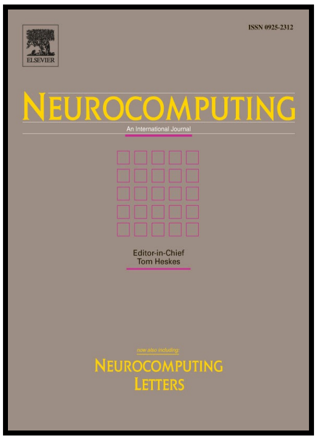
Author's Accepted Manuscript



www.elsevier.com/locate/neucom

COSNet: an R package for label prediction in unbalanced biological networks

Marco Frasca, Giorgio Valentini

PII: S0925-2312(16)00048-5
DOI: http://dx.doi.org/10.1016/j.neucom.2015.11.096
Reference: NEUCOM16624

To appear in: *Neurocomputing*

Received date: 29 June 2015
Revised date: 9 November 2015
Accepted date: 14 November 2015

Cite this article as: Marco Frasca and Giorgio Valentini, COSNet: an R packag for label prediction in unbalanced biological networks, *Neurocomputin* http://dx.doi.org/10.1016/j.neucom.2015.11.096

This is a PDF file of an unedited manuscript that has been accepted fo publication. As a service to our customers we are providing this early version o the manuscript. The manuscript will undergo copyediting, typesetting, an review of the resulting galley proof before it is published in its final citable forn Please note that during the production process errors may be discovered whic could affect the content, and all legal disclaimers that apply to the journal pertain

ACCEPTED MANUSCRIPT

COSNet: an R package for label prediction in unbalanced biological networks

Marco Frasca and Giorgio Valentini

Dipartimento di Informatica, Università degli Studi di Milano
Via Comelico 39 Milano, 20137, Italy
frasca@di.unimi.it

Abstract

Several problems in computational biology and medicine are modelled as learning problems in graphs, where nodes represent the biological entities to be studied, e.g. proteins, and connections different kinds of relationships among them, e.g. protein-protein interactions. In this context, classes are usually characterized by a high imbalance, i.e. positive examples for a class are much less than those negative. Although several works studied this problem, no graph-based software designed to explicitly take into account the label imbalance in biological networks is available. We propose *COSNet*, an **R** package to serve this purpose. *COSNet* deals with the label imbalance problem by implementing a novel parametric model of Hopfield Network (HN), whose output levels and activation thresholds of neurons are parameters to be automatically learnt. Due to the quasi linear time complexity, *COSNet* nicely scales when the number of instances is large, and application examples to challenging problems in biomedicine show the efficiency and the accuracy of the proposed library.

Keywords: Biological Network, Label Imbalance, Node Label Prediction, R package, Protein Function Prediction

Accepted manuscript

1. Introduction

Network Biology and Network Medicine opened new avenues for the discovery of the underlying biological and pathological properties of biological systems [1]. In this context, several prediction problems, such as the *automated prediction of protein functions* (*AFP*) [2], the *drug-repositioning* and *gene prioritization* problems [3], and the *prediction of gene-abnormal phenotypes associations* [4], can be modelled as network-based supervised or semi-supervised learning tasks, where a specific node property or class can be inferred from the topology and the a priori knowledge coded in the network. Nodes in the network represent the bio-molecular entities and connections their relationships; moreover, nodes usually are only partially labelled, and the aim is to extend the labelling to unclassified nodes. In many cases, only a small number of positive items is available, thus resulting in largely unbalanced classification problems.

For these purposes, we have developed *COSNet*, an **R** package available at the Bioconductor (`http://www.bioconductor.org/packages/release/bioc/html/`

ACCEPTED MANUSCRIPT

COSNet.html) and GitHub (https://github.com/m1frasca/COSNet_GitHub.git) repositories, which implements the homonym algorithm [5, 6], favourably applied to gene expression and function prediction problems [7, 8]. This method introduces a family of parametric HNs which can provide both binary predictions and discriminant scores to rank nodes for the class being predicted. Through a cost-sensitive strategy, which ensures the minimization of the energy function, the neuron thresholds and activation values are efficiently learnt to face the disproportion between positive and negative instances, thus preventing trivial solutions made up by almost all negative predictions.

In this paper we shortly review the *COSNet* algorithm, describe the functionalities of the homonym Bioconductor package, and present some examples of application to computational biology problems.

2. Problem and Background

To meet the increasing need of graph-based prediction tools for guiding researchers in clinical experimentations on biological networks, and in interpreting the corresponding results, several strategies over the years have been developed to assign node labels in graphs, ranging from *guilt-by-association* approaches to methods based on k-nearest neighbourhood, random walks and label propagation [9, 10, 11]. Moreover, in the network biology and medicine context, various network-based tools and packages covering different programming languages are available, including *GeneMANIA* (http://www.genemania.org), a web-server and MATLAB standalone application for gene function prediction and prioritization; *Aleph* (http://aleph-ml.sourceforge.net), a Java tool for machine learning on graphs including node label prediction with *label propagation* (*LP*) and *random walks* (*RW*); *DTHybrid* (http://alpha.dmi.unict.it/dtweb/dthybrid.php), an **R** package to infer Drug-Target interactions implementing the homonym algorithm [12]; *GUILDify* (http://sbi.imim.es/web/GUILDify.php) and *PhenoPred* (http://www.phenopred.org), web-servers to discover gene-disease associations.

Despite their proven effectiveness, these tools do not adopt imbalance-aware strategies, and they often fail when the positive class is largely under-represented; indeed, when classes are largely unbalanced, imbalance-aware strategies are required to prevent remarkable decay in performance [13].

3. Software Framework

COSNet is a semi-supervised method that embeds the input network, composed of positive, negative and unlabeled nodes, in a parametric Hopfield network $H(k, \rho)$, where k is the neuron activation threshold and $\rho \in (0, \frac{\pi}{2})$ a real number used to determine the values for neuron activation. Namely, activation values and neuron labels are conceptually separated, i.e. instead of the classical $\{0(-1),\ 1\}$, the activation values are $\{-\cos\rho,\ \sin\rho\}$ for unfired and fired neurons respectively, thus allowing to counterbalance the disproportion between positive and negative neurons by appropriately learning k and ρ [6].

COSNet can be summarized in two main steps (Figure 1):

Accepted manuscript

ACCEPTED MANUSCRIPT

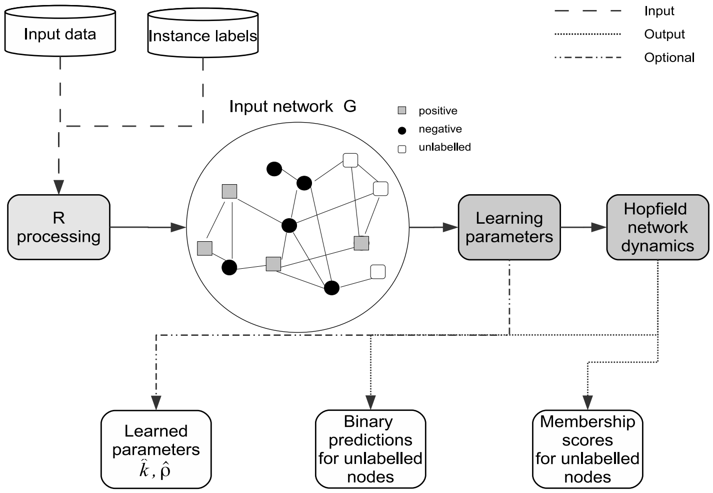

Input data
Instance labels
Input
Output
Optional
Input network G
positive
negative
unlabelled
R processing
Learning parameters
Hopfield network dynamics
Learned parameters $\hat{k}, \hat{\rho}$
Binary predictions for unlabelled nodes
Membership scores for unlabelled nodes

Figure 1: The *COSNet* framework. *COSNet* is based on two main modules (*gray boxes*): the automatic learning of model parameters and the network dynamics to provide binary and real prediction for unlabelled nodes.

- *Step 1: Parameter learning.* The optimal parameters $(\hat{k}, \hat{\rho})$ are automatically estimated from the labelled part S of the network to achieve a network state close to the equilibrium of the labeled sub-network (see [6])

- *Step 2: Network dynamics.* The sub-network restricted to unlabelled nodes with the learnt parameters is simulated. The achieved equilibrium state is then used to infer both binary predictions and discriminant scores.

COSNet takes overall time $\mathcal{O}(|S| \log |S| + |W|)$, where W is the weighted adjacency matrix of the graph, thus resulting in a computational complexity quasi-linear in the number of instances when W is sparse. The documentation, the vignette and the on-line Supplementary Information (available at `http://frasca.di.unimi.it/cosnet.html`) of the package provide the details about the usage and the **R** code for each step of the method.

The **R** package builds upon efficient **C** code implementing the computationally intensive parts in both Steps 1 and 2. The package also provides **R** functions to randomly partition the input data set and to implement k-fold cross-validation procedures to assess the generalization capabilities of the method. The software is conceived in a modular way, thus allowing, e.g., to compute separately or to provide different implementations of Step 1 (Parameter learning) and Step 2 (Network dynamics) of *COSNet*, or to apply different normalization and pre-processing network procedures, without affecting the rest of the code (see the Bioconductor documentation for more details).

4. Empirical results

We applied *COSNet* to two challenging problems in Computational Biology and Medicine: the *automated protein function prediction* and the *drug repo-*

Accepted manuscript

ACCEPTED MANUSCRIPT

Method	AUPRC			F		
	YEAST					
	BP	MF	CC	BP	MF	CC
GeneMANIA	0.178	0.217	0.263	0.241	0.307	0.340
LP	0.280	0.401	**0.427**	0.269	0.392	0.386
RW	0.119	0.237	0.179	0.105	0.185	0.132
GAIN	0.015	0.022	0.024	0.003	0.007	0.002
COSNet	**0.348**	**0.446**	0.383	**0.423**	**0.487**	**0.451**
	FLY					
GeneMANIA	0.097	0.235	0.187	0.145	0.306	0.243
LP	**0.106**	**0.258**	0.195	0.099	0.206	0.161
RW	0.069	0.175	0.139	0.076	0.160	0.128
GAIN	0.006	0.014	0.013	0.006	0.017	0.015
COSNet	**0.106**	0.251	**0.223**	**0.196**	**0.353**	**0.321**

Table 1: Area under the precision/recall curve (AUPRC) and F-score (F) for the AFP problem: 10-folds cross-validation results averaged across each GO branch separately (BP: Biological Process; MF: Molecular Function; CC: Cellular Component), for a total of 3469 and 4350 GO terms for respectively yeast and fly. Best results are in boldface, those statistically significant according to the Wilcoxon signed-rank test at $\alpha = 0.05$ significance level are underlined. The `cost` parameter, after a tuning on small–sized labeled data, has been set to 10^{-5}.

sitioning problems. In both the experiments we compared *COSNet* with the *LP* and *RW* algorithms, provided by the *Aleph* tool, and with *GAIN* [14], a "vanilla" HN. Moreover, we also tested two problem-specific methods recently proposed: *GeneMANIA*, the winner method in the *MouseFunc* mouse protein function prediction challenge, and *DTHybrid*, a method recently proposed and successfully applied to predict Drug-Target interactions. For evaluating the performance, we adopted two measures suitable for unbalanced problems: the Area Under the Precision-Recall Curve (AUPRC), and the F-measure, the harmonic mean of precision and recall, to respectively assess ranking and classification performances of the compared methods.

Automated protein function prediction. We predicted the GO [15] functions of 5775 yeast and 9361 fly proteins, by considering the most unbalanced GO functions, that is those with 3-300 positive annotations (Table 4). The input network (available at `http://frasca.di.unimi.it/cosnetdata/`) is obtained by integrating several sources covering various types of data, ranging from co-expression to genetic interactions and protein domain annotations.

To provide an idea of the computational efficiency of *COSNet*, with the yeast model organism only about 5 seconds are required to complete on the average an entire cycle of 10-fold cross-validation on each GO term, using an Intel i7-860 CPU 2.80 GHz, while *GeneMANIA* requires about 7 seconds and *RW* about 10 seconds. *LP* (4.5 seconds) and *GAIN* (0.7 seconds) are faster, but significantly less accurate than *COSNet* (except for *LP* AUPRC results on yeast CC data).

Prediction of drug-therapeutical categories associations. In the second experiment, we considered a network of 1253 `DrugBank` drugs, and the corresponding binary labels for 45 `DrugBank` categories, provided by the **R** package `bionetdata`

Accepted manuscript

ACCEPTED MANUSCRIPT

(`http://cran.r-project.org/web/packages/bionetdata`). The number of positive instances for Drug categories varies from 18 to 105. Also in this case, *COSNet* (`cost`=0.03) significantly outperformed the compared methods: the achieved AUPRC and F results, averaged across 45 DrugBank categories and estimated through 10-fold cross-validation procedures repeated 20 times, were respectively {0.301, 0.398} vs. {0.288, 0.388} (*DTHybrid*), {0.247, 0.276} (*LP*), {0.034, 0.055} (*RW*) and {0.030, 0} (*GAIN*).

5. Conclusions

COSNet is an efficient **R** Bioconductor package originally conceived for the accurate analysis of complex biomolecular networks. Its cost-sensitive features and the quasi-linear time complexity makes *COSNet* a versatile tool for the analysis of large-size networks characterized by unbalanced labellings.

References

[1] A. Barabasi, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nature Rev. Genet. 12 (2011) 56–68.

[2] R. Sharan, I. Ulitsky, R. Shamir, Network-based prediction of protein function, Molecular Systems Biology 3 (1). doi:10.1038/msb4100129.

[3] G. Valentini, et al., An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods, Artif. Intell. in Med. 61 (2) (2014) 63–78.

[4] G. Valentini, et al., Prediction of human gene-phenotype associations by exploiting the hierarchical structure of the human phenotype ontology, in: IWBBIO, Vol. 9043, 2015, pp. 66–77.

[5] A. Bertoni, M. Frasca, G. Valentini, COSNet: a cost sensitive neural network for semi-supervised learning in graphs, in: ECML, 2011, pp. 219–234. doi:10.1007/978-3-642-23780-5_24

[6] M. Frasca, et al., A neural network algorithm for semi-supervised node label learning from unbalanced data, Neural Networks 43 (0) (2013) 84 – 98. doi:10.1016/j.neunet.2013.01.021.

[7] M. Frasca, G. Pavesi, A neural network based algorithm for gene expression prediction from chromatin structure., in: IJCNN, IEEE, 2013, pp. 1–8. doi:10.1109/IJCNN.2013.6706954.

[8] M. Frasca, Automated gene function prediction through gene multifunctionality in biological networks, Neurocomputing 162 (2015) 48–56. doi:doi:10.1016/j.neucom.2015.04.007.

[9] S. Kohler, S. Bauer, D. Horn, P. Robinson, Walking the interactome for prioritization of candidate disease genes, Am. J. Human Genetics 82 (4) (2008) 948–958.

[10] L. Lan, N. Djuric, Y. Guo, S. Vucetic, MS-kNN: protein function prediction by integrating multiple data sources, BMC Bioinformatics 14 (Suppl 3:S8).

[11] X. Zhu, Z. Ghahramani, J. Lafferty, Semi-supervised learning using gaussian fields and harmonic functions, in: In ICML, 2003, pp. 912–919.

[12] S. Alaimo, et al., Drug-target interaction prediction through domain-tuned network-based inference., Bioinformatics 29 (16) (2013) 2004–2008.

[13] C. Elkan, The foundations of cost-sensitive learning, in: In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, 2001, pp. 973–978.

[14] U. Karaoz, et al., Whole-genome annotation by using evidence integration in functional-linkage networks, Proc. Natl Acad. Sci. USA 101 (2004) 2888–2893.

[15] M. Ashburner, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nature genetics 25 (1) (2000) 25–29. doi:10.1038/75556.

Accepted manuscript

ACCEPTED MANUSCRIPT

Required Metadata

Current executable software version		
Nr.	**(executable) Software metadata description**	
S1	Current software version	1.4.1
S2	Permanent link to executables of this version	*http : //www.bioconductor.org/packages/release/bioc/src/contrib/COSNet_1.4.1.tar.gz* *https : //github.com/m1frasca/COSNet_GitHub/blob/master/linux_package/COSNet_1.4.1.tar.gz*
S3	Legal Software License	GPL (>= 2)
S4	Computing platform/Operating System	Linux, Mac OS X 10.6 (Snow Leopard), Mac OS X 10.9 (Mavericks), Microsoft Windows
S5	Installation requirements & dependencies	**R** software environment version 3.1 or higher, *http : //www.r − project.org/*
S6	If available, link to user manual - if formally published include a reference to the publication in the reference list	*http : //www.bioconductor.org/packages/release/bioc/manuals/COSNet/man/COSNet.pdf* *http : //www.bioconductor.org/packages/release/bioc/vignettes/COSNet/inst/doc/COSNet_v.pdf*
S7	Support email for questions	frasca@di.unimi.it
Current code version		
Nr.	**Code metadata description**	
C1	Current code version	1.4.1
C2	Permanent link to code/repository used of this code version	*http : //www.bioconductor.org/packages/release/bioc/html/COSNet.html* *https : //github.com/m1frasca/COSNet_GitHub.git*
C3	Legal Code License	GPL (>= 2)
C4	Code versioning system used	svn, git
C5	Software code languages, tools, and services used	**C**, **R**
C6	Compilation requirements, operating environments & dependencies	**R** software environment version 3.1 or higher
C7	If available Link to developer documentation/manual	*http : //www.bioconductor.org/packages/release/bioc/manuals/COSNet/man/COSNet.pdf* *http : //www.bioconductor.org/packages/release/bioc/vignettes/COSNet/inst/doc/COSNet_v.pdf*
C8	Support email for questions	frasca@di.unimi.it

Table 2: Software and code metadata

Accepted manuscript