Negative Example Selection in Protein Function Prediction

Noah Youngs*, Richard Bonneau, Dennis Shasha
Department of Computer Science, New York University, New York, USA
Center for Genomics and Systems Biology, Department of Biology, New York University, New York, USA

*To whom correspondence should be addressed: nyoungs@nyu.edu

1. INTRODUCTION

There has been a surge of interest lately in a subset of semi-supervised machine learning problems known as Positive-Unlabeled (PU) learning scenarios, in which the only known labels are of the positive class. This situation presents an obvious problem for the vast majority of machine learning techniques, which require examples of both the positive and negative class in order to train a predictor, and while some PU algorithms attempt to learn in this one-class scenario, the majority instead proceed by pre-classifying a set of reliable negative examples before applying a traditional machine learning classifier to the enriched data as usual. While the main focus of the literature has been on applying these algorithms in the context of text classification (3), where labeling documents is time-intensive and it is much easier to label a document's topic than all of the topics it does not contain, the analogies to protein function are obvious.

Choosing negative examples for protein function-prediction has been a little-studied problem, despite the recent influx of machine learning algorithms applied to the function prediction. The situation is quite similar to that of text classification: proteins are rarely labeled with the functions they do NOT possess, and proteins are nearly always multi-topic, in that the possession of one function does not exclude the potential for several other functional classifications. Therefore PU algorithms are highly applicable to the function prediction problem, and hold great potential for improvements in machine learning algorithms applied in this context. Indeed Youngs et al. (6) showed how more-reliable negative examples can boost the predictive power of function prediction algorithms. Our current work focuses more directly on the first step of the PU learning task, namely generating a reliable set of negative examples for protein function.

Past methodologies for choosing negative examples in protein function prediction have largely involved heuristics, including: designating all genes that don’t have a particular label as being negative for that label (2), randomly sampling genes and assuming the probability of getting a false negative is low (often done when predicting protein-protein interactions, as in (1)), and (iii) using genes with annotations in sibling categories of the category of interest as negative examples (4), although this final heuristic was later withdrawn. To these heuristics we add our new technique, based on a PU algorithm known as the “1-DNF” negative example selection. This technique operates in the context of text classification, by identifying terms which appear more frequently in the positive class than the unlabeled, and using as negatives all unlabeled documents that do not contain these words. We extend this idea to protein function by calculating the pairwise empirical conditional probability of a given GO function 'f' appearing, given the presence of all other GO functions across all three branches of GO. We then select as negatives for 'f' the genes with the lowest average probability of being annotated to 'f' in the future, given the the other annotations currently present for those genes.

Evaluating methods for predicting negative examples for protein function presents all the of the same problems and biases as evaluating function prediction algorithms, except in reverse. Where function prediction results are biased negatively by the fact that a positive prediction without a corresponding validation annotation might simply indicate lack of study rather than an incorrect prediction, negative example validations are biased positively by the same effect. Just because a gene is not annotated with the function in the validation set doesn't guarantee that it was correctly identified as a negative example. In order to attempt to rigorously evaluate potential negative example selection algorithms, we used the following criteria: validation data was obtained two years after training data, and methods were evaluated by the average number of mis-classified negative examples. For further stringency, a negative example is considered misclassified even if the validation annotation has an IEA evidence code. Lastly, since the trivial solution (predicting no negative examples at all) would perform the best under this evaluation, we present results in a two dimensional representation, where the x-axis is the number of negative examples chosen.

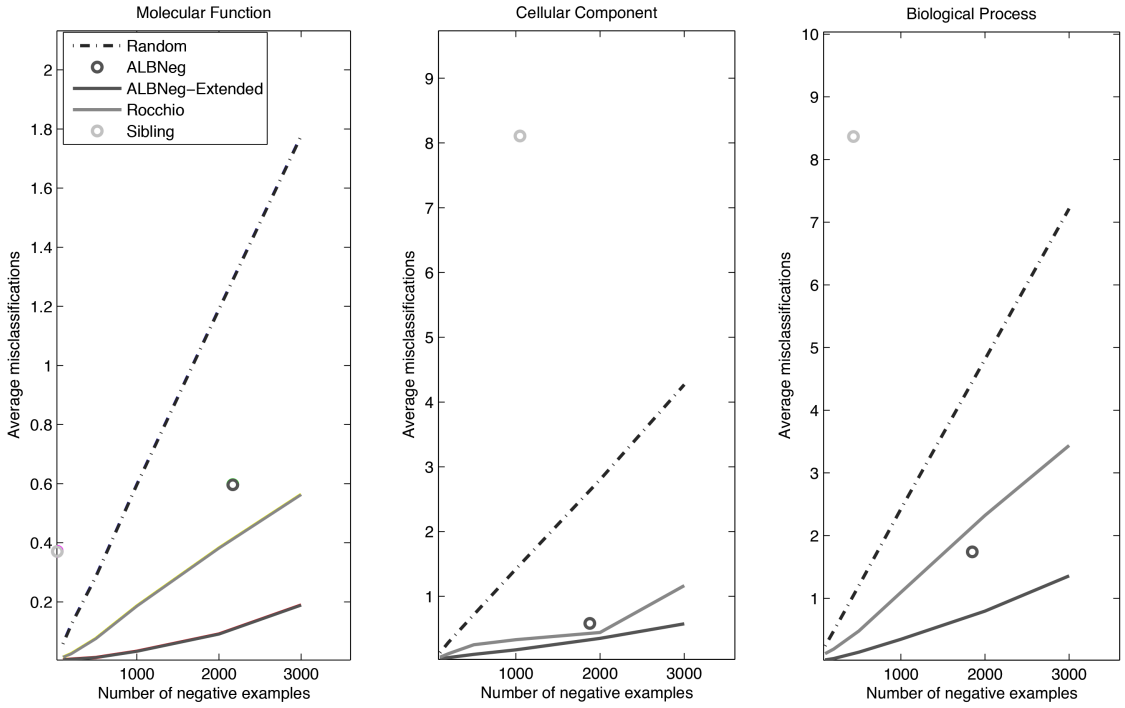
For comparison purposes we have included the heuristic methods previously mentioned (Random selection, and siblings), noting that the choice of all-non positive genes as negative is a special case of the

random selection where he sampling percentage is allowed to approach 100%. To these heuristics we add the negative example selection algorithm of Youngs et al, noted as “ALBNeg”, as well as another standard PU technique known as “Rocchio” selection (5), and our latest algorithm “ALBNeg_Extended”. All experiments were conducted in the human genome, with training annotations from June 2010, and validation annotations from June 2012.

Our results indicate that our negative example selection method has a significantly lower misclassification rate than all the alternative algorithms we tested.

2. FIGURES

Negative Example Algorithm Performance



Performance statistics for the trial negative example selection algorithms, averaged over all categories with between 3 and 300 annotations in the human genome. Results are presented for all three branches of GO.

3. REFERENCES

1. Gomez, S. M., *et al.* 2003. Learning to predict protein-protein interactions. *Bioinformatics:* Oxford University Press. **19,** 1875-1881.

2. Guan, Y., *et al.* 2008. Predicting gene function in a hierarchical context with an ensemble of classifiers. *Genome Biology:* Biomed Central Press. 9(Suppl. 1), S3.

3. Liu, B., Dai, Y., Li, X., Lee, W. S., & Yu, P. S. 2003. Building text classifiers using positive and unlabeled examples. *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on* (pp. 179-186). IEEE.

4. Mostafavi, S., and Morris, Q. 2009. Using the Gene Ontology hierarchy when predicting gene function. *UAI Conference Proceedings:* AUAI Press 2009**.**

5. Rocchio, J. (1971). Relevant feedback in information retrieval. In G. Salton (ed.). The smart retrieval system- experiments in automatic document processing, Englewood Cliffs, NJ.

6. Youngs, N., *et al.* 2013. Parametric Bayesian priors and better choice of negative examples improve protein function prediction. *Bioinformatics:* Oxford University Press. *(*Preprint)