BITS '09 Sixth Annual Meeting of the Bioinformatics Italian Society

Genome-wide hierarchical classification of gene function.

Nicolò Cesa-Bianchi, Giorgio Valentini

DSI, Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano

Extended abstract

Motivation.

Gene function prediction, a central problem of computational biology, has three distinguishing features. First, the number of functional classes is large: in the hundreds for the Functional Catalogue (FunCat) taxonomy and in the thousands for the more refined Gene Ontology (GO). Moreover, each gene or gene product may belong to more than one class, and class frequencies are typically largely unbalanced.

Second, classes have complex relationships: FunCat is structured as a tree forest and GO as a directed acyclic graph. As a consequence, graph-structured gene function annotations exist at different level of resolution, depending both on biological characteristics and on the information available for the gene under investigation.

Third, multiple and heterogeneous sources of bio-molecular data are available, where each source captures different and complementary characteristics of the genes: we may thus expect that integrating these different sources can significantly improve the accuracy of predictions.

Although several works pointed out the central role played by the integration of heterogeneous biomolecular data for gene function prediction, less efforts have been devoted to the development of methods for the hierarchical classification of genes. This contribution focuses on this latter issue, taking explicitely into account the relationships between gene functional classes and the unbalance between the positive and negative examples in each class.

Methods.

We propose a cost-sensitive version of the hierarchical top-down classification algorithm, a cost-sensitive Bayesian bottom-up method, and a hierarchical algorithm based on the "true path rule" mutuated from the GO. In the training phase, all the proposed hierarchical algorithms learn a base classifier for each class node of the hierarchy. These classifiers discriminate between genes annotated or not to a specific class node.

The three algorithms differ in the evaluation phase. The classical hierarchical top-down algorithm classifies an unknown gene starting from the roots of the hierarchy: if a gene is predicted to belong to a class node, then the children nodes are evaluated; otherwise, the evaluation process stops at that node.

In the cost-sensitive version, the threshold used to discriminate between positive and negative examples is optimized to improve the precision/recall performance of the classifier.

The Bayesian bottom-up hierarchical algorithm starts from the leaf nodes at the bottom of the hierarchy and implements an approximation of the Bayes-optimal hierarchical classifier with respect to the hierarchical loss (a loss function specific to hierarchical multilabel classification problems) and a simple stochastic model for the labels. Uncertain predictions of the children influence the decision of the parent node, propagating information from bottom to top of the hierarchy. In the cost-sensitive version we add a balance factor to vary the cost of positive and negative predictions.

In the "true path rule" hierarchical algorithm, a two-way asymmetric flow of information traverses the tree-structured ensemble: positive predictions for a node influence in a recursive way its ancestors, while negative predictions influence its offsprings. This general behaviour is a consequence of the "true path rule". According to this rule, if an example belongs to a class, it belongs to all its ancestors, and if does not belong to a class it does not belong to all its offsprings.

Results.

The three proposed hierarchical algorithms, and the "flat" multiclass multilabel algorithm (used as baseline method) are applied to the classification of yeast genes based on the FunCat taxonomy. For comparing performances we used a hierarchical F-measure, which takes into account both the unbalance between positive and negative examples and the hierarchical nature of the prediction problem. Hierarchical methods are seen to largely outperform the basic "flat" approach, but there is no clear winner among the proposed algorithms.

Hierarchical F-measure results are comparable among the different hierarchical methods: their F-measure varies between 0.30 and 0.45, depending on the data set used. But for specific subsets of classes we obtain markedly higher F-measures and precisions. In order to significantly improve prediction performance, we conjecture that class-specific model selection and data fusion techniques are needed (in our experiments we applied a very limited model selection and used only single sources of biomolecular data). With these enhancements, hierarchical methods can really become reliable tools for the "in silico" prediction of gene function.