Regularized network-based algorithm for predicting gene functions with high-imbalanced data

M. Frasca✉, A. Bertoni, G. Valentini
Department of Computer Science, University of Milan, Italy

Motivations

The gene function prediction problem is a real-world problem consisting in finding new bio-molecular functions of genes/gene products and characterized by hundreds or thousands of functional classes structured according to a predefined hierarchy.

This problem can be formalized as a semi-supervised multi-class, multi-label classification problem where the biological functions of new genes can be predicted by exploiting their connections with genes whose biological functions are known.

Many different approaches have been proposed to address this problem, including "guilt- by-association" [1], "label propagation" [2], module-assisted techniques [3], SVMs [4]. Nevertheless, these methods usually suffer a decay in performance when input data are highly unbalanced, that is positive examples are sig-

F-score			
Dataset	Level 4	Level 5	Method
Expr	0,052	0,033	COSNet
	0,095	0,071	R-COSNet
	0,043	0,017	LP-Zhu
	0,032	0,015	SVM
PPI-BG	0,363	0,281	COSNet
	0,370	0,297	R-COSNet
	0,292	0,268	LP-Zhu
	0,132	0,130	SVM
Pfam	0,349	0,258	COSNet
	0,350	0,283	R-COSNet
	0,268	0,212	LP-Zhu
	0,051	0,028	SVM
Sim-SW	0,323	0,265	COSNet
	0,333	0,279	R-COSNet
	0,254	0,239	LP-Zhu
	0,050	0,023	SVM

Figure 1. Average F-score across levels 4 and 5 of the FunCat classes using four data sources of S.cerevisiae organism: Pfam (protein domain data obtained from the Pfam data base), Expr (gene expression data from Spellman and Gasch experiments), PPI-BG (protein-protein interaction data obtained from the BioGRID databases), and Sim-SW (sequence similarities obtained through Smith Waterman algorithm).

nificantly less than negatives. This scenario characterizes in particular the most specific classes of the ontology, which are the classes more far from the root classes and that better describe the functions of genes.

Methods

To address these items, we propose a regularization of a Hopfield-based cost- sensitive algorithm, COSNet, recently proposed to predict gene functions [5]. This algorithm, although designed to manage the imbalance in labeled data, tends to predict an excessively high proportion of positives when data are particularly unbalanced (that is in particular on most specific classes). By adding a term to the energy function of the network, we are able in modifying the dynamics in order to prevent the number of positives becomes too large. This energy term is minimized when the proportion of positive neurons (current positive rate) resembles the rate of positive labels in the training set (expected positive rate). The higher the difference between current and expected positive rates, the more the penalty to the energy function. We call this regularized version R-COSNet.

Results

We tested R-COSNet on the prediction of yeast genes, by using four different data sets and the classes of the FunCat ontology [6]. This ontology is structured in forest of trees, in which each node belong to one of the six levels of specificity. Level 1 refers to the root nodes, level i to nodes at distance i from the root. The considered classes are those with at least 20 positives and are spanned from level 1 to level 5. We compared our methods with a label propagation algorithm, LP-Zhu [2], and Support Vector Machine (SVM) with probabilistic output [4].

In Figure 1 we report the results in terms of F-score averaged across the functional classes belonging to the level 4 and level 5 of the hierarchy.

References

1. Oliver, S. Guilt-by-association goes global. Nature 2000, 403: 601-603.
2. Zhu, X, Ghahramani, Z, and Lafferty, J. Semi-supervised learning using gaussian fields and harmonic functions. In ICML 2003, 912-919.
3. Sharan, R, Ulitsky, I, and Shamir, R. Network-based prediction of protein function. Molecular Systems Biology 2007, 3:88.
4. Lin, HT, Lin, CJ, Weng, R. A note on platt's probabilistic outputs for support vector machines. Machine Learning 2007, 68(3): 267-276.
5. Bertoni, A, Frasca, M, Valentini, G. Cosnet: A cost sensitive neural network for semi- supervised learning in graphs. ECML/PKDD (1) 2011, Lecture Notes in Computer Science, 6911: 219-234.
6. Ruepp, A, et al. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Research 2004, 32(18): 5539-5545.