analytics

MDPI

Editorial

Beyond the ROC Curve: The IMCP Curve

Jesus S. Aguilar-Ruiz

School of Engineering, Pablo de Olavide University, ES-41013 Seville, Spain; aguilar@upo.es

The ROC curve [1–4] is currently one of the most widely used methods for evaluating the quality of classification models in the literature. Its major competitive advantage over many other metrics—accuracy, F_1-measure, Matthews correlation coefficient (MCC), etc.—is that it provides a curve of classification performance. The curve—understood as a sequence of points that form it—is always richer and more informative than a simple scalar value, as it allows for more varied interpretation. However, it also has a significant limitation: it is not applicable to datasets that contain more than two classes, meaning its use is restricted to binary classification problems. This limitation does not affect other performance measures, such as accuracy or even the MCC multiclass extension, known as R_k [5].

In 2023, Chicco & Jurman [6] proposed retiring the ROC curve in favor of MCC (specifically for binary datasets). While there are mathematical justifications for choosing a specific measure over the ROC curve, the role it has played—and continues to play—in the field of machine learning is undeniable. This is especially true because the most popular metric, accuracy, suffers from significant biases that could lead to overestimation, particularly in the presence of class imbalance.

Several efforts have been made to explain and extend the ROC curve for multiclass contexts [7,8], but they have not gained widespread acceptance in the scientific community. In any case, the interpretation of the ROC curve—or its possible extensions—starts from an inherent deficiency: half of the unit square where it is represented is uninformative. The area of interest is confined to the triangle formed by the points (0, 0), (0, 1), and (1, 1). Since the diagonal line connecting the points (0, 0) and (1, 1) indicates randomness, anything below it is not informative.

It is noteworthy that all measures derive from a common element: the confusion matrix. However, the confusion matrix is a binary simplification of the probabilities of assignment to a particular class. For instance, if we have two classes {A, B}, and the classifier outputs probabilities {0.8, 0.2} for a test instance of class A, we will record a 1 in the confusion matrix (cumulatively) in the TP (True Positives) cell. This approach does not account for whether the probability was 0.8, 0.7, or any other value above 0.5, in which cases we would still accumulate a 1. Thus, the confusion matrix omits information (the probabilities of class assignment) that could be very informative for understanding the classifier's performance.

For a binary case with a dataset comprising two classes {A, B} with n instances (n_A, n_B), the number of different confusion matrices rises to $T = (n_A + 1)(n_B + 1)$. For example, the popular Titanic dataset, with 891 instances (549 deceased and 342 survived), could generate a total of 188,650 possible confusion matrices. However, if we consider the probabilities, that number would rise to infinity, thereby increasing both its richness and complexity. Moreover, all these confusion matrices were obtained by transforming probabilities to binary values, generating a large loss of information that is transferred to all the metrics that use the elements of the confusion matrix in their mathematical formulation. For instance, for a binary classification task that provides a confusion matrix with TP = 100, we know that all the 100 values come from a probability higher than 0.5, but in the worst case they could be slightly greater than 0.5, which would give a large weight to False Positives

check for updates

Citation: Aguilar-Ruiz, J.S. Beyond the ROC Curve: The IMCP Curve. *Analytics* **2024**, *3*, 221–224. https://doi.org/10.3390/analytics3020012

Received: 20 May 2024
Accepted: 23 May 2024
Published: 27 May 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

(FP) (even if in the confusion matrix FP = 0), providing an unreliable predictive model with high values for accuracy or MCC.

In any case, when the datasets are multiclass (more than two classes), the graphical representation of classifier performance has suffered from a significant gap and has been replaced by scalar measures. This problem is especially notable in genomic datasets in the context of tumor prediction, where the need is most compelling. However, in 2022, a measure called MCP (Multiclass Classification Performance) [9] was introduced, which is not based directly on the confusion matrix but on the probabilities of class assignment. The most interesting aspect of this measure is that it is independent of the number of classes in the dataset and its informative capacity extends over the entire unit square, unlike the ROC curve. While the ROC curve indicates the False Positive Rate (FPR) on the X-axis and the True Positive Rate (TPR) on the Y-axis, in the MCP curve, all instances of the dataset are distributed in the range [0, 1] on the X-axis, and the Y-axis represents the distance between the predicted class and the true class in terms of probability.

To eliminate the impact of class imbalance in classification, the MCP curve has been subsequently extended into the so-called IMCP (Imbalanced Multiclass Classification Performance) curve [10]. The result can be observed in Figure 1, where the left side shows the ROC curve for the Titanic dataset, and the right side shows the IMCP curve (see Supplementary Materials for downloading the IMCP Python package). These results were obtained using Random Forest with 10-fold cross-validation. The ROC curve tends to be more optimistic than the IMCP curve.

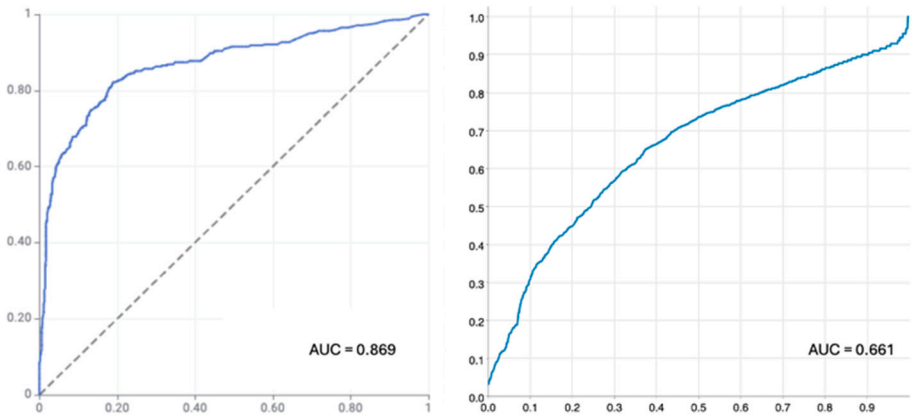


Figure 1. Titanic dataset, with two classes. Classification results from 10-fold cross-validation with Random Forest. **Left**: the ROC curve; **right**: the IMCP curve.

The advantage of the IMCP curve over MCP lies in the fact that the instances are uniformly distributed on the X-axis with respect to the classes. Therefore, all classes have equal opportunities to cover the area of the unit square.

The IMCP curve allows us to delve deeper into the behavior of each class by analyzing the probabilities independently. In this way, we can observe for which classes the classifier is performing better. In the case of the Titanic dataset, as shown in Figure 2, the classifier clearly performs better for the "deceased" class (green), with a median around 0.79, compared to 0.63 for the "survived" class (red). However, it is also evident that the classifier exhibits irregular performance, as indicated by the whiskers of the box plot for both classes.

In Figure 3, an example with a multiclass dataset (the popular Iris dataset, with three classes) is shown. On the left is the IMCP curve, indicating excellent performance of Random Forest (10-fold cross-validation). On the right, the unequal performance of the classifier for the three classes is evident, although with a very high median (between 0.94 and 1). This situation is common in multiclass problems and often, irregular classifier performance can be critical, especially in medical contexts. The classifier might show good overall performance, but upon examining the individual classes, one might have poor performance. If it concerns tumor prediction, we cannot rely on scalar measures alone. We

must caution that the classifier performs well for some classes but be wary of its predictions for other classes.

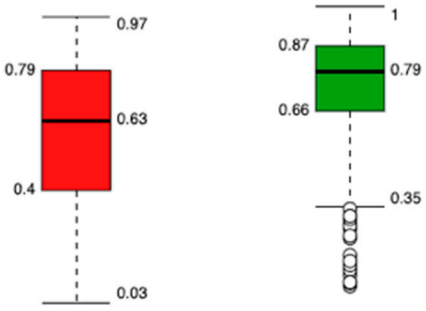


Figure 2. Titanic dataset. Box-plot from values of the IMCP for each class: survived (**left**, red) and deceased (**right**, green).

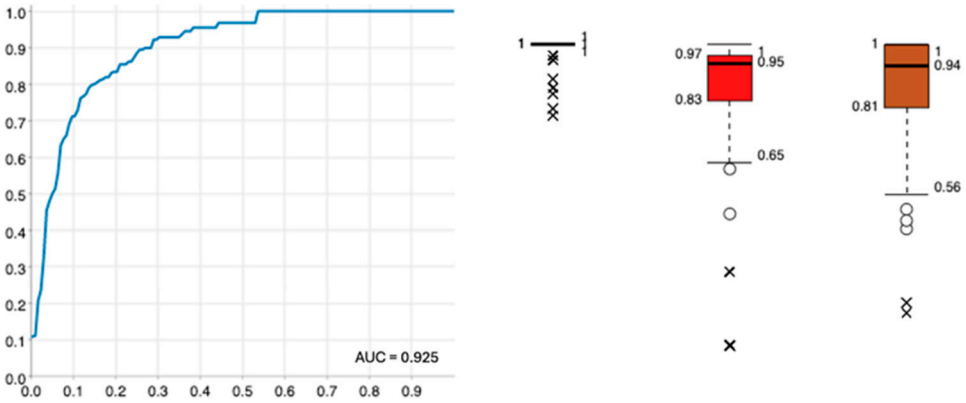


Figure 3. Iris dataset, with three classes. **Left**: the IMCP curve; **right**: conditional box-plot from the IMCP values for each class.

The ROC curve is old, but not as old as the mathematical foundation that led Matthews to define a correlation coefficient (MCC) between the prediction and observation of protein secondary structure [11]. However, replacing a method that provides both a curve and a scalar (area under the curve) with one that only provides a scalar is a step backward. A step forward would have been to define a curve whose area is equivalent to the MCC. Furthermore, it should be extendable to any number of classes, both for the curve and for the scalar (it already exists: R_k). This is precisely the purpose of the IMCP curve: regardless of the number of classes, the IMCP curve allows for the analysis of classifier performance, while the scalar (area under the IMCP curve) provides a comprehensive quantification.

In summary, the ability to graphically display classification performance has been predominantly represented by the ROC curve for binary datasets for many years. The IMCP curve emerges as a promising method for illustrating classification quality in multi-class contexts.

Supplementary Materials: The Python package to generate the IMCP curve is publicly available at https://github.com/adaa-polsl/imcp, accessed on 19 May 2024.

Funding: This work was supported by Grant PID2020-117759GB-I00 funded by MCIN/AEI/10.13039/501100011033, the Andalusian Plan for Research, Development, and Innovation.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Fox, W.; Peterson, W.; Birdsall, T. The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory* **1954**, *4*, 171–212.
2. Swets, J.A. The Relative Operating Characteristic in psychology. *Science* **1973**, *182*, 990–1000. [CrossRef] [PubMed]

3. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [CrossRef] [PubMed]
4. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
5. Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367–374. [CrossRef] [PubMed]
6. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. [CrossRef]
7. Hernández-Orallo, J.; Flach, P.; Ferri, C. A unified view of performance metrics: Translating threshold choice into expected classification loss. *J. Mach. Learn. Res.* **2012**, *13*, 2813–2869.
8. Hand, D.J.; Till, R.J. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach. Learn.* **2001**, *45*, 171–186. [CrossRef]
9. Aguilar-Ruiz, J.S.; Michalak, M. Multiclass Classification Performance Curve. *IEEE Access* **2022**, *10*, 68915–68921. [CrossRef]
10. Aguilar-Ruiz, J.S.; Michalak, M. Classification performance assessment for imbalanced multiclass data. *Sci. Rep.* **2024**, *14*, 10759. [CrossRef] [PubMed]
11. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) Protein Struct.* **1975**, *405*, 442–451. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.