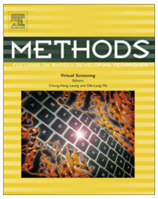
Methods xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



Predicting protein function and other biomedical characteristics with heterogeneous ensembles

Sean Whalen[a], Om Prakash Pandey[b], Gaurav Pandey[b,c,*]

[a] Gladstone Institutes, University of California, San Francisco, CA, USA
[b] Icahn Institute for Genomics and Multiscale Biology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[c] Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

ARTICLE INFO

Article history:
Received 21 May 2015
Received in revised form 3 August 2015
Accepted 23 August 2015
Available online xxxx

Keywords:
Protein function prediction
Heterogeneous ensembles
Nested cross-validation
Diversity-performance tradeoff
Ensemble calibration
Distributed machine learning

ABSTRACT

Prediction problems in biomedical sciences, including protein function prediction (PFP), are generally quite difficult. This is due in part to incomplete knowledge of the cellular phenomenon of interest, the appropriateness and data quality of the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor for specific problems. In such scenarios, a powerful approach to improving prediction performance is to construct heterogeneous ensemble predictors that combine the output of diverse individual predictors that capture complementary aspects of the problems and/or datasets. In this paper, we demonstrate the potential of such heterogeneous ensembles, derived from stacking and ensemble selection methods, for addressing PFP and other similar biomedical prediction problems. Deeper analysis of these results shows that the superior predictive ability of these methods, especially stacking, can be attributed to their attention to the following aspects of the ensemble learning process: (i) better balance of diversity and performance, (ii) more effective calibration of outputs and (iii) more robust incorporation of additional base predictors. Finally, to make the effective application of heterogeneous ensembles to large complex datasets (*big data*) feasible, we present *DataSink*, a distributed ensemble learning framework, and demonstrate its sound scalability using the examined datasets. DataSink is publicly available from https://github.com/shwhalen/datasink.

1. Introduction

Prediction problems in biomedical sciences, including protein function prediction (PFP) [41,52], are generally quite difficult. This is due in part to incomplete knowledge of how the cellular phenomenon of interest is influenced by the variables and measurements used for prediction, as well as a lack of consensus regarding the ideal predictor for specific problems. Even from a data perspective, the frequent presence of extreme class imbalance, missing values, heterogeneous data sources of different scales, overlapping feature distributions, and measurement noise further complicate prediction. Indeed, given these challenges, several community-wide exercises, most notably CAFA [47] and MouseFunc [44], have been organized to assess the state of the art in PFP. The approaches used to participate in these exercises use a variety of biological information, such as the amino acid sequence and three-dimensional structure of proteins, as well as systems-level data like gene expression and protein–protein interactions. They also use a diverse array of prediction methodologies from machine learning, statistics, network theory and others. Although these exercises indicated general principles that can help enhance PFP, they were generally unable to determine the *best method* for this problem out of the participating approaches, partly due to the problems listed above.

In scenarios like these, a powerful approach to improving prediction performance is to construct ensemble predictors that combine the output of individual predictors [49,51]. These predictors have been immensely successful in producing accurate predictions for many biomedical prediction tasks [62,1,33,28,42], including protein function prediction [57,63,19]. The success of these methods is attributed to their ability to reinforce accurate predictions as well as correct errors across many diverse base predictors [56]. Diversity among the base predictors is key to ensemble performance: if there is complete consensus (no diversity) the ensemble cannot outperform the best base predictor, yet an ensemble lacking any consensus (highest diversity) is unlikely to perform well due to weak base predictors. Successful ensemble methods strike a balance between the diversity and accuracy of the ensemble [29,14].

[*] Corresponding author at: Icahn Institute for Genomics and Multiscale Biology and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

E-mail addresses: shwhalen@gmail.com (S. Whalen), omprakash.pandey@mssm.edu (O.P. Pandey), gaurav.pandey@mssm.edu (G. Pandey).

http://dx.doi.org/10.1016/j.ymeth.2015.08.016
1046-2023/© 2015 Elsevier Inc. All rights reserved.

A wide variety of methods have been proposed to create ensembles consisting of diverse base predictors that benefit from both their consensus and disagreement [49,51]. Popular methods like bagging [6], boosting [50] and random forest [7] generate this diversity by sampling from or assigning weights to training examples. However, they generally utilize a single type of base predictor to build the ensemble. Such *homogeneous* ensembles may not be the best choice for biomedical problems like PFP where the ideal base prediction method is unclear, as discussed earlier. A more potent approach in this scenario is to build ensembles from the predictions of a wide variety of *heterogeneous* base prediction methods, such as the predictions from a variety of PFP methods submitted to CAFA [47].

Two commonly used heterogeneous ensemble methods include a form of meta-learning called *stacking* [35,61], and the *ensemble selection* method [10,9]. Stacking constructs a higher-level predictive model over the predictions of base predictors, while ensemble selection uses an iterative strategy to select base predictors for the ensemble while balancing diversity and performance. Due to their ability to utilize heterogeneous base predictors, these approaches have produced superior performance across several application domains [1,40].

In this paper, we present a comparative analysis of several heterogeneous ensemble methods applied to large and diverse sets of base predictors. This comparison is carried out in the context of the prediction of protein function, as well as that of genetic interactions [4,42], a problem intimately related to PFP. In addition to assessing the overall performance of these methods on these problems, we will investigate the following critical aspects of heterogeneous ensembles that have not been investigated before:

1. How these methods try to achieve the diversity-performance tradeoff inherent in ensemble learning.
2. The importance of base predictor calibration for effective ensemble performance.
3. Dependence of heterogeneous ensemble performance on the number and type of base predictors constituting them.

We explored the first two aspects in our previous work on this problem [60]. Other than this work, there have been few, if any, such analyses of ensembles constructed from such a large set of diverse base predictors, and we expect our results will shed light on the inner dynamics of ensemble predictors, especially heterogeneous ones. These insights are expected to have wide applicability across diverse applications of ensemble learning beyond PFP.

Finally, we present DataSink, a scalable distributed implementation of the heterogeneous ensemble methods analyzed in this study (available at https://github.com/shwhalen/datasink). This implementation is built on the insight that efficient (nested) cross-validation is critical for robust ensemble learning, and utilizes several parallelization opportunities in this process to enhance scalability. DataSink is implemented in Python and uses Weka [20] for base predictor learning (unless they are provided, such as in CAFA), and the pandas/scikit-learn analytics stack [43,34] for ensemble construction. We illustrate how DataSink can be used for PFP and other similar prediction problems, and through this illustration, we demonstrate its scalability capabilities and other salient features.

2. Materials and methods

2.1. Problem definitions and datasets

To assess the potential of heterogeneous ensembles for enhancing PFP, we assess their performance on three PFP instances. We also evaluate their performance on the closely related problem of prediction of genetic interactions.

2.1.1. Protein function prediction

Gene expression data are a commonly used data source for predicting protein function, as the simultaneous measurement of gene expression across the entire genome enables effective inference of functional relationships and annotations [41,52]. Thus, for the PFP assessment, we use the gene expression compendium of [26] to predict the functions of roughly 4000 baker's yeast (*Saccharomyces cerevisiae*) genes. The three most abundant functional labels (GO terms) from the list of most biologically interesting and actionable Gene Ontology Biological Process terms compiled by [38] are used in our evaluation. These labels are GO:0051252 (regulation of RNA metabolic process), GO:0006366 (transcription from RNA polymerase II promoter) and GO:0016192 (vesicle-mediated transport). We refer to these prediction problems as PF1, PF2 and PF3 respectively (details in Table 1). Note that we demonstrate on only these three large labels due to the substantial amount of computation needed even for a single label (detailed in subsequent sections), and hope to report much more extensive results in future work. We expect the results presented here to be representative and the methodology to be applicable to other (functional) labels as well.

2.1.2. Genetic interaction prediction

Genetic interactions (GIs) are a category of cellular interactions that are inferred by comparing the effect of the simultaneous knockout of two genes with that of knocking them out individually [21]. Since these interactions represent cases of functional buffering and inter-connections, their knowledge is very useful for understanding cellular pathways and their interactions, and predicting gene function [23,4,27,37,36]. However, despite their utility, and substantial progress in experimental mapping of GIs in model organisms [23,12], there is a general paucity of GI data for several organisms important for biomedical research. To address this problem, some of us [42] and several others [4] have developed various computational approaches to predict GIs. In the current study, we wish to assess how heterogeneous ensembles can help advance this GI prediction effort. For this, we use the dataset developed in our previous work [42], which focuses on predicting GIs between genes from *S. cerevisiae* (baker's yeast) using features that denote functional relationships between gene pairs. Some such features include correlation between expression profiles, extent of co-evolution, and the presence or absence of physical interactions between their corresponding proteins (see Table 2 for an illustration). This particular experiment enables us to assess the predictive ability of heterogeneous ensembles on a problem that is related but complementary to PFP, and the scalability of our implementation DataSink, given the much larger size of this dataset (Table 1).

Finally, note that both the types of datasets we considered involve binary labels, so the results presented are in the classification context. However, the concepts and methods discussed apply to the general *prediction* scenario, such as multi-class and regres-

Table 1
Details of genetic interaction (GI) and protein function (PF) datasets, including the number of features, number of examples in the minority (positive) and majority (negative) classes, and total number of examples.

Problem	#Features	#Positives	#Negatives	#Total
GI	152	9994	125,509	135,503
PF1	300	382	3597	3979
PF2	300	344	3635	3979
PF3	300	327	3652	3979