IMCP curve Python package — the mathematical backgrounds

1 Introduction

Classifier performance measures are essential for evaluating the effectiveness and reliability of machine learning models. These measures provide objective insights into how well a classifier is performing and enable comparisons between different models or configurations of the same model. Performance measures provide a quantitative assessment of how well a classifier performs in making predictions. Accuracy is a commonly used measure that indicates the proportion of correct predictions. However, accuracy alone may not be sufficient, especially in imbalanced datasets, where the number of instances of one class is significantly higher than the others. Measures like precision, recall, and F1 score offer a more nuanced evaluation by considering true positives, false positives, and false negatives.

Some performance measures provide insights into the prediction confidence or uncertainty of a classifier. ROC curves illustrate the trade–off between true positive rate (sensitivity) and false positive rate ($1 - \text{specificity}$) across various decision thresholds. By examining these curves, decision–makers can understand how the performance varies based on different thresholds and make informed decisions on the appropriate trade–off between true and false positives or negatives.

ROC curves are primarily designed for binary classification tasks. While they can be extended to multiclass classification by using one–vs–all or one–vs–one approaches, interpreting and comparing ROC curves for multiclass problems can be more complex. Specialized evaluation measures, such as multiclass AUC or macro/micro-average precision and recall, may provide more meaningful insights for multiclass scenarios. In addition, ROC curves treat each class equally and do not take into account the underlying class distribution, which can be problematic with dealing with imbalanced datasets. The curve may appear to show good performance even if the classifier struggles to correctly identify the minority class (which might have higher importance).

These two important issues (not designed for multiclass datasets and insensitive to class distribution) have been addressed in the next contributions: the MCP (Multiclass Classification Performance) curve and the IMCP (Imbalanced Multiclass Classification Performance) curve. Overall, the MCP curve provides a visual representation of the classifier performance across multiple classes, enabling a more comprehensive evaluation of the classifier effectiveness in multiclass scenarios. The IMCP curve extends the MCP properties to deal with imbalance in multiclass datasets (very common in the biomedical context).

The MCP curve is a highly intuitive method to visualize the classification performance of one or more classifiers applied to a dataset [1]. Unlike many approaches, such as the ROC curve, it can handle multiclass datasets, i.e.,

datasets with any number of class labels. Furthermore, it is not based on the confusion matrix but rather on the prediction probabilities generated by the classifier. The method relies on the Hellinger distance [3, 4] to compare the test–object classifier output with the ground true one–hot vector.

2 Methodology

With such a tool it becomes possible to provide a more consistent way of comparing classifier outputs with the object class, including the margin of certainty. Let us consider a sample object from three–class (A, B, and C) data that belongs to class A, and two different classifier outputs for that object. The one–hot (ground truth) vector would then look as follows: $[1.00, 0.00, 0.00]$. Two properly assigned classifier outputs could look like the following two: $[1.00, 0.00, 0.00]$ and $[0.40, 0.30, 0.30]$. In both cases, if we assign the object to the class with the highest probability coming from the classifier output we would get a correct result (the highest probability is for class A). However, we should notice that the difference between the ground truth distribution and the classifier distributions is significantly different. The level of difference becomes more significant if we consider the following classifier output: $[0.34, 0.33, 0.33]$. The Hellinger distances for all three cases are presented in Table 1.

Ground truth	Original class	Classifier output	Predicted class	HD
$[1.00, 0.00, 0.00]$	A	$[1.00, 0.00, 0.00]$	A	0.00
$[1.00, 0.00, 0.00]$	A	$[0.40, 0.30, 0.30]$	A	0.61
$[1.00, 0.00, 0.00]$	A	$[0.34, 0.33, 0.33]$	A	0.64

Table 1: Comparison of Hellinger distance between three correctly classified objects, however based on completely different classifier outputs.

The Hellinger distance helps also to distinguish between completely misclassified objects and objects with a small margin of misclassification. Let us consider the same three–class problem: there are three classifier responses for the ground truth vector $[0.00, 1.00, 0.00]$, as presented in Table 2. Despite the same misclassification, the second classifier output seems much closer to the origin distribution than the first classifier output.

Ground truth	Original class	Classifier output	Predicted class	HD
$[0.00, 1.00, 0.00]$	B	$[1.00, 0.00, 0.00]$	A	1.00
$[0.00, 1.00, 0.00]$	B	$[0.40, 0.30, 0.30]$	A	0.67

Table 2: Comparison of Hellinger distance between two incorrectly classified objects.

Intuitively, it is possible to sort all test points according to the $1 - HD$ value. Such a defined curve is called a Multiclass data Classifier Performance curve. A more detailed description of the mathematical background as well as some properties can be found in [1].

The methodology described above was developed to present a graphical measure of classification performance for multiclass datasets, without taking into account the possible data imbalance. When imbalance is present in data, the distribution of classes has a great impact on the performance measures, as the number of instances of the majority class might be much higher than that of the minority class. In order to mitigate this difference, the IMCP curve assigns the same relevance to each class by rescaling the IMCP curve X axis, in such a way that the sum of widths of each class is equal for all the classes. Overrepresented classes will have the same weight in the IMCP curve as underrepresented classes.

The difference between the MCP and the IMCP curves is illustrated by means of the following example: let us consider an eight-sample dataset containing three classes (one instance of class A, three instances of class B, and four instances of class C). Let us also assume that a classifier provided the output leading to the following $1 - HD$ values, as presented in Table 3 (2^{nd} column). 3^{rd} column — the same bar width for all objects — refers to MCP approach – while in the 4^{th} column the bar width varies in accordance to the class imbalance (the sum for each class is equal — 1/3 — and the class width is equally distributed for each object belonging to this class).

class	$1 - HD$	MCP bar width	IMCP bar width
A	0.10	1/9	1/3
B	0.15	1/9	1/9
B	0.25	1/9	1/9
B	0.33	1/9	1/9
C	0.55	1/9	1/12
C	0.88	1/9	1/12
C	0.95	1/9	1/12
C	0.99	1/9	1/12

Table 3: $1-HD$ values for classification results and corresponding bar widths for the MCP curve (not considering imbalance) and for the IMCP curve (considering imbalance).

This example shows the situation when the minority class A has a relatively small level of confidence, while the majority class objects are correctly classified. That should imply that the MCP and IMCP curves should differ significantly. Both of them are presented in Fig. 1. The blue curve represents the MCP curve while the red one represents the IMCP curve.

The main reason for such a difference comes directly from bar widths presented in Table 3 — interior points (i.e. excluding $X = 0$ and $X = 1$) of each curve are midpoints of the corresponding upper bar. A more detailed representation of both curves — with bars — is shown in Fig. 2. Mathematical calculations of the IMCP curve points can be found in [2].

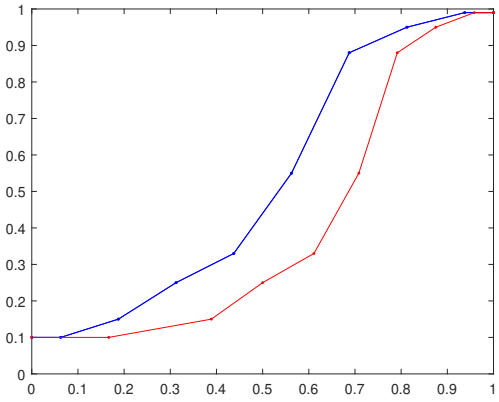

Figure 1: Comparison between the MCP (blue) and the IMCP (red) curves for the same imbalanced and multiclass data.

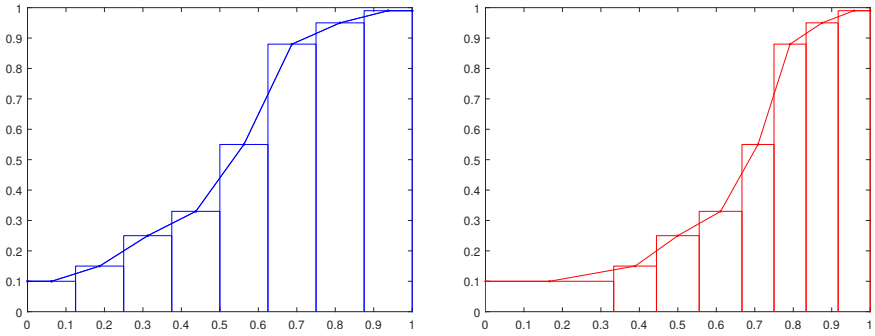

Figure 2: Constant widths of the MCP curve bars (left) and class distribution dependent widths of the IMCP curve bars (right).

References

[1] J. S. Aguilar-Ruiz and M. Michalak. Multiclass classification performance curve. *IEEE Access*, 10:68915–68921, 2022.

[2] J. S. Aguilar-Ruiz and M. Michalak. Classification performance assessment for imbalanced multiclass data. *Scientific Reports*, 14:10759, 2024.

[3] E. Hellinger. *Die Orthogonalvarianten quadratischer Formen von unendlich vielen Variablelen*. PhD thesis, University of Göttingen, 1907.

[4] E. Hellinger. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.