Computing and Visualizing Precision-Recall Curves and Receiver Operating Characteristic Curves for Soft-labeled and Hard-labeled Data

Ivo Grosse[1,2], Jan Grau[1] and Jens Keilwagen[3]
[1] *Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany*
[2] *German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Germany*
[3] *Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany*
grosse@informatik.uni-halle.de

Introduction

The assessment of classifier performance is of fundamental importance in many bioinformatics applications. For instance, measures of classification performance are used to select appropriate models for solving classification problems. Performance evaluation is also inevitable for demonstrating the utility of novel approaches. In general, it assists researchers in identifying the most promising approach for the classification problem at hand. This implies that the choice of appropriate performance measures may influence the results of downstream analyses.

For binary classification tasks, the receiver operating characteristic (ROC) curve and the area under this curve (AUC-ROC) are widely accepted as a general measure of classifier performance. In many bioinformatics applications, however, positive examples are substantially less abundant than negative examples, resulting in a highly imbalanced class ratio. For instance, the number of true donor splice sites is substantially smaller than the number of genomic sequences with central GT consensus, and the number of target genes of a microRNA is substantially smaller than the number of non-target genes. In such cases, the precision-recall (PR) curve and the area under this curve (AUC-PR) is better suited for comparing the performance of individual classifiers than the ROC curve and AUC-ROC [DBR$^+$05].

Often, the decision for the true class labels of a given data point is ambiguous and partly subjective. For instance, class labels may be based on an arbitrary threshold for some continuous measurement, e.g., fold changes of differentially expressed genes. Uncertain class labels may also arise from multiple, possibly contradictory, expert labelings. However, the decision for a specific labeling decisively influences classifier training and assessment. One solution to this problem is the transition from hard-labeling to soft-labeling, where each data point is assigned to both classes with a certain probability that reflects confidence in the labeling. For instance, Grau *et al.* [GPGK13] develop a schema for deriving soft-labels from peak statistics for ChIP-seq data, or Mihaljevic *et al.* [M$^+$14] determine soft-labels from expert labelings of interneurons. While soft-labeling has been used extensively for classifier training in the past, it has been neglected for classifier assessment [KGG14].

Computing empirical AUC-PR and AUC-ROC values from test data points requires interpolation between discrete supporting points corresponding to a series of classification thresholds. AUC-ROC can be computed by linear interpolation between the supporting points of the curve for hard-labeled and soft-labeled data. In contrast, Davis & Goadrich [DG06] show that for AUC-PR an interpolation along the true positives is more accurate than linear interpolation for hard-labeled data, while Boyd *et al.* [BEP13] and Keilwagen *et al.* [KGG14] propose a more fine-grained, continuous interpolation between the supporting points of the PR curve. Only the latter can also be used for soft-labeled data and weighted data in general.

We make this interpolation available to the scientific community in the R package PRROC [GGK15], which is available from CRAN and may be used to compute and visualize PR and ROC curves.

Results

To illustrate the efficacy of the developed method, we investigate the influence of soft-labeled test data on classifier performance. To this end, we compare the classifier performance of published classifiers using AUC-PR on hard-labeled and soft-labeled test data for predicting transcription factor binding affinities.

We perform a reassessment of classifiers from Weirauch *et al.* [WCN+13], who evaluate the performance of classifiers for 66 protein binding microarray (PBM) data sets. PBMs measure the *in-vitro* binding affinity of transcription factors to DNA sequences using microarrays in an unbiased manner, where double-stranded probe sequences are chosen such that they contain all k-mers up to a given k with identical frequency. The goal of that study was to assess different classifiers for their ability to distinguish bound from unbound probes and for the correspondence of their classification scores to measured microarray intensity values.

Weirauch *et al.* introduce a hard labeling based on the intensity values for all probes sequences in each of the 66 experiments. For each individual experiment, they define the threshold separating foreground and background data points. Based on this labeling, they compare classifiers using different performance measures including the mean AUC-ROC over all experiments.

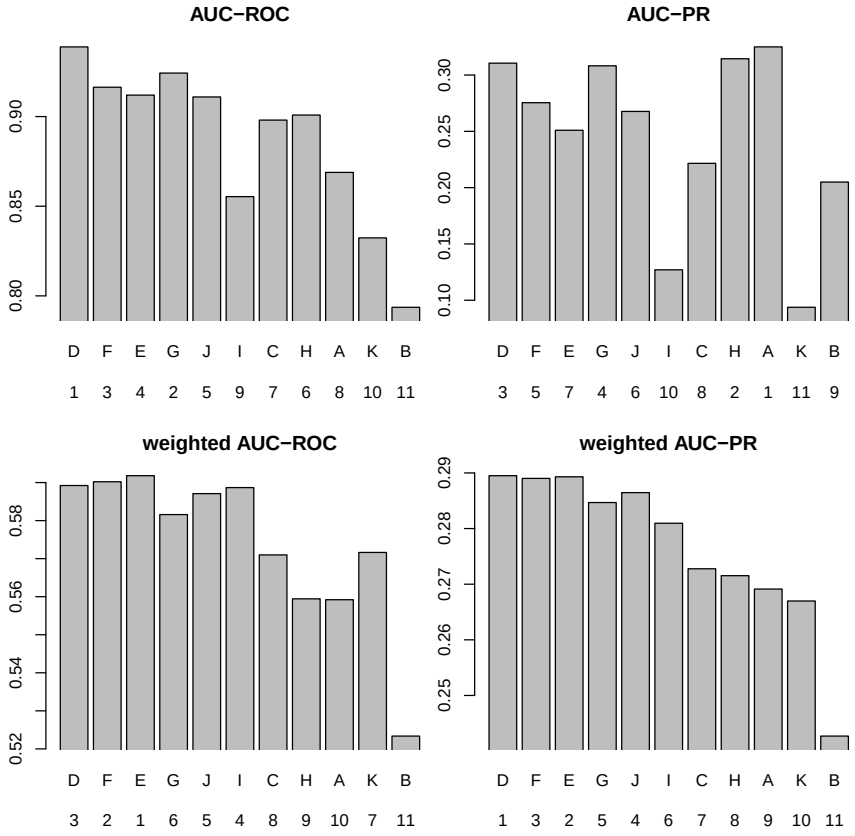

Figure 1: Mean results for AUC-ROC and AUC-PR on PBM data sets using hard-labeled or soft-labeled (i.e., weighted) test data. Letters (A,B,...,K) on the abscissa indicate the team names of approaches in the original publication of Weirauch *et al.* [WCN+13] and appear in the order of the original ranking. Rankings according to the different performance measures are shown below the team names, while the mean values for AUC-ROC and AUC-PR are depicted on the ordinate.

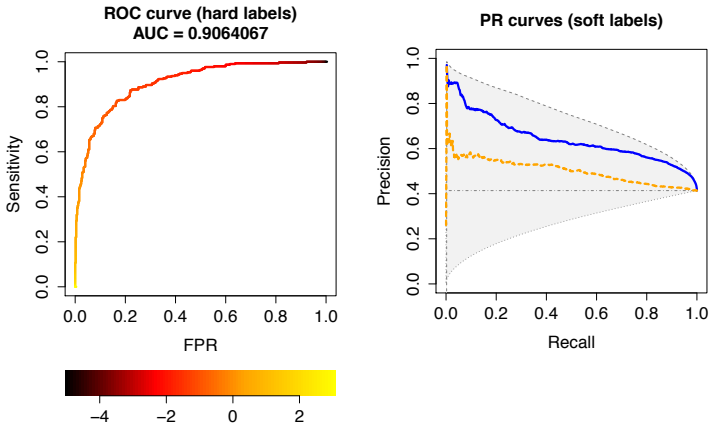

Figure 2: Plots of ROC (left) and PR (right) curves generated by PRROC. For the ROC curve, we consider hard-labeled data and show the plotting variant with a color scale that indicates classification thresholds yielding the points on the curve. For the PR curve, we consider soft-labeled data and show a comparative plot for two classifiers as solid blue and dashed orange lines. We also include the maximal and minimal possible curves and the curve of a random classifier for the given soft-labels.

In Figure 1, we compare the mean AUC-ROC, the mean AUC-PR, and the corresponding rankings for hard-labeled and soft-labeled test data. In the hard-labeled case, we take the class labels suggested by Weirauch *et al.* [WCN+13]. We find that the rankings for both mean AUC-ROC and mean AUC-PR change considerably when considering soft-labeled test data instead of less informative hard-labeled test data. Focusing on the mean AUC-PR, we find that the ranking obtained by AUC-PR using soft-labeled test data are in better accordance to the original ranking of Weirauch *et al.* than the ranking using hard-labeled test data.

PRROC R-package

We have developed a user-friendly and well-documented R package called PRROC [GGK15], which allows for computing PR and ROC curves as well as the areas under these curves for soft-labeled and hard-labeled data. Optionally, PRROC also computes curves and AUC values for the optimal, the worst, and the random classifier as a reference. These references are particularly useful for (i) PR curves and (ii) ROC and PR curves in case of soft-labeled data, where the minimum and maximum AUC may differ from 0 and 1, respectively. In addition, PRROC allows for visualizing PR and ROC curves as exemplarily shown in Figure 2. PRROC is available from CRAN (`http://cran.r-project.org/web/packages/PRROC/index.html`) and provides R documentation files and a vignette.

Talk outline

In the talk, we will first motivate why appropriate performance measures are important for classification problems in bioinformatics and why these should be chosen in a problem-specific manner. Second, we introduce AUC-PR as a useful performance measure for problems with highly imbalanced class ratios, which are prevalent in bioinformatics. Third, we will provide examples for bioinformatics applications that may profit from performance evaluation using soft-labels. Finally, we will show how researchers can use the PRROC R-package to evaluate classifier performance for soft-labeled and hard-labeled test data, and to produce publication-quality plots of PR and ROC curves using PRROC.

References

[BEP13] Kendrick Boyd, Kevin H. Eng, and C. David Page. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, volume 8190 of *LNCS*, pages 451–466. Springer Berlin Heidelberg, 2013.

[DBR+05] Jesse Davis, Elizabeth Burnside, Raghu Ramakrishnan, Vitor Santos Costa, and Jude Shavlik. View learning for statistical relational learning: With an application to mammography. In *Proceeding of the 19th International Joint Conference on Artificial Intelligence*, pages 677–683, 2005.

[DG06] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 233–240, New York, NY, USA, 2006. ACM.

[GGK15] Jan Grau, Ivo Grosse, and Jens Keilwagen. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*, 2015.

[GPGK13] Jan Grau, Stefan Posch, Ivo Grosse, and Jens Keilwagen. A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197, 2013.

[KGG14] Jens Keilwagen, Ivo Grosse, and Jan Grau. Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS ONE*, 9(3):e92209, 03 2014.

[M+14] Bojan Mihaljevic et al. Multi-dimensional classification of GABAergic interneurons with Bayesian network-modeled label uncertainty. *Frontiers in Computational Neuroscience*, 8(150), 2014.

[WCN+13] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, DREAM consortium, Harmen J. Bussemaker, Quaid D. Morris, Martha L. Bulyk, Gustavo Stolovitzky, and Timothy R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31:126–134, 2013.