

FACTORS INFLUENCING CLASSIFICATION PERFORMANCE AND THE RELIABILITY OF PERFORMANCE SCORES

Attila Fazekas

*University of Debrecen
4032 Debrecen, Egyetem square 1*

Abstract

In the rapidly expanding world of artificial intelligence, where intelligent systems increasingly permeate our daily lives, classification stands as one of the most fundamental, yet most consequential tasks. It concerns the assignment of samples to predefined classes based on the available measurements. Consider, for example, autonomous vehicles: the objects captured and processed by on-board cameras raise critical, even life-or-death questions—does the system detect a pedestrian, a cyclist, or a car? Medical applications bear similarly high stakes, where diagnostic tools must determine whether a disease is caused by malignant processes.

Formally, the problem can be described as follows. Let $X = \{x_i \in \mathbb{R}^d \mid i = 1, \dots, n\}$ be a set of n elements, where each element (d -dimensional sample) constitutes part of a dataset. In addition, let $Y = \{0, \dots, k\}$ denote the set of class labels. We assume that the class label of every sample is known, that is, for each $x_i \in X$ there exists a corresponding $y_i \in Y$. The goal of classification is to find a classifier $c: X \rightarrow Y$ that assigns to each x_i , its corresponding label y_i . However, a perfect classifier virtually never exists, as errors are inevitable. Consequently, the task becomes the search for a classifier that produces as few misclassifications as possible. To this end, we must first define precisely how to measure the "quality" of a classifier, and then assess the reliability of these performance scores. Finally, we must understand the factors that influence classifier performance. Such clarity is essential, as our trust in AI-based systems depends heavily on how well we understand the behaviour of fundamental classifiers and the reliability of the performance scores used to evaluate them.

In our presentation, we aim to demonstrate that, although performance scores appear to give a favourable description of classifier behaviour, they often rely on flawed or misleading calculations. To address this issue, we introduce a consistency test developed by our research group, together with a concrete case study illustrating both the presence of these problems and the effectiveness of our method. It is important to emphasize that, beyond reproducibility, the results themselves strongly depend on the datasets used in evaluation (and, in the case of supervised classification, in training as well). Thus, it is essential to examine which properties of a dataset—and in what manner modify the classification performance. Two particularly influential factors are label noise (incorrectly assigned class labels) and class imbalance (differences in class sample sizes), both of which frequently pose challenges for classification. In the second part of our presentation, we investigate the effect of these two factors in binary classification, demonstrating the substantial difficulties they can introduce.

Our conclusions are twofold. First, rigorous scrutiny of experimental results is indispensable. Second, the selection of an appropriate classifier is shaped by numerous dataset-specific parameters, among which noise and class imbalance are especially decisive. Consequently, it becomes clear that selecting suitable machine-learning-based solutions is a highly complex task, one that requires careful theoretical consideration and thorough experimental validation.