A cost-sensitive neural algorithm to predict gene functions using large biological networks.

Marco Frasca, Alberto Bertoni,Giorgio Valentini
DSI – Dept. of Computer Science – Università degli Studi di Milano, Italy

Biological networks can represent different types of relationships between biomolecular entities (e.g. genes or proteins), ranging from genetic or physical interactions, to gene expression correlations, chemical reactions, or co-occurences in bio-medical literature.
In this context, a central problem is the integration of different networks and the development of algorithms to infer the underlying biological properties of the biological entities, i.e. the functional classes of genes, or the potential protein targets of a drug, with relevant applications in functional genomics, proteomics and pharmacogenomics.
In particular, the gene function prediction problem can be formalized as a semi-supervised multi-class, multi-label classification problem where the unknown labels of the unlabeled part of the network can be predicted by exploiting the known labels of the labeled part and the relationships connecting the nodes of the network.
Several approaches have been proposed to address this problem, including simple "guilt-by-association" methods (Marcotte et al. 1999), "label propagation" algorithms (Zhou et al. 2003), Markov (Deng et al., 2004) and Gaussian Random fields (Tsuda et al. 2005, Mostafavi et al. 2008).
Unfortunately none of these methods has been specifically designed to manage the unbalance which often characterizes gene functional classes, with negative examples that largely outnumber positives. Moreover, most of these methods do not preserve the prior knowledge coded in the labeling of genes.
To address these items, we propose a Hopfield-based cost sensitive neural algorithm which preserves the prior information and introduces an efficient cost-sensitive strategy to learn the appropriate parameters of the network (neuron states and their thresholds) in order to manage the unbalance between positive and negative examples in functional classes. Our method factorizes the solution of the problem in two parts: 1) the sub-network composed of the labelled vertices is considered, and the network parameters are estimated through an efficient supervised algorithm; 2) the estimated parameters are extended to the subnetwork composed of the unlabeled vertices, and the attractor reached by the dynamics of this subnetwork allows to predict the labeling of the unlabeled vertices. Moreover our method allows to efficiently integrate multiple sources of data and significantly reduces the computational complexity by restricting the network dynamics to the unlabeled part of the network. The algorithm is fast, scales nicely when new sources of data are added, and can be efficiently applied to large biological networks.
We tested our method on the yeast and mouse model organisms at genome-wide level, using both the FunCat and Gene Ontology taxonomies, integrating different data sources, including protein domain, gene expression, and protein interactions.
Cross-validated results show that our integrated approach achieves competitive results with state-of-the-art semi-supervised and supervised methods on the MouseFunc benchmark data (Pena-Castillo et al., 2008). Moreover our cost-sensitive approach allows to significantly outperform state-of-the-art hierarchical ensemble methods (Cesa-Bianchi et al, 2010) using multiple sources of data and the whole FunCat taxonomy with the yeast model organism.