Neuroscience 2011, Washington DC, USA
12.11. – 16.11.2011

Choosing an Appropriate Performance Measure: Classification of EEG-Data with Varying Class Distribution

Sirko Straube, Jan Hendrik Metzen, Anett Seeland, Mario Krell, Elsa Kirchner

Workgroup Robotics, Faculty of Mathematics and Computer Science, University of Bremen,
Robert Hooke Str. 5, D-28359 Bremen, Germany

Robotics Innovation Center, DFKI GmbH,
Robert Hooke Str. 5, D-28359 Bremen, Germany

ABSTRACT

A popular method to judge the behavior of a system that has to make decisions is to come up with a measure of performance. Such a system could be a human, an animal, an artificial agent or a classification algorithm. In case of multiple decisions, e.g. between two classes, a straightforward measure is often the Accuracy (i.e., the rate of correct decisions). However, Accuracy is misleading when the true class distribution is unbalanced. This situation is not only common in natural environments, but also often intended experimentally (e.g. in oddball paradigms). Consider, e.g., a situation where we have 990 examples of class A and just 10 examples of class B. Even if we know nothing about how to separate the two classes, we are able to obtain an Accuracy of 99% if we always decide for class A. The problem gets even worse when we want to compare two situations with different class distributions (e.g. when distributions change over time). To deal with such difficulties a lot of measures exist, but none of them is "perfect", and so advantages and disadvantages remain.
In the present work, we investigate the differences and effects when using different performance measures commonly used in decision-making, signal detection and machine learning. We classify single-trial data from the electroencephalogram (EEG) of 5 subjects, which has been recorded in an oddball paradigm. Besides the fact that we have unbalanced classes (due to the nature of the oddball), we also vary the ratios between two experimental conditions: The classifier is trained on classical oddball data and then used in an application where the ratio between the occurence of class instances has strongly changed.
We compare the performance of the classifier (here, we use a support vector machine) using different performance measures all relying on the confusion matrix (consisting of numbers of True Positives, True Negatives, False Positives and False Negatives). These measures include Accuracy, Balanced Accuracy, F-measure, Area under ROC-Curve (AUC) and Mutual Information (MI). Since Accuracy, F-measure and MI are sensitive to the prior class distribution, while Balanced Accuracy and AUC are not, our results indicate that selecting an appropriate measure is important to avoid drawing misleading conclusions. Our results demonstrate the importance of choosing the correct performance measure in the light of the evidence one wants to give.

Keywords: performance, classification, decision making, oddball, EEG, brain reading, brain-computer interfaces, classification, P300