Computational Biology and Chemistry 64 (2016) 263–270

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification

Gholam-Hossein Jowkar*, Eghbal G. Mansoori

School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

ARTICLE INFO

Article history:
Received 27 March 2016
Received in revised form 25 June 2016
Accepted 8 July 2016
Available online 12 July 2016

Keywords:
Disease gene identification
Biological networks
Positive-unlabeled learning
Ensemble of classifiers
Perceptron

ABSTRACT

Identification of disease genes, using computational methods, is an important issue in biomedical and bioinformatics research. According to observations that diseases with the same or similar phenotype have the same biological characteristics, researchers have tried to identify genes by using machine learning tools. In recent attempts, some semi-supervised learning methods, called positive-unlabeled learning, is used for disease gene identification. In this paper, we present a Perceptron ensemble of graph-based positive-unlabeled learning (PEGPUL) on three types of biological attributes: gene ontologies, protein domains and protein-protein interaction networks. In our method, a reliable set of positive and negative genes are extracted using co-training schema. Then, the similarity graph of genes is built using metric learning by concentrating on multi-rank-walk method to perform inference from labeled genes. At last, a Perceptron ensemble is learned from three weighted classifiers: multilevel support vector machine, *k*-nearest neighbor and decision tree. The main contributions of this paper are: (i) incorporating the statistical properties of gene data through choosing proper metrics, (ii) statistical evaluation of biological features, and (iii) noise robustness characteristic of PEGPUL via using multilevel schema. In order to assess PEGPUL, we have applied it on 12950 disease genes with 949 positive genes from six class of diseases and 12001 unlabeled genes. Compared with some popular disease gene identification methods, the experimental results show that PEGPUL has reasonable performance.

1. Introduction

In biomedical research, identification of genes underlying human hereditary is essential for prenatal and postnatal diagnosis and treatment (Piro and Cunto, 2012). Huntington as the first genetic disease, on the 4th chromosome of human DNA, was discovered by using polymorphism information (Bromberg, 2013). After that, the biologists focused on gene associated diseases and mutation on genes to identify genetic disorders and gene associated diseases. By screening them, the vulnerabilities of a child for inherited diseases before his/her birth can be determined. Also, the prognosis and counselling of affected families are discussed, and in some cases, this can lead to the development of therapeutic strategies (Piro and Cunto, 2012).

Since the abnormal function of genes, in the body, causes some diseases, it is necessary to identify the molecular pathway of these disorders (Bromberg, 2013). In this regard, the study on the properties of disease genes showed that the genes with the same or similar diseases stay in the same neighborhood in molecular networks (Piro and Cunto, 2012). Moreover, the traditional tools were expensive and time-consuming. These observations lead to the development of computational approaches for prediction or priorization of candidate disease genes (Wang et al., 2011). These approaches rely on the observations that diseases with the same or similar phenotype have the same biological characteristic. In this regard, computational analysis is used to combine different data sources, functional information of genes is used to extract disease gene knowledge, and machine learning methods are used to predict the disease genes.

Disease genes identification, in terms of learning type, can be categorized into three groups of unsupervised, supervised and semi-supervised learning. Traditionally, researchers face the classification of disease genes as a supervised learning method (Kohler et al., 2008; Smalter et al., 2007; Radivojac et al., 2008), though it is regarded as a semi-supervised problem in some researches (Yang et al., 2012; Yang et al., 2014). Since in semi-supervised methods, the learning starts with a small set of labeled (positive and negative) samples, the obtained model faces a small subset of positive samples and a huge subset of unlabeled samples

* Corresponding author.
E-mail addresses: hjowkar@shirazu.ac.ir (G.-H. Jowkar), mansoori@shirazu.ac.ir (E.G. Mansoori).

http://dx.doi.org/10.1016/j.compbiolchem.2016.07.004
1476-9271/

(possibly negative and/or positive) (Cerulo et al., 2010). This model is known as positive-unlabeled (PU) learning since some disease genes have not been identified yet, but have been treated as negative in the unlabeled set. In PU learning, these unidentified genes are used to identify positive genes. The main question to be answered is: Does unlabeled data help? The answer depends on the problem status which is tested.

In this paper, we present a Perceptron ensemble of graph-based positive-unlabeled learning (PEGPUL) on three biological networks: gene ontology (GO), protein domain (PD) and protein-protein interaction (PPI) networks. In this regard, after a brief explanation of building biological networks, a reliable set of negative genes with the helps of co-training schema are extracted in order to form a two-class problem. Next, the similarity graph is built using metric learning by concentrating on modified random walk method to perform inference from labeled gene based on the guilt-by-association rule. At last, a Perceptron ensemble is learned from three classifiers: weighted multilevel support vector machine (SVM), weighted *k*-nearest neighbor (KNN) and weighted classification and regression tree (CART).

The rest of this paper is organized as follows. In Section 2, some recent related works are reviewed and discussed. Materials and our proposed PEGPUL method are explained in Section 3. In Section 4, the experimental setting and results are presented. And Section 5 concludes the paper.

2. Related works

In early attempts, unsupervised clustering methods were used on GO (Freudenberg and Propping, 2002). Diseases of known genetic origin were clustered based on their phenotype similarities. Each candidate gene was scored according to its similarity to clusters. And this score showed the association degree of that gene when searching for a mutation in monogenic diseases. As a supervised method, Adie et al. proposed an algorithm (based on decision tree) called PROSPECTR which uses a variety of genome sequence-based features (Adie et al., 2005). Smalter et al. (2007) used SVM classifier with topological and sequence-based features of PPI networks to classify disease genes. Radivojac et al. (2008) proposed a SVM-based method called PhenoPred which uses three types of features including PPI network, protein sequences and protein functional information. PhenoPred combines three individual SVM classifiers, designed on three feature sets, to form the final classifier.

Selecting a subset (priorization) from candidate of disease genes has been recently studied. Kohler et al. (2008) used PPI data to build the similarity network and then to prioritize the candidate genes by random walk. They showed that PPI data is a valuable resource for this problem and is far better than direct interaction or shortest path measures for capturing relation between global similarity networks of genes. Vanunu et al. proposed PRINCE as a network-based method for prioritizing disease genes and inferring protein complex associations using PPI and disease-disease similarity measure (Vanun et al., 2010). By categorizing genes based on type of evidence, some disease identification methods are used through functional annotations (Yang et al., 2012; Yang et al., 2014), gene expression data (Adie et al., 2005) and ontologies (Yang et al., 2012; Yang et al., 2014; Freudenberg and Propping, 2002; Wang et al., 2007). New technologies make use of PPI networks (Kohler et al., 2008; Yang et al., 2012; Yang et al., 2014) as a precious source for candidate gene priorization.

In most of the researches mentioned, the confirmed disease genes have been considered as positive genes and the unconfirmed genes considered as negative genes. Although known genes can safely be assumed to be positive, obtaining negative genes is not straightforward. This is because the non-involvement of these genes in hereditary disease has not been proved. No matter which supervised method is used, training with wrongly potential positive genes affects the performance of the classifier. With regard to these limitations, PU method was proposed as a suitable procedure, by considering unconfirmed disease genes as an unlabeled (beside negative) set (Yang et al., 2012; Yang et al., 2014; Cerulo et al., 2010; Mordelet and Vert, 2011). There are two main approaches in PU learning (Cerulo et al., 2010): probability estimate correction which learns without negative samples, and selection of reliable negatives which extracts negative samples (Yang et al., 2012; Yang et al., 2014; Cerulo et al., 2010; Mordelet and Vert, 2011).

As an example of the first approach, Cerulo et al. (2010) presented a method called PosOnly. It trains SVM classifier on positive and unlabeled gene regulatory networks to predict the probabilities that differ by only a constant factor from the true conditional probabilities of being positive. In the other approach, the aim is to extract negative samples from unlabeled genes. One way of choosing reliable negatives is selecting random subset of unlabeled genes. Mordelet et al. proposed a bagging method called ProDiGe which repeatedly selects random subset from unlabeled genes and runs different SVM classifiers on each bootstrap (Mordelet and Vert, 2011). They used nine sources of information about the genes which can be categorized to three types of features including PPI network, protein sequences and protein functional information. The final result was gained by aggregating all of the presented results with the confidence score.

In early attendee of PU learning in disease gene identification, Yang et al. proposed a method called PUDI that used three biological networks: GO, PPI network and PDs (Yang et al., 2012). Based on PU selection of reliable negatives and also the similarity of being positive/negative genes, each unlabeled gene was divided into multiple subsets called reliable negative, likely positive, likely negative and weak negative. According to their similarity to positive or negative classes, they got different weights in order to be presented to multi-level weighted SVMs. In a new work (Yang et al., 2014), Yang et al. have extended their previous work and have proposed Ensemble based PU learning (EPU). They have added gene expression data and phenotypic similarity networks to their previous data sources. For ensembling classifiers, they have used KNN, SVM and naïve Bayes.

3. Materials and methods

Disease gene identification methods typically involve two stage, extracting a list of candidate genes and the criteria for learning, such as the involvement in a particular disease and learning procedure (Moreau and Tranchevent, 2012). In this section, the specification and statistics of disease datasets are described. Also, the related biological networks, the extracted features and their meanings are presented. Then, our proposed approaches for disease gene identification are explained.

3.1. Data and biological networks

In this work, by using online Mendelian inheritance in man (OMIM) database, positive genes are marked and other genes are treated as unlabeled. Six groups of confirmed diseases are selected as positive genes, based on (Goh et al., 2007) namely cardiovascular disease, endocrine disease, cancer disease, metabolic disease, neurological disease and ophthalmological disease. With respect to the quality and the performance of disease gene identification methods, the data are derived from multiple biological sources (Gill et al., 2014). Attempted to follow the methods outlined in (Yang et al., 2012), PD, PPI and GO data have been used as feature vector of each gene.