Evaluating Learning Classifier System Performance In Two-Choice Decision Tasks: An LCS Metric Toolkit

John H. Holmes

Center for Clinical Epidemiology and Biostatistics
University of Pennsylvania School of Medicine
Philadelphia, PA 19104
jholmes@cceb.med.upenn.edu

Abstract

A "metric toolkit" to evaluate learning classifier system performance is proposed. The metrics are shown to be superior to crude accuracy in evaluating classification performance, especially for data with unequal numbers of positive and negative cases. In addition, these metrics provide information to the researcher that is not available from crude accuracy. When used appropriately, these metrics provide accurate depictions of learning classifier system performance during training and testing in supervised learning environments.

1 INTRODUCTION

Applying a learning classifier system (LCS) to two-choice decision problems requires a special approach to performance evaluation. This paper presents a suite of quantitative tools that addresses the requirements of two-choice problems in supervised learning environments, especially where the *base rate* (proportions of the two classes is unequal. These metrics, borrowed from the domain of medical decision making, are proposed as adjuncts to commonly used evaluation methods such as crude accuracy ("percent correct").

2 METHODS

The components of the toolkit include: *Sensitivity* (true positive rate), which indicates a LCS's ability to classify correctly gold standard-positive cases; *Specificity* (true negative rate), which measures the ability of a LCS to classify correctly gold-standard-negative cases; and negative and positive *predictive value*, which measure the probability of a negative (positive) when a LCS detects a negative (positive); and, the area under the receiver-operating characteristic (ROC) curve, which indicates the tradeoff between sensitivity and specificity for a given test as a single measure. In addition, the indeterminant rate is presented as a metric for correcting these measures for cases that cannot be classified (correctly or erroneously). A stimulus-response LCS, EpiCS, was used as the testbed system. Four datasets consisting of 15 predictor variables and one outcome were created at four different positive base rates. Each dataset was partitioned into training and testing sets of equal size. EpiCS was trained over 30,000 iterations, and then tested once using the testing set; its learning performance during training and classification performance on testing were evaluated using the metric toolkit.

3 RESULTS

During training, progressive separation between the area under the curve and crude accuracy were seen with decreasing base rates. As the base rate decreased, crude accuracy overestimated the classification performance of EpiCS. At the lowest base rate, these two measures never converged during the training period.

On testing with novel data from the testing sets, EpiCS demonstrated the same divergence between crude accuracy and area under the ROC curve that was observed during training. In addition, as the base rate decreased, the sensitivity and positive predictive value decreased, indicating that at lower base rates, EpiCS would not be a good choice for classification, and that its decisions would be less accurate in detecting positives. This information is not available from crude accuracy.

4 CONCLUSIONS

This paper demonstrates the use of metrics that will be useful in evaluating learning classifier system performance in two-choice problems where the base rates are unequal. These metrics are more meaningful and robust indicators of performance than traditional measures such as crude accuracy, especially in forced two-choice decision tasks.