An unbalance-aware network integration method for gene function prediction

Marco Frasca, Alberto Bertoni and Giorgio Valentini
Dipartimento di Informatica
Università degli Studi di Milano, Italy
{frasca, bertoni,valentini}@di.unimi.it

Abstract

Data integration and the unbalance between functionally annotated and unannotated genes are relevant items in the context of network-based gene function prediction. Even if both these topics have been analyzed in recent works, to our knowledge no network integration methods, specific for unbalanced functional classes have been proposed in this context. We introduce an unbalance-aware network integration method based on the recently proposed *COSNet* algorithm, and we apply it to the genome-wide prediction of Gene Ontology terms with the *M. musculus* model organism.

Introduction

In silico prediction of gene function through computational analysis of a variety of genomic and proteomic data is a central goal of molecular biology and raises challenging bioinformatics problems [1]. Two of the main items that characterize the gene function prediction problem (GFP) are the unbalance between annotated and unannotated genes and the integration of multiple sources of evidence for functional annotation [4]. In the context of network-based GFP, both these problems have been addressed (see, e.g., [2, 5, 6]), but to our knowledge no network integration methods specific for unbalanced classification problems have been proposed.

We address this item by proposing a network integration method unbalance-aware, that explicitly takes into account the relatively low number of gene annotations to properly integrate multiple networked data. Extensive experiments with the *MouseFunc* benchmark [7] show the effectiveness of the proposed approach.

Methods

Our proposed network integration method leverages the cost-sensitive capabilities of *COSNet* (COst-Sensitive neural Network), a semi-supervised algorithm for learning node labels in graphs with unbalanced data [3].

Let $G =< V, E >$ be a undirected graph, where V is the set of nodes corresponding to genes, E the set of edges, and $\boldsymbol{W} : V \times V \longrightarrow [0, 1]$ the corresponding symmetric connection matrix, whose weights w_{ij} represents similarities between genes i and j. The *COSNet* algorithm is based on parametrized Hopfield networks $H =< \boldsymbol{W}, \gamma, \alpha >$, where γ is the neuron threshold and α is a

real number in $[0, \frac{\pi}{2}]$ by which the neuron states $\{\sin\alpha, -\cos\alpha\}$ are automatically learned from the data. By exploiting the bipartition (U, S) of V, where S is the set of labeled and U the set of unlabeled nodes, *COSNet* learns the "optimal" parameters (γ, α) from the data and computes a bipartition (U^+, U^-) of U through a cost-sensitive network dynamics by which neuron states are propagated across the graph. The conceptual separation between node labels and neuron activation values allows us to effectively deal with data imbalance [5].

Given a set of undirected graphs $G^{(d)}$, $1 \leq d \leq m$, represented through the corresponding adjacency matrices $\boldsymbol{W}^{(d)}$, our proposed approach combines the networks by weighting each matrix $\boldsymbol{W}^{(d)}$ with an unbalance-aware coefficient $h^{(d)}$, computed through a supervised procedure performed on the projections of the set of nodes V into the plane. More precisely, let $L : \mathbb{V} \longrightarrow \{+, -\}$ be a labeling function, where $\mathbb{V} = \bigcup_{d=1}^{m} V^{(d)}$; $V^{(d)}$ is halved in $(V_+^{(d)}, V_-^{(d)})$ where $V_+^{(d)} = \{k \in V^{(d)} | L(k) = +\}$ is the set of positive and $V_-^{(d)} = \{k \in V^{(d)} | L(k) = -\}$ is the set of negative examples.

The unbalance-aware network integration method can be set out in three main steps:

1. **Network projection to a plane**. For each network $G^{(d)}$, each node $k \in V^{(d)}$ is associated with a point $\Delta^{(d)}(k) \equiv \left(\Delta_+^{(d)}(k), \Delta_-^{(d)}(k)\right) \in \mathbb{R}^2$, where

$$\Delta_+^{(d)}(k) = \sum_{j \in V_+^{(d)}} w_{kj}^{(d)}, \qquad \Delta_-^{(d)}(k) = \sum_{j \in V_-^{(d)}} w_{kj}^{(d)}.$$

The bipartition $(V_+^{(d)}, V_-^{(d)})$ of $V^{(d)}$ induces in a natural way a bipartition $(I_+^{(d)}, I_-^{(d)})$ of the points $I^{(d)} = \{\Delta^{(d)}(k) \mid k \in V^{(d)}\}$, where:

$$I_+^{(d)} = \{\Delta^{(d)}(k) \mid k \in V_+^{(d)}\} \qquad I_-^{(d)} = \{\Delta^{(d)}(k) \mid k \in V_-^{(d)}\}.$$

2. **Learning a parametric line to separate positive and negative examples**. Consider now an arbitrary straight line in the plane of equation:

$$f_{\alpha,\gamma}(\Delta_+^{(d)}(k), \Delta_-^{(d)}(k)) = \cos\alpha \cdot \Delta_-^{(d)}(k) - \sin\alpha \cdot \Delta_+^{(d)}(k) + \gamma = 0$$

It separates the points of $I^{(d)}$ in $I_{\alpha,\gamma,+}^{(d)}$ and $I_{\alpha,\gamma,-}^{(d)}$:

$$I_{\alpha,\gamma,+}^{(d)} = \{\Delta^{(d)}(k) \mid f_{\alpha,\gamma}(\Delta^{(d)}(k)) > 0\} \qquad I_{\alpha,\gamma,-}^{(d)} = \{\Delta^{(d)}(k) \mid f_{\alpha,\gamma}(\Delta^{(d)}(k)) \leq 0\}.$$

Positive and negative points are linearly separated through an efficient quasi-linear two-steps approximated algorithm to maximize the F-score: at first the "optimal" slope $\tan\hat{\alpha}^{(d)}$ of the lines crossing the origin is computed and then, fixing the slope $\tan\hat{\alpha}^{(d)}$, the "optimal" intercept $\hat{\gamma}^{(d)}$ is selected. In both steps the maximization of the F-score is performed, thus obtaining an estimation $F^{(d)}$ of the maximal F-score achieved by the linear separator $f_{\hat{\alpha}^{(d)},\hat{\gamma}^{(d)}}$.

3. **Computation of net-weights and network combination**. The values $F^{(d)}$ estimated at step 2 are then used to compute the weights $h^{(d)} = \frac{F^{(d)}}{\sum_i F^{(i)}}$ for combining the networks $\boldsymbol{W}^{(d)}$ according to a weighted integration schema:

$$\boldsymbol{W}^* = \sum_d h^{(d)} \boldsymbol{W}^{(d)}$$

Intuitively the slope $\tan\alpha$ depends on the relationships between the positive Δ_+ and the negative Δ_- "neighborhood values" that determine the geometric location of each projected node into

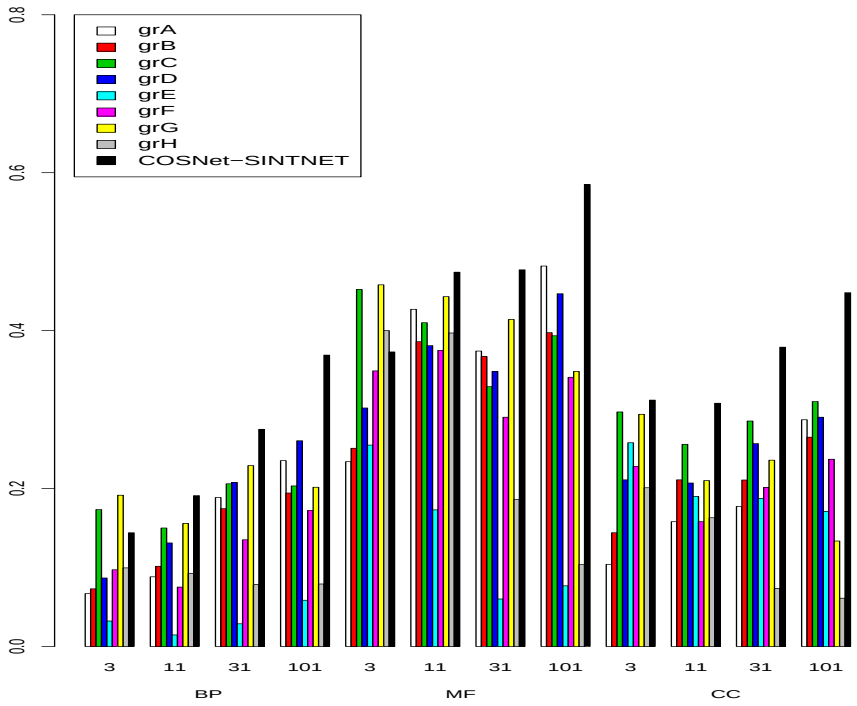

Figure 1: Comparison of the MouseFunc I challenge methods and COSNet in terms of F-score averaged across all the twelve considered GO categories. GrA: Calibrated ensembles of SVMs, proposed by Obozinski et al; GrB: Integrated Kernel-Logistic Regression (Lee et al.); GrC: GeneMANIA (Mostafavi et al); GrD: Multi-label hierarchical classification (Guan et al); GrE: Combination of classifier ensemble and gene networks (Kim et al); GrF: GeneFAS (Joshi et al); GrG: Query Retrieval Methods (Qi et al); Last black bar: unbalance-aware COSNet integration.

the plane. By learning the "optimal" parameters in terms of the F-score, we implicitly take into account the topological unbalance between positive Δ_+ and negative Δ_- neighborhoods.

Moreover we experimentally achieved a large and statistically significant positive Pearson correlation between the supervised classification of projected nodes performed through $f_{\alpha,\gamma}$ and the F-score achieved by *COSNet* in the corresponding node label prediction problem: as a consequence the F-scores obtained at step 2 are reliable estimates of the informativeness of each network, and thus well-suited to weight the network (step 3 of the unbalance-aware network integration method).

Finally, as shown in [5], by optimizing the linear separator $f_{\alpha,\gamma}$ we obtain the "near-optimal" parameters for *COSNet*: the parameters $(\hat{\gamma}, \hat{\alpha})$ that maximize the F-score achieved by $f_{\alpha,\gamma}$ move the corresponding parametrized Hopfield network $H =< \boldsymbol{W}, \hat{\gamma}, \hat{\alpha} >$ towards an equilibrium state (local minimum of energy).

Results and Conclusion

We applied our unbalance-aware network integration method using *COSNet* to the *MouseFunc* benchmark [7]. In this setting we combined 17 sources of evidence including expression data, sequence patterns, protein interactions, phenotype annotations, phylogenetic profiles, and other types of genomic data to predict gene functions of 21603 genes annotated to 2815 GO terms in *M. musculus*. These GO terms have a number of annotations ranging form 3 to 300, and for each GO domain, Biological Process(BP), Molecular Function (MF) and Cellular Component (CC), four categories of terms have been considered: the categories with 3–10, 11–30, 31–100 and 101–300 annotations respectively. The F-score and precision at fixed recall results achieved by our proposed approach are in most cases significantly better than those obtained by the 8 methods participating to the *MouseFunc* I challenge[1], according to the Wilcoxon rank sum test at 0.01 significance level. In Figure 1 we show the results in terms of F-score averaged by GO category. Our method achieves the best performance in all the categories, except the BP 3–10 and MF 3–10, where it is the third and the fourth best method respectively. These categories are among the most difficult to be predicted, since they have a low number of annotations, thus reducing the effectiveness of the cost–sensitive approach of our algorithm. To deal with these cases a regularized variant of the *COSNet* algorithm could be applied, as recently shown in [5].

The results suggest that learning strategies for unbalanced classification problems should be embedded into data integration algorithms to significantly boost gene function prediction methods, confirming results recently reported in literature [4].

References

[1] T.M. Murali, C.J. Wu and S. Kasif. The art of gene function prediction. *Nature Biotechnology*, 24:1474–1475, 2006.

[2] H Shin, K. Tsuda and B. Scholkopf. Protein functional class prediction with a combined graph. *Expert Systems with Applications*, 36:3284–3292, 2009.

[3] A. Bertoni, M. Frasca and G.Valentini. COSNet: a cost sensitive neural network for semi-supervised learning in graphs. In :European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 219–234, 2011.

[4] N. Cesa-Bianchi, M. Re, and G. Valentini. Synergy of multi-label hierarchical ensembles, data fusion, and cost-sensitive methods for gene functional inference. *Machine Learning*, 88(1):209–241, 2012.

[5] M. Frasca, A. Bertoni, M. Re, and G. Valentini. A neural network algorithm for semi-supervised node label learning from unbalanced data. *Neural Networks*, 43:84–98, 2013.

[6] S. Mostafavi and Q. Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 26(14):1759–1765, 2010.

[7] L. Pena-Castillo et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biology*, 9:S1, 2008.

[1]For the M. Leone and A. Pagnani group the predictions do not include all the functional classes, and for this reason we excluded this group from the experimental comparison.