Neurocomputing 145 (2014) 44–52

Contents lists available at ScienceDirect

Neurocomputing

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

Active learning for protein function prediction in protein–protein interaction networks

Wei Xiong [a], Luyu Xie [a], Shuigeng Zhou [a,*], Jihong Guan [b]

[a] Shanghai Key Laboratory of Intelligent Information Processing, and School of Computer Science Fudan University, Shanghai 200433, China
[b] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form
22 May 2014
Accepted 24 May 2014
Available online 1 July 2014

Keywords:
Protein function prediction
Active learning
Collective classification
Protein–protein interaction network
Centrality

ABSTRACT

The high-throughput technologies have led to vast amounts of protein–protein interaction (PPI) data, and a number of approaches based on PPI networks have been proposed for protein function prediction. However, these approaches do not work well if annotated or labeled proteins are scarce in the networks. To address this issue, we propose an active learning based approach that uses graph-based centrality metrics to select proper candidates for labeling. We first cluster a PPI network by using the spectral clustering algorithm and select some informative candidates for labeling within each cluster according to a certain centrality metric, and then apply a collective classification algorithm to predict protein function based on these labeled proteins. Experiments over two real datasets demonstrate that the active learning based approach achieves a better prediction performance by choosing more informative proteins for labeling. Experimental results also validate that betweenness centrality is more effective than degree centrality and closeness centrality in most cases.

1. Introduction

In recent years, the rapid development of high-throughput experimental biology has led to huge amounts of unannotated protein sequences. Meanwhile, experimentally determining protein function is expensive and time-consuming. So there is a wider and wider gap between the pace of discovery of protein sequences and that of functional annotation of known proteins. Therefore, protein function prediction has been a fundamental challenge of biology in the post-genomic era. Although many efforts have been made to solve this problem, the proportion of annotated proteins is still very low. Among the 13 million protein sequences, there are only 1% sequences having experimentally validated annotations [1]. Even for the most well-studied model organisms, taking yeast as an example, approximately one-fourth of the proteins have no annotated functions [2].

Due to high cost and long duration of experimentally annotating protein function, there is increasing research on using computational approaches to predict protein function [3]. The recent development of high-throughput experimental biology and computational biology has generated vast amounts of protein–protein interaction (PPI) data [4], which are represented as networks, where a node corresponds to a protein and an edge corresponds to the interaction between a pair of proteins. Following that, a number of computational approaches for protein function prediction based on PPI networks have been proposed. These approaches make use of the observation that proteins with short distance to each other in a PPI network are more likely to have similar functions.

However, current network-based approaches do not work well when there are not enough labeled proteins in the PPI networks, which unfortunately is true in most scenarios. To address this issue, in this paper we propose an active learning [5] based approach that uses graph-based centrality metrics to select good candidates for labeling. Our approach consists of two steps: we first cluster a PPI network by using spectral clustering algorithm and select proper candidates for labeling within each cluster according to a certain node centrality metric, and then apply a collective classification algorithm to predict protein function based on these annotated proteins. To the best of our knowledge, this is the first study where active learning is employed to predict protein functions in PPI networks. The key idea behind active learning is that a machine learning algorithm can achieve higher accuracy with fewer training labels if it is allowed to choose the proper data for labeling from which it learns. Therefore, we let the learning algorithm pick a set of unannotated proteins to be labeled by an oracle (*i.e.*, a lab experiment), which will then be used as the labeled data set. In other words, we let the learning algorithm tell us which proteins to label, rather than select them randomly.

We conduct experiments on the *S. cerevisiae* and *M. musculus* PPI datasets, The experimental results demonstrate that the active

[*] Corresponding author.
E-mail addresses: wxiong@fudan.edu.cn (W. Xiong), taoistly@gmail.com (L. Xie), sgzhou@fudan.edu.cn (S. Zhou), jhguan@tongji.edu.cn (J. Guan).

http://dx.doi.org/10.1016/j.neucom.2014.05.075
0925-2312/© 2014 Elsevier B.V. All rights reserved.

learning based approach achieves a better prediction performance by choosing more informative proteins for labeling. Experimental results also validate that betweenness centrality is more effective than degree centrality and closeness centrality in most cases.

The rest of this paper is organized as follows: Section 2 describes background, Section 3 presents our approach, Section 4 gives the experimental evaluation results, and finally Section 5 concludes the paper.

2. Background

In a review [2], the existing network-based methods for protein function prediction were categorized into two main groups: direct methods and module-assisted methods. Direct methods propagate functional information through a PPI network and use the propagated information for functional annotation, examples include neighborhood counting methods and graph theoretic methods.

The majority method [6] and the indirect neighbors method [7] are two typical direct network-based approaches. Majority method [6] is the simplest direct method, it utilizes the biological hypothesis that interacting proteins probably have similar functions, it ranks each candidate function based on the function's occurrences in the immediate neighbors. Indirect neighbors method [7] assumes that proteins interacting with the same proteins may also have some similar functions, It exploits both indirect and immediate neighbors to rank each candidate function. Functional flow method [8] is a graph theoretic method, it simulates a discrete-time flow of functions from all proteins. At each time step, the function weight transferred along an edge is proportional to the edge's weight and the direction of transfer is determined by the functional gradient.

Module-assisted methods first identify functional modules in the network and then assign functions to all the proteins in each module, representatives are hierarchical clustering-based methods and graph clustering methods. A key problem of this kind of methods is how to define the similarity between two proteins. Arnau et al. [9] used the shortest path between proteins as a distance measure and apply hierarchical clustering to detecting functional modules. Up to now, numerous graph-clustering algorithms have been applied to detecting functional modules, such as clique percolation [10], edge-betweenness clustering [11], overlapping clustering [12] and Graphlet-based edge clustering [13].

Recently, Chua et al. [14] presented a simple framework for integrating large amount of diverse information for protein function prediction by using simple weighting strategies and a local prediction method. Hu et al. [15] hybridized the PPI information and the biochemical/physicochemical features of protein sequences to predict protein function. The prediction is carried out as follows: if the query protein has PPI information, the network-based method is applied; otherwise, the hybrid-property based method is employed. Additionally, network alignment approaches have been applied to predict protein function across species, such as GRAAL algorithm [16] and IsoRank algorithm [17].

Active learning [5] is a form of supervised machine learning in which a learning algorithm is able to interactively query the user (or some other information source) to obtain the desired outputs at some unlabeled data points. The key issue is to design the query strategy such that as few data points as possible are queried to achieve as large learning performance improvement as possible. The simplest and most commonly used query strategy is *uncertainty sampling* [18]. In this framework, an active learner queries the instance that the classifier is most uncertain. This strategy is often straightforward for probabilistic learning models. The *query-by-committee* (QBC) [19] strategy maintains a committee, each committee member is allowed to vote on the labelings of query candidates, the most informative query is considered to be the instance about which they most disagree. The fundamental premise behind the QBC strategy is minimizing the version space. The *expected model change* [20] strategy uses a decision-theoretic approach, it selects the instance that would impart the greatest change to the current model. The *expected error reduction* [21] strategy aims to measure not only how much the model is likely to change, but also how much its generalization error is likely to be reduced. It selects the instance that offers maximal expected error reduction to the classifier. The *density-weighted* [22] strategy suggests that the informative instances should not only be those which are uncertain, but also those which are representative of the underlying distribution (*i.e.*, inhabit dense regions of the input space).

Active learning has been applied to some bioinformatic problems, such as cancer classification [23], DNA microarray data analysis [24] and protein–protein interaction prediction [25], *etc.* However, to the best of our knowledge, there is no work on active learning for protein function prediction in the literature.

3. Methodology

3.1. Notation and problem definition

In this paper, a PPI network is represented as an indirected graph $G=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}=(V_1,\dots,V_n)$ is a set of n vertices and $\mathcal{E}$ is a set of weighted edges. Each vertex $V_i \in \mathcal{V}$ represents a protein and each edge $E_{i,j} \in \mathcal{E}$ represents an interaction between proteins V_i and V_j. Edge $E_{i,j}$ is labeled with a weight $w_{i,j}$ that indicates the interaction confidence. $\mathcal{F}=(F_1,\dots,F_m)$ is the set of m functions assigned to the proteins, and each vertex $V_i \in \mathcal{V}$ is assigned with at least one function. The functions of vertex $V_i \in \mathcal{V}$ are denoted by

$$\Phi(V_i)=[f_{i,1},f_{i,2},\dots,f_{i,j},\dots,f_{i,m}]^T \tag{1}$$

where

$$\begin{cases} f_{i,j}=1 & \text{if } V_i \text{ has the function } F_j, \\ f_{i,j}=0 & \text{otherwise.} \end{cases} \tag{2}$$

$\mathcal{V}$ can further be divided into two sets: $\mathcal{X}$ — the labeled vertices and $\mathcal{Y}$ — the vertices whose functions need to be determined.

In this paper, our goal is to label as few vertices $\{Y_i\} \subset \mathcal{Y}$ as possible with at least one of the functions in $\mathcal{F}$ based on the available information of the corresponding PPI network, so that the labeled vertices $\{Y_i\}$ and $\mathcal{X}$ together constitute the training set, which can be used to train an as good as possible classifier. Here, active learning is used for data selection to be labeled, the collective classification method is employed for classifier training.

3.2. Active learning strategies for protein function prediction

As we point out above, experimentally annotating protein function is expensive in terms of cost and effort, and current network-based approaches do not work well if annotated proteins are scarce. Therefore, strategies that minimize the amount of labeled data required in the supervised learning task would be useful. Active learning attempts to overcome the labeling bottleneck by asking queries in the form of unlabeled instances to be labeled by an oracle (*i.e.*, a lab experiment). In this way, the active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. The key idea behind active learning is that a machine learning algorithm can achieve higher accuracy with fewer training labels if it is allowed to choose the most proper data for labeling from which it learns.

In this study, the PPI network is represented as a graph, so it seems reasonable that we leverage graph structure to identify the nodes (proteins) in the graph that are important (central) for labeling. That is, we expect that such central nodes are proper candidates to label. Furthermore, we also note that nodes of the