Neurocomputing 162 (2015) 48–56

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Automated gene function prediction through gene multifunctionality in biological networks

Marco Frasca

Dipartimento di Informatica, Università degli Studi di Milano, Via Comelico 39, Milano 20137, Italy

ARTICLE INFO

Article history:
Received 18 March 2014
Received in revised form
22 January 2015
Accepted 4 April 2015
Communicated by M. Chetty
Available online 14 April 2015

Keywords:
Gene multifunctionality
Biological networks
Hopfield networks
Gene function prediction
Gene ranking
Cost-sensitive learning

ABSTRACT

As the number of sequenced genomes rapidly grows, Automated Prediction of gene Function (AFP) is now a challenging problem. Despite significant progresses in the last several years, the accuracy of gene function prediction still needs to be improved in order to be used effectively in practice. Two of the main issues of AFP problem are the imbalance of gene functional annotations and the 'multifunctional properties' of genes. While the former is a well studied problem in machine learning, the latter has recently emerged in bioinformatics and few studies have been carried out about it. Here we propose a method for AFP which appropriately handles the label imbalance characterizing biological taxonomies, and embeds in the model the property of some genes of being 'multifunctional'. We tested the method in predicting the functions of the Gene Ontology functional hierarchy for genes of yeast and fly model organisms, in a genome-wide approach. The achieved results show that cost-sensitive strategies and 'gene multifunctionality' can be combined to achieve significantly better results than the compared state-of-the-art algorithms for AFP.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

High throughput biomolecular technologies have been made available a vast amount of genomic, proteomic and transcriptomic data and the experimental determination of gene functions is the most reliable way to characterize genes and their products. However, due to its inherent difficulty and expense, the experimental characterization of functions cannot appropriately scale up and the automated annotation of gene functions has therefore emerged as a challenging problem in computational and molecular biology [1]. The Automated Prediction of gene Functions (AFP) is a complex problem, with several distinctive features: functional classes (biological functions) are structured in a hierarchy with different levels of specificity (e.g. the Gene Ontology (GO) [2]) and labelings are not independent; each gene may have multiple labels (multi-label classification); classes are thousands (GO) and often highly unbalanced, with few positive and much more negative genes; negative instances are not uniquely defined, since usually only positive gene memberships are known for functional classes, and negatives in principle can be chosen with different strategies [3]; data are noisy and usually large-scale and high-dimensional; several heterogeneous sources of biological data are available, each one describing specific properties of genes, and, to achieve more reliable predictions, their integration with suitable methods is needed [4].

In this work we take into consideration two of these issues: the imbalance of class labelings and the multiple annotation of genes. Many attempts have been proposed in the literature for AFP. More general approaches characterize genes by a set of features, which in turn are exploited by machine learning algorithms to typically address a set of binary classification problems: predict whether or not a gene should be associated with a functional class [5]. Another commonly used approach is based on sequence homology, which adopts sequence alignment tool, e.g. BLAST [6], to find sequences of gene products (such as proteins) similar to the target sequence, and then transfers their known functional annotations to the target sequence as predictions [7,8]. Moreover, the availability of large-scale networks of genetic and physical interactions, where nodes are genes/gene products and connections are among nodes the gene pairwise relationships, has focused the investigation also on the design of network-based algorithms for AFP. The first network-based approaches have been based on the so-called *guilt-by-association* (GBA) rule, which makes predictions based on the interacting genes, assuming that interacting genes are likely to share similar functions [9–11]. Indirect neighbors have also been exploited to modify the notion of pairwise-similarities among nodes by accounting for pairs of nodes connected through intermediate ones [12–14].

Furthermore, gene functions can be predicted by propagating node labels through the network with an iterative process until

E-mail address: frasca@di.unimi.it

http://dx.doi.org/10.1016/j.neucom.2015.04.007
0925-2312/© 2015 Elsevier B.V. All rights reserved.

convergence [15,16], by tuning the amount of propagation we allow in the graph through Markov Random Walks [17,18], by evaluating the functional flow through the nodes [19]. Other relevant studies also adopted techniques based on Global graph consistency [20], on Hopfield networks [21–23], on Markov [24] and Gaussian Random Fields [25–27].

Despite their proved effectiveness, these methods totally or partially neglect two main issues of AFP. First, they do not appropriately handle the label imbalance affecting classes in biological taxonomies. The Gene Ontology is the most popular repository for biological functions and structures genes in three major ontologies (direct acyclic graphs): Molecular Function (MF), Biological Process (BP), and Cellular Component (CC). The most specific classes, which are those better describe the functions of genes, have usually very few annotations (genes that previous studies have shown having the function). This lack of information makes the prediction task very difficult, and cost-insensitive algorithms may suffer high decay in performance [28,29]. Second, when predicting in a flat setting, i.e. without considering the hierarchical structure of GO, such methods do not embed in their framework the multi-functional properties of genes. Indeed, recent works have introduced the concept of *gene multifunctionality* [30], which regards the property of some genes which are annotated with many classes of being really multifunctional in the cell cycle. The authors have shown that multifunctionality drives most computational predictions made by GBA-based methods, and, furthermore, that there exists a relationship between multifunctionality and the number of interacting partners in the network. Overall, multifunctionality has been investigated as a possible limitation of generalization capabilities of algorithms that infer gene functions exploiting solely the GBA rule, and some strategies have been suggested to prevent this limitation (e.g. avoiding the network sparsification). On the other hand, they disregard high gene degree may be a good indicator of the gene cell activity, and do not develop any strategy which exploits gene multifunctionality to improve the reliability of functional predictions.

In this work we propose an approach to cast gene multifunctionality in the prediction model of a network-based imbalance-aware algorithm, *COSNet* [23], recently proposed to predict node labels in partially labeled graphs. We analyzed biomolecular networks from model organisms to investigate the role multifunctionality has on the predictive capability of *COSNet*. Interestingly, we found that for almost all GO functions, the considered networks (that as suggested, we do not sparsify) contain several *exceptional genes*, which are genes annotated with the function c being predicted, without interacting partners annotated with c, but with high node degree (i.e. expected high multifunctionality). Such genes are likely to be wrongly predicted by most of all network-based AFP methods. Our strategy is designed to explicitly take into account the presence of exceptional genes and to exploit their multifunctionality to improve the accuracy of the prediction. The experimental validation carried out on two eukaryotic organisms in a genome-wide approach shows our method favorably compares with the state-of-the-art algorithms for AFP.

In the following, the AFP problem is formalized in Section 2, Section 3 introduces the multifunctionality in gene networks, whereas Sections 4.1 and 4.2 are dedicated to the description of *COSNet* and its extension to multifunctionality, respectively. The experimental validation of the proposed algorithm is discussed in Section 5.

2. Automated Function Prediction in gene networks

In the *Automated Function Prediction* (AFP) problem, genes are represented by a set of vertices V, and the relationships among genes are encoded in the symmetric matrix $\boldsymbol{W}: V \times V \longrightarrow [0,1]$, where W_{ij} is a precomputed measure of 'functional similarity' between genes $i,j \in V$. For a given functional class c (e.g. a term of the Gene Ontology), a labeling function $L_c: S \longrightarrow \{+,-\}$ is known, where $S \subset V$ is the set of labeled vertices. Moreover, a bipartition (S_+, S_-) of S is given, where $S_+ = \{i \in S | L(i) = +\}$ is the set of positive vertices and $S_- = \{i \in S | L(i) = -\}$ the set of those negative.

The aim is to derive a score function $\psi: U \longrightarrow \mathbb{R}$, which ranks unlabeled nodes according to the values of $\psi(i)$: the higher the score, the higher the likelihood that a gene belongs to the given functional class. $U = V \backslash S$ is the set of unlabeled vertices.

3. Multifunctionality in gene networks

The concept of 'multifunctionality' has been recently introduced in the scientific community to analyze the role 'multifunctional genes' have in the computational prediction of gene functions [30]. The *gene multifunctionality* can be defined as 'the number of molecular functions a gene is involved in', depending on the context and the interacting partners (other gene products). From a computational standpoint, multifunctionality is the number of classes an instance is classified as member.

Using GO as a source of functional annotations for genes, *Ranking by multifunctionality* means assigning higher rank to genes annotated with more GO terms. Indeed, if a gene is involved in many biological functions, the degree to which the gene has also a chosen function is higher than another gene which is, for example, annotated just with one GO term. In other words, algorithms which predict a GO term by assigning higher scores to multifunctional genes (genes already annotated with many other GO terms) are expected to achieve good performance for the majority of GO terms. Gillis and Pavlidis [30] define the multifunctionality score of gene i as follows:

$$\text{Score}_c(i) = \sum_{c \in GO | i \in c} \frac{1}{I_c * O_c} \tag{1}$$

where I_c and O_c are respectively the number of genes annotated and not annotated with term c. If we ignore the normalization by the product of I_c and O_c, $\text{Score}_c(i)$ is simply the number of functions the gene i has. This score provides a ranking which makes correct predictions for almost all the considered GO terms, achieving a mean Area Under the receiver operating characteristic Curve (AUC) of 0.9. Unfortunately, when predicting a single term with a flat approach, this score cannot be computed. On the other hand, the authors have also shown that gene multifunctionality is related to the node degree in gene networks, that is the number of genes interacting in a particular context. A greater number of interaction partners reflects (at least partially) the involvement in the biomolecular functions the partners have (hence expected higher multifunctionality). They show that if the data used for prediction is in some way a proxy for multifunctionality, and the algorithm used for classification can exploit this, very good prediction performance can result. Accordingly, the node (gene) degree can be assumed as suitable estimate of gene multifunctionality, and it can be exploited as prior knowledge to improve the predictive capabilities of network-based AFP methods.

4. Methods

In this section we first recall a recently proposed method for AFP, *COSNet*, designed to properly handle the class label imbalance, then we describe our approach to take into account the gene multifunctionality.

4.1. COSNet

COSNet (COst-Sensitive neural Network) [23] is a semi-supervised learning algorithm for predicting node labels in graphs with

Download English Version:

https://daneshyari.com/en/article/406094

Download Persian Version:

https://daneshyari.com/article/406094

Daneshyari.com