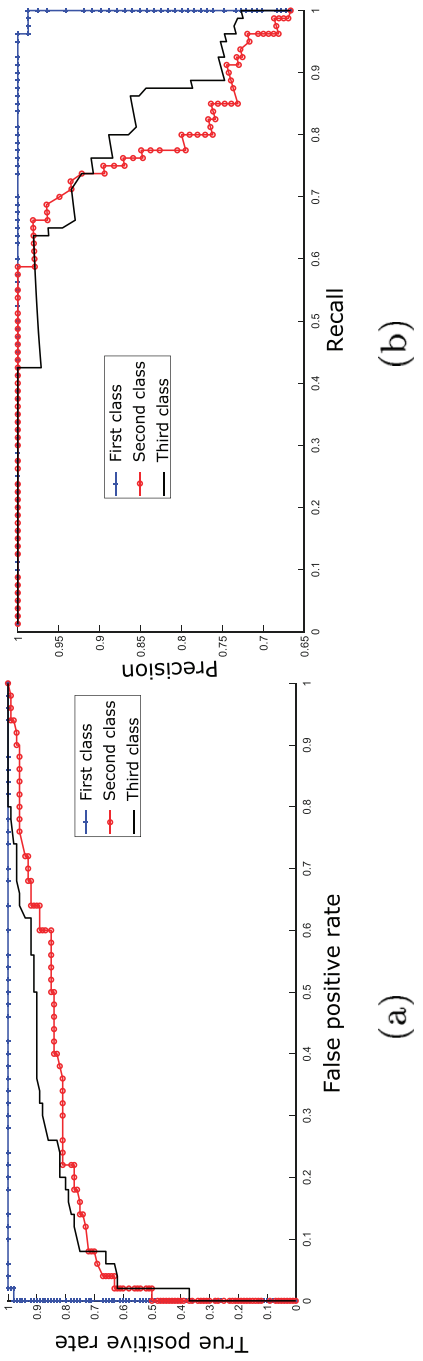

Figure 13.
Results of our experiment. (a) ROC curve, (b) Precision-Recall curve.

represents the failure rate when the values of *FMR* and *F NMR* are equal. Figure 11 shows the *FAR* and *FRR* curves and also the *EER* measure.

Detection Error Trade-off (DET) curve is used for evaluating biometric models. In this curve, as in the ROC and PR curves, the threshold value is changed and the values of *FAR* and *FRR* are calculated at each threshold. Hence, this curve shows the relation between *FAR* and *FRR*. Figure 12 shows an example of the DET curve. As shown, as in the ROC curve, the DET curve is plotted by changing the threshold on the confidence score; thus, each threshold generates only one point in the DET curve. The ideal point in this curve is the origin point where the values of both *FRR* and *FAR* are zeros and hence the perfect classification performance in the DET curve is represented in Figure 12 by a green curve. As shown, this curve starts from the point (0,1) vertically to (0,0) and then horizontally to (1,0), where (1) the point (0,1) represents a classifier that achieves 100% *FAR* and 0% *FRR*, (2) the point (0,0) represents a classifier that obtains 0% *FAR* and *FRR*, and (3) the point (1,0) represents a classifier that indicates 0% *FAR* and 100% *FRR*. Thus, we can say that the closer a DET curve is to the lower left corner, the better the classification performance is.

7. Experimental results

In this section, an experiment was conducted to evaluate the classification performance using different assessment methods. In this experiment, we used Iris dataset which is one of the standard classification datasets and it is obtained from the University of California at Irvin (UCI) Machine Learning Repository [1]. This dataset has three classes, each class has 50 samples, and each sample is represented by four features. We used (1) the Principal component analysis (PCA) [23] for reducing the features to two features and (2) Support vector machine (SVM)[6] for classification.

In our experiment, we used different assessment methods for evaluating the learning model. Figure 13 shows the ROC and Precision-Recall curves. As shown, there are three curves, one curve for each class and as shown, the first class obtained results better than the other two classes. Figure 14 shows the confusion matrix for each class. From these confusion matrices we can calculate different metrics as mentioned before (see Figure 3). For example, the results of the first class were as follows, *Acc*, *TPR*, *TNR*, *PPV*, and *NPV* were 99.33, 100, 98.0, 99.01, 100, respectively. Similarly, the results of the other two classes can be calculated.

8. Conclusions

In this paper, the definition, mathematics, and visualizations of the most well-known classification assessment methods were presented and explained. The paper aimed to give a detailed overview of the classification assessment measures. Moreover, based on the confusion matrix, different measures are introduced with detailed explanations. The relations between these measures and the robustness of each of them against imbalanced data are also introduced. Additionally, an illustrative numerical example was used for explaining how to calculate different classification measures with binary and multi-class problems and also to show the robustness of different measures against the imbalanced data. Graphical measures such as ROC, PR, and DET curves are also presented with illustrative examples and visualizations. Finally, various classification measures for evaluating biometric models are also presented.

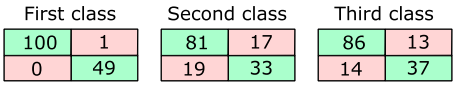

Figure 14.
Confusion matrices of the three classes in our experiments.

Notes

[1] More details about these two metrics are in Sections 2.2 and 2.5.

[2] More details about these two metrics are in Section 2.8.

[3] As mentioned before $TPR = \frac{TP}{TP+FN} = \frac{TP}{P}$ and both TP and FN are in the same column, and similarly FNR.

[4] The AUC metric will be explained in Section 4.

[5] A trapezoid is a 4-sided shape with two parallel sides.

[6] More details about SVM can be found in [24].

References

[1] C. Blake, Uci repository of machine learning databases, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html.

[2] R.M. Bolle, J.H. Connell, S. Pankanti, N.K. Ratha, A.W. Senior, Guide to biometrics, Springer Science & Business Media, 2013.

[3] S. Boughorbel, F. Jarray, M. El-Anbari, Optimal classifier for imbalanced data using matthews correlation coefficient metric, PLoS One 12 (6) (2017) e0177678.

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

[5] J. Davis, M. Goadrich, The relationship between precision-recall and roc curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 233–240.

[6] J.J. Deeks, D.G. Altman, Diagnostic tests 4: likelihood ratios, Brit. Med. J. 329 (7458) (2004) 168–169.

[7] R.O. Duda, P.E. Hart, D.G. Stork, et al., Pattern Classification, vol. 2, Wiley, New York, 2001. second ed.

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

[9] V. Garcia, R.A. Mollineda, J.S. Sanchez, Theoretical analysis of a performance measure for imbalanced data, in: 20th International Conference on Pattern Recognition (ICPR), IEEE, 2010, pp. 617–620.

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

[11] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowledge Data Eng. 21 (9) (2009) 1263–1284.

[12] V. López, A. Fernández, S. Garća, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.

[13] A. Maratea, A. Petrosino, M. Manzo, Adjusted f-measure and kernel scaling for imbalanced data learning, Inf. Sci. 257 (2014) 331–341.

[14] B.W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochim. Biophys. Acta 405 (2) (1975) 442–451.

[15] C.E. Metz, Basic principles of roc analysis, in: Seminars in nuclear medicine, vol. 8, Elsevier, 1978, pp. 283–298.

[16] D.M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation 2 (1) (2011) 37–63.

[17] T. Saito, M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, PLoS One 10 (3) (2015) e0118432.

[18] A. Shaffi, Measures derived from a 2 x 2 table for an accuracy of a diagnostic test, J. Biometr. Biostat. 2 (2011) 1–4.

[19] S. Shaikh, Measures derived from a 2 x 2 table for an accuracy of a diagnostic test, J. Biometr. Biostat. 2 (2011) 128.

[20] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, in: Australasian Joint Conference on Artificial Intelligence, Springer, 2006, pp. 1015–1021.

[21] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Inf. Process. Manage. 45 (4) (2009) 427–437.

[22] A. Srinivasan, Note on the location of optimal classifiers in n-dimensional roc space. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory, Oxford, England, 1999.

[23] A. Tharwat, Principal component analysis-a tutorial, Int. J. Appl. Pattern Recogn. 3 (3) (2016) 197–240.

[24] A. Tharwat, A.E. Hassanien, Chaotic antlion algorithm for parameter optimization of support vector machine, Appl. Intelligence 48 (3) (2018) 670–686.

[25] A. Tharwat, Y.S. Moemen, A.E. Hassanien, Classification of toxicity effects of biotransformed hepatic drugs using whale optimized support vector machines, J. Biomed. Inf. 68 (2017) 132–149.

[26] K.H. Zou, Receiver operating characteristic (roc) literature research, 2002. On-line bibliography available from: http://splweb.bwh.harvard.edu8000.

Corresponding author
Alaa Tharwat can be contacted at: aothman@fb2.fra-uas.de