Improving Classification Accuracy Using Gene Ontology Information

Ying Shen and Lin Zhang*

School of Software Engineering, Tongji University, Shanghai, China
{yingshen,cslinzhang}@tongji.edu.cn

Abstract. Classification problems, e.g., gene function prediction problem, are very important in bioinformatics. Previous work mainly focuses on the improvement of classification techniques used. With the emergence of Gene Ontology (GO), extra knowledge about the gene products can be extracted from GO. Such kind of knowledge reveals the relationship of the gene products and is helpful for solving the classification problems. In this paper, we propose a new method to integrate the knowledge from GO into classifiers. The results from the experiments demonstrate the efficacy of our new method.

Keywords: Gene Ontology, Semantic Similarity, Distance Metric Learning.

1 Introduction

In the post-genomics era with the availability of large-scale gene expression data, gene function prediction becomes an emergent task. Computational approaches with novel classification techniques have been used to address this problem [3]. Despite of the success achieved by them, the improvement for the classification accuracy remains limited, because they only deal with the data obtained from the biological experiments, which contains noise and missing values. If additional information can be referred to in the prediction process, the classification accuracy should be improved. Fortunately, the Gene Ontology (GO) [9] provides us with such kind of information, which has been tentatively used for the gene function prediction [6, 14].

GO characterizes the functional properties of gene products using standardized terms. Based on GO, the semantic similarities are defined to quantitatively measure the relationships between two GO terms/gene products. Several methods have been proposed for this purpose [8, 10, 11]. Compared with the expression data, the semantic similarity information is more reliable and reflects the true relationships between the terms/gene products.

Several approaches have been proposed to make use of the semantic similarity information in the gene function prediction problems. Initially, researchers only used the semantic similarity to predict the functions for genes [7]. The problems is, because Gene Ontology is still under development, novel functions for some gene products

* Corresponding author.

D.-S. Huang et al. (Eds.): ICIC 2013, CCIS 375, pp. 171–176, 2013.

may be masked by their known functions if the classifier only relies on the current semantic similarity information. Later, some improved methods combining both the semantic similarity and the experimental data are proposed [6, 14]. The similarities based on the expression data and the semantic similarities are weighted and together form the final combined similarities. The likelihood of a gene *g* having a function represented by the term *t* is computed using the combined similarities. Term *t* with the largest likelihood will be assigned to *g* as its potential function.

In this paper, we propose a novel method which integrates the semantic similarity information into the existing classification techniques. Specifically, in the training process, our new algorithm will learn a distance metric using the semantic similarity information. In the prediction process, classifiers can use the learned distance metric to predict functions for genes. The experimental results demonstrate that the learned distance metric can enhance the performance of the classifier.

The rest of the paper is organized as follows. Section 2 provides some background knowledge about the global distance metric learning. Section 3 introduces our new algorithm. Section 4 reports the experimental results. Finally, Section 5 concludes the paper with a summary.

2 Global Distance Metric Learning

Intuitively, the distance metric learned from the training data would be more suitable than a generic distance metric for solving a specific problem. Global supervised distance metric learning aims to solve the following problem: given a set of pairwise constraints, to find a global distance metric that best satisfies these constraints. It has been shown that the learned distance metric can significantly enhance the classifier's accuracy [4, 5].

Pairwise Constraint. can be represented by a similarity constraint set S and a dissimilarity constraint set D. Given a set of points $\{x_k \mid k = 1,\ldots, n\}$, $(x_i, x_j) \in S$ if x_i and x_j are in the same class; and $(x_i, x_j) \in D$ if they are in the different classes, where $i, j \in \{1, \ldots.., n\}$. Given the two sets S and D, how can we learn a distance metric that satisfies both kinds of constraints? An algorithm proposed by Xing *et al.* [12] solves this problem by minimizing the sum of distances between the samples in S:

$$\begin{aligned} &\min_{A} \sum_{(x_i, x_j) \in S} \left\| x_i - x_j \right\|_A^2 \\ &s.t. \sum_{(x_i, x_j) \in D} \left\| x_i - x_j \right\|_A \geq 1,\ A \succeq 0 \end{aligned} \tag{1}$$

A is a positive semi-definite matrix used by the Mahalanobis distance. To solve the problem formulated in Eq. (1), two solutions can be found in [12].

3 Distance Metric Learning with GO Information

In this section, we describe a novel algorithm which integrates the semantic similarity information into the existing classification technique. Specifically, in the training