Value of the Area Under the Receiver Operating Characteristic Curve for Validation of Prediction Models on Imbalanced Data Sets

Luis H. John, MSc[1], Jenna Reps, PhD[2], Peter R. Rijnbeek, PhD[1]
[1]Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Janssen Research and Development, Raritan, New Jersey, United States

Background

The area under the Receiver Operating Characteristic curve (AUROC) is a popular metric to report and compare model performance. However, AUROC is based on sensitivity and specificity and, thus, is not sensitive to class imbalance. For that reason, Jeni et al. suggest that AUROC may mask poor model performance[1]. In patient-level prediction, class imbalance is a common phenomenon, as a medical condition often shows low prevalence in a target population. The area under the Precision-Recall curve (AUPRC) has been proposed as a superior metric for imbalanced data, because it compares false-positives to true-positives rather than true-negatives (precision). Also, Saito and Rehmsmeier recommend precision-recall plots as the most informative visual tool, because they express the susceptibility of classifiers to imbalanced data sets with clear visual cues[2].

In this study we aim to compare model performance measured by AUROC and AUPRC on three datasets with different disease prevalence. Our problem is defined as predicting heart failure within one year in patients with type 2 diabetes. Furthermore, we perform subsampling to study the effect of sample size in both metrics.

Method

AUROC is a discrimination metric that measures the ability to distinguish between classes, based on specificity and sensitivity (Table 1). Measurements range from 0.5 to 1.0, where higher is better. AUPRC describes the ability to correctly detect a relevant sample by trading off precision and recall (Table 1). Measurements range from 0.0 to 1.0, where higher is better.

Measure	Formula	Explanation
Specificity	$\frac{TN}{TN+FP}$	Measures the proportion of negatives that are correctly identified as such
Sensitivity	$\frac{TP}{TP+FN}$	Measures the proportion of positives that are correctly identified as such
Recall		
Precision	$\frac{TP}{TP+FP}$	Measures the proportion of true-positives among all retrieved positives

Table 1. Definitions of Sensitivity/Recall, Specificity, and Precision, in terms of true-positive (TP) and true-negative (TN) predictions.

We train a Lasso Logistic Regression model to predict heart failure in patients with type 2 diabetes on data from three Truven Health MarketScan claims databases: Medicaid (MDCD), Medicare Supplemental (MDCR), and Commercial Claims And Encounters (CCAE). We choose Lasso Logistic Regression, because it is thought to be fairly robust to imbalanced data sets[3].

Database	Population count	Outcome count	Prevalence
MDCD	273183	20518	7.51%
MDCR	501806	38836	7.74%
CCAE	1788773	29805	1.67%

Table 2. Data set overview for predicting the outcome of heart failure in patients with type 2 diabetes

The Patient-Level Prediction package contains functionality for generating a learning curve, by training a hyper-parameter reduced model on successively larger subsets of the training data. We use learning curves to get an estimate of the learning rates and model saturation. All models have been trained with default parameters available in the patient-level prediction package.

Results

Model performance measured by AUROC and AUPRC are plotted in Figure 1 and Figure 2, respectively.

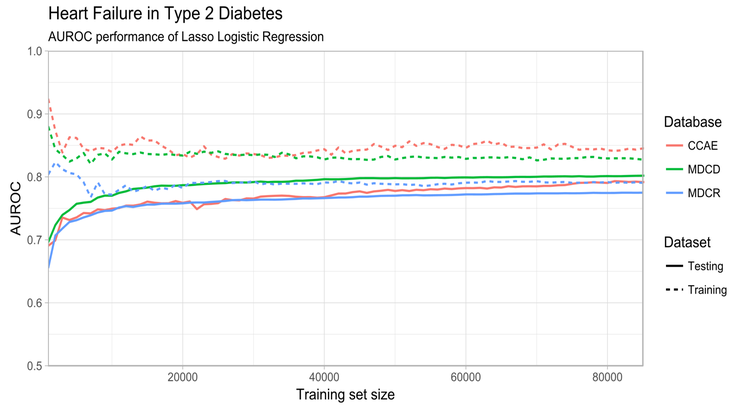

Figure 1. AUROC performance of prediction models.

All three models seem to perform well on the testing set considering the AUROC (0.75 - 0.80). Model performance can be ranked from best to worst: MDCD, CCAE, MDCR. This figure also shows that adding more data might improve performance a little further, as learning curves have not yet plateaued. However, a training set size of approximately 40000 seem to suffice in all three datasets to achieve satisfactory results.

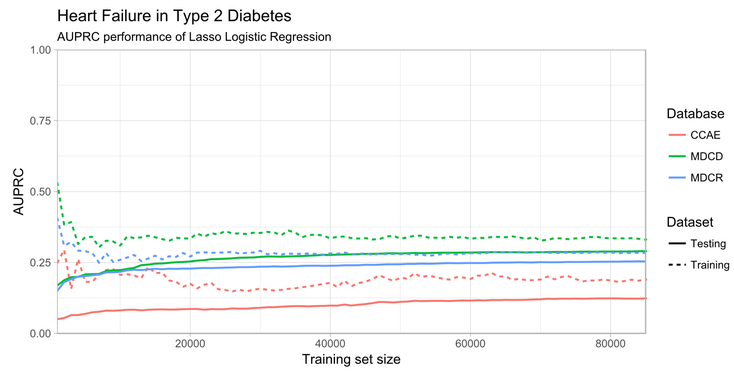

Figure 2. AUPRC performance of prediction models.

The same three models perform poorly in Figure 3, when we are comparing AUPRC, which suggests a high bias. In addition, the ranking has changed from best to worst: MDCD, MDCR, CCAE. The reason why CCAE performs poorly may be explained by the larger class imbalance, for which AUPRC is sensitive. As to be expected, the same effects of training set size as for AUROC are observed.

Conclusion

Our results suggest that it is useful to inspect AUPRC as an additional performance metric when assessing patient-level prediction models. In medical practice, discrimination may not be the ideal metric to optimize for, whereas a tradeoff between precision and recall appears practically relevant. Our preliminary results stimulate us to further investigate these metrics for patient-level prediction models at a larger scale.

1. Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013, September). Facing imbalanced data--recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*(pp. 245-251). IEEE.
2. Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced data sets. *PloS one*, *10*(3), e0118432.
3. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453.

Contact: l.john@erasmusmc.nl