Artificial Dilution Series for Comparison of Classifier Evaluation Metrics

Petri Törönen*[#], Ilja Pljusnin[#], Liisa Holm
Structural Genomics Group, Institute of Biotechnology
PO Box 56, 00014 University of Helsinki, Finland
[#]These authors contributed equally
*To whom correspondence should be addressed: petri.toronen@helsinki.fi

1. INTRODUCTION

The comparison of competing methods is central to bioinformatics research. Honest evaluation, however, requires benchmark data sets as well as appropriate evaluation metrics. In the evaluation of Gene Ontology (GO) predictions, there is no consensus as to what is the most appropriate metric and, therefore, the metrics used tend to vary between publications. The selection of evaluation metrics for GO prediction is a challenge: many have difficulties related to highly unbalanced classes and others fail to account for the complex correlation structure between GO classes. Furthermore, when evaluation metrics are misused, a reasonable set of GO predictions can be outperformed by simply ranking GO classes in order of size. This demonstrates the need to benchmark classifier evaluation metrics with respect to the features specific to the Gene Ontology.

2. METHODS

We have developed a framework for testing the performance of classifier evaluation metrics called Artificial Dilution Series (ADS). ADS takes a GO annotated set of proteins and generates artificial predictions sets with a defined level of noise by permuting labels in the original data set. These artificial predictions inherently respect the ontological structure of the GO and are propagated towards parent nodes in the usual manner. Next, the evaluation metric under test is applied to assess the quality of the artificial predictions. This procedure is repeated many times with different noise levels creating a series of diluted versions of the original data set. Finally, we assess the performance of different evaluation metrics by their ability to separate different noise levels from one another.

To complement ADS, we perform additional tests based on False Positive Data (FPD). Such data sets have all of the original signal removed. Instead, they represent corner cases that cause some evaluation metrics to give unreasonably high scores. These were created by, for example, annotating each protein with the largest GO classes. The resultant evaluation metric score for each FPD is then compared against the results from the dilution series, to see whether it is more similar to ADS results with a high or low noise level. A good evaluation metric would naturally match all FPD results next to high noise level results. This presents a worst case scenario test for evaluation metric and ensures that it is not biased towards some error signals.

3. RESULTS AND DISCUSSION

We tested several GO prediction evaluation metrics using ADS and FPD. Our results show dramatic differences between evaluation metrics and we replicate the weak performance of metrics known to perform badly with the GO structure. Furthermore, our framework for evaluation assessment allowed us to experiment with variations of popular evaluation metrics. For each tested metric, we show how each one can be improved, increasing performance significantly.

In conclusion, we present a simple framework to assess and improve classifier evaluation metrics. Researchers can apply ADS to their own data sets to identify the best performing evaluation metrics. Our ultimate goal is to develop standards for evaluation of GO prediction tools. Despite the focus on GO prediction, the ADS and FPD methods are very general and can therefore be applied to other classification evaluation problems. The ADS and FPD methods were implemented as a C++ programs with additional scripts written in Perl. All source code will be freely available upon publication.