The Loss of Optimizing the Correct Metric

David B. Skalak
Highgate Predictions, LLC
Ithaca, NY 14850 USA
skalak@cs.cornell.edu

Alexandru Niculescu-Mizil
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
alexn@cs.cornell.edu

Rich Caruana
Department of Computer Science
Cornell University
Ithaca, NY 14853 USA
caruana@cs.cornell.edu

Most machine learning research has assumed that it is best to train and select a classifier according to the metric upon which it ultimately will be evaluated. There is however, little empirical evidence that this is the case. Is this assumption really justified? Moreover, what if we don't know the metric upon which the classifier will be judged? What if the classification objective is not optimal performance, but simply robust performance across several metrics? Does it make any difference how much data is available on which to base model performance estimates?

These questions are very relevant in a large number of applications. Classifiers that are deployed in the field can be used and evaluated in ways that were not anticipated when the model was trained. The ultimate evaluation metric may not have been known to the modeler at training time, additional performance criteria may have been added, the evaluation metric may have changed over time, or the real-world evaluation procedure may have been impossible to simulate.

Our objective is to test the well established conjecture that it is always best to optimize the "correct" metric. We also aim to provide experimental support for modelers who face potential "cross-metric" optimization problems. We tackle these issues in the case of model selection for eight widely used performance metrics.

The experimental results lead to a number of quite surprising conclusions. First, even if the true evaluation metric is known at training time, the common belief that it is always better to optimize the evaluation metric is not warranted in the case of model selection. In fact one can incur a significant loss by optimizing the correct metric, especially if the data is scarce.

Second, the most robust selection metric for scarce data regimes is by far cross entropy (log-loss). Regardless of the evaluation metric, if the validation set contains only a few hundred cases, it is always better to perform model selection using cross-entropy. This result allows us to make a powerful recommendation to machine learning practitioners that face cross-metric challenges: if the data is scarce do not worry about the final evaluation metric; just use cross-entropy for model selection. For larger validation sets squared error and area under the ROC curve also have a strong cross-metric performance.

Third, metrics such as lift, accuracy, F-score and precision-recall breakeven point should be avoided when performing model selection. These metrics proved to be the least robust, leading to poor cross-metric performance.

Topic: Classifier evaluation
Preference: oral/poster