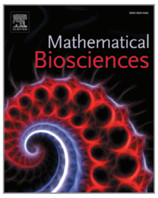
Contents lists available at ScienceDirect

Mathematical Biosciences

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



Classification by integrating plant stress response gene expression data with biological knowledge

Jun Meng[a], Rui Li[a], Yushi Luan[b,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning 116023, China.
[b] School of Life Science and Biotechnology, Dalian University of Technology, Dalian, Liaoning 116023, China.

ARTICLE INFO

Article history:
Received 24 March 2015
Revised 3 May 2015
Accepted 5 June 2015
Available online 17 June 2015

Keywords:
Biological knowledge
Information fusion
Neighborhood rough set
Plant stress response

ABSTRACT

Classification of microarray data has always been a challenging task because of the enormous number of genes. In this study, a clustering method by integrating plant stress response gene expression data with biological knowledge is presented. Clustering is one of the promising tools for attribute reduction, but gene clusters are biologically uninformative. So we integrated biological knowledge into genomic analysis to help to improve the interpretation of the results. Biological similarity based on gene ontology (GO) semantic similarity was combined with gene expression data to find out biologically meaningful clusters. Affinity propagation clustering algorithm was chosen to analyze the impact of the biological similarity on the results. Based on clustering result, neighborhood rough set was used to select representative genes for each cluster. The prediction accuracy of classifiers built on reduced gene subsets indicated that our approach outperformed other classical methods. The information fusion was proven to be effective through quantitative analysis, as it could select gene subsets with high biological significance and select significant genes.

1. Introduction

Plants are severely affected by various biotic and abiotic stresses. Biotic stress caused by other organisms, such as viruses, bacteria, insects, parasites and weeds, are the mainly contribution of cash crop loss, but hard to diagnose accurately. Whereas the abiotic stress is related with the non-living factor in environment like water, wind, sunlight, temperatures and radiation. It is the most harmful factor on the growth and productivity of crops worldwide, as plants are especially dependent on environmental factors [1]. Discriminative gene selection is crucial for the future development of stress-tolerant crops. So in this study, we focus on the analysis of plant stress response using gene microarray data.

Compared with tens of thousands of features, the number of samples in gene microarray data is relatively small. Furthermore, only a few of genes are highly associated with classification. Many gene selection approaches were introduced on the purpose of improving precision for microarray data analysis and reducing computation cost. Chuang et al. [2] proposed a hybrid system combined Tabu search and binary particle swarm optimization for feature selection using microarray data. Chopra et al. [3] used gene pair combination approach to solve the cancer classification problem. Gene selection method based on heuristic breadth-first search algorithm could find minimum gene subsets with high classification accuracy [4]. A two-stage procedure was proposed by Nguyen [5], partial least squares method was used for dimensionality reduction and survival prediction was made by proportional hazard regression. Recently, Shreem et al. [6] introduced a measure which is a hybrid of harmony search and Markov blanket for gene selection. Although those methods are efficient and the results were prominent, it is far from satisfactory in the view of biology, because of the poor biological significance in results [7].

Some previous researches for genomic analysis have successfully used prior biological knowledge in data mining methods, which make the result more acceptable to biologist. For example, Chen and Wang [8] linked gene identifiers in gene expression dataset with gene ontology, and used principal component analysis to select important genes. Reboiro-Jato et al. [9] integrated biological knowledge into classification model for the prediction analysis of microarray data. Bandyopadhyay et al. [10] developed a pathway based feature selection method for microarray data and a human breast cancer classification method using pathway was introduced by Gatza et al. [11]. Kim et al. [12] used pathway to improve accuracy of classification for cancer subtypes. Recently, a pathway-level for disease classification based on hyper-box principles was presented by Yang et al. [13]. Combination of the microarray data with biological information may highlight tissue- and process-specific biological processes underlining the added value of the developed method.

* Corresponding author. Tel.: +86 411 84706356; fax: +86 411 84706365.
E-mail addresses: mengjun@dlut.edu.cn (J. Meng), peter7rui@163.com (R. Li), luanyush@dlut.edu.cn (Y. Luan).

http://dx.doi.org/10.1016/j.mbs.2015.06.005
0025-5564/© 2015 Elsevier Inc. All rights reserved.

In the analysis of microarray data, clustering is an effective technique for finding groups of genes with similar expression pattern among samples. In addition, the search dimension of data mining algorithm is reduced by clustering genes. It is particularly important due to the high dimension of attributes and small number of samples for gene expression data [14]. Pollard and van der Laan [15] introduced a statistical framework for simultaneous clustering of gene expression data. An efficient gene selection technique based on fuzzy c-means clustering and neighborhood rough set is proposed by Xu et al. [16]. Tsai et al. [17] presented a multi-class clustering and prediction approach to analyze microarray data. Biological data fusion in clustering could help biologist to identify potentially meaningful relationships among genes. A method combined gene expression data and GO-derived information in clustering was proposed by Kustra and Zagdanski [18]. In order to enhance the biological value of clustering results, Milone et al. [19] incorporated pathway knowledge into self-organizing map training. In the application of microarray data classification, domain knowledge based on GO was conjunct with fuzzy clustering to help finding biological meaningful partitions [20].

However, some knowledge-integrated method only applied in clustering analysis among all the measures above. Others for gene selection didn't consider the internal biological relation between genes, and couldn't adjust the two types of information flexibly. In other words, in order to get a better classification performance, it is necessary to find out an appropriate proportion of biological knowledge in the information fusion approach. Additionally, many genes currently are unannotated by GO terms or pathways, clustering could be used to handle this situation rather than simply excluding them from the analysis. On the basis of our previous work [21], a clustering method combined GO term semantic similarity was proposed, gene selection based on neighborhood rough set was applied on the clusters. Moreover, we did a quantitative analysis to observe the impact of biological similarity on the prediction result and ensemble learning framework was built to enhance the robustness and generalization.

This paper is organized as follows: Section 2 reviews the existing researches on GO term semantic similarity. Detailed process of the proposed method is described in Section 3. Experiment results about the analysis of plant stress response are presented in Section 4. Finally, Section 5 provides the conclusion and future work.

2. Methods

2.1. Similarity incorporating

We use an information-fusion metric in clustering which involves both numerical similarity and biological similarity, based on gene expression data and GO knowledge. GO [22] is a gene annotation database (http://geneontology.org/) providing an ontology of defined terms representing gene product properties, which consists of three categories: molecular function (MF), biological process (BP) and cellular component (CC). The GO is structured as directed acyclic graph, and each term has defined relationships to one or more other terms in the same category, and sometimes to other categories. Plenty of approaches have been proposed to measure the semantic similarity between GO terms, based on the structure information in GO consortium.

Yu et al. [23] considered the topological distance and lowest common ancestors (LCAs) from GO to compute the gene functional similarity directly. Resnik [24] proposed a method based on information content (IC) to compute semantic similarity between GO terms. Another measure taking both the distance from LCA to the target terms and the distance from LCA to root into account was introduced by Schlicker et al. [25], while Wang et al. [26] considered all of the parent terms of the target GO term. Recently, a novel method based on semantic overlap ratio of annotations was proposed to measure gene functional similarity in GO context [27], and Liu et al. [28] introduced another new method called weighted multipath measurement.

An integrative approach, InteGO [29] was proposed, which is a gene functional similarity measurement which integrates three state-of-the-art gene semantic similarity measures. Its performance got a significant improvement on yeast, *Arabidopsis* and human.

However, limitations still exist in InteGO method, because it is sensitive to the selection of low performance measures, and its integration strategy may not be suitable for all gene pairs. In this study, we use InteGO2 [30], which involves choosing the most appropriate seed measures for each gene pair from a pool of candidate measures using a grouping method, and a metaheuristic search method is used to integrate the selected seed measures. The incorporated similarity is defined as:

$$\text{Similarity} = -(\alpha(1 - \text{Sim}_{\text{bio}}) + (1 - \alpha)\text{Sim}_{\text{num}}) \quad (1)$$

Where α is biological similarity weight that can be varied between 0 and 1. Sim_{bio} is biological similarity based on InteGO2 lying in [0, 1]. Sim_{num} is numerical similarity which is derived from microarray expression data using classical Euclidean distance. It is normalized from 0 to 1 in line with Sim_{bio}. The greater value of Sim_{bio} stands for the higher functional similarity, which is contrary to Sim_{num}. So $1 - \text{Sim}_{\text{bio}}$ is used. In this formula, we take the inverse of the incorporated similarity, in order to explain that with the value descending, the distance between two genes actually keeps increasing. We use parameter α to regulate these two types of similarity, when $\alpha = 0$, it is a classical similarity metric based on gene expression data. And when $\alpha = 1$, the functional similarity is used alone, the expression information is disregarded.

2.2. Affinity propagation

The affinity propagation (AP) algorithm was proposed by Frey and Dueck [31]. It's a clustering algorithm applied in many fields of data mining. Every data point in AP is considered as potential cluster center (exemplar), iteratively updating information between data points until the end of iteration or algorithm convergence. Pick out exemplars according to the result of message passing, and assign the rest of data points to the nearest exemplar. Compared with *K*-means and self-organizing map, affinity propagation has three strengths: (1) The number of clusters is determined by AP algorithm automatically. (2) AP could produce more stable and precise clustering result. (3) AP needs less time to achieve the same clustering accuracy.

Affinity propagation is based on similarity matrix, element $s(i, j)$ represents the distance between data point i and j. Usually, it is the negative value of Euclidean distance, so the greater value shows the closer distance. The value on diagonal $s(k, k)$ is called preference, data points with larger preference value are more applicable to be chosen as an exemplar. In general, we set all the data point's preference value the same, ensuring all data points are equally suitable as exemplars.

In each iteration of the clustering algorithm, AP transmits two kinds of message, responsibility and availability. "Responsibility" $r(i, k)$ shows how well-suited point k is to serve as the exemplar for point i, and "Availability" $a(i, k)$ means how appropriate it would be for point i to choose point k as its exemplar. $r(i, k)$ and $a(i, k)$ are calculated by the following rules:

$$r(i, k) = s(i, k) - \max\{a(i, j) + s(i, j)\}(j \in \{1, 2 \ldots N, j \neq k\}) \quad (2)$$

$$a(i, k) = \min\{0, r(k, k) + \sum_j \{\max(0, R(j, k))\}\} \quad (3)$$

In the process of iteration, sometimes two or more data points are suitable for the exemplar in a cluster at the same time, so the algorithm cannot converge. In this case, damp factor λ is introduced to improve the stability of AP, $r(i, k)$ and $a(i, k)$ are constrained by their