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

process, our algorithm learns a distance metric under the supervision of a semantic similarity matrix. In the prediction process, the learned distance metric is fed into the classifier to classify the testing samples.

3.1 Distance Based on the Expression Data

Given a set of gene products $\{g_k \mid k = 1,\ldots, n\}$, the distance between a pair of gene products g_i and g_j ($i, j \in \{1, \ldots.., n\}$) is defined by the Mahalanobis distance:

$$d_{exp}(g_i, g_j) = \| g_i - g \|_A = \sqrt{(g_i - g_j)^T A (g_i - g_j)} \tag{2}$$

A symmetric distance matrix D_{exp} can be formed consequently:

$$D_{exp} = \{d_{exp}(g_i, g_j)\}_{n \times n}, i, j \in \{i,...,n\} \tag{3}$$

3.2 Semantic Similarity over Terms

Wang's method [10] is adopted here to compute the semantic similarity between terms. In [10], a GO term A is represented as $\mathrm{DAG}_A = (A, T_A, E_A)$, where T_A is a set of terms consisting of A and all its ancestors, and E_A is a set of edges in GO that connects the terms in T_A. The contribution S of term t in T_A to term A is

$$\begin{cases} S_A(t) = 1 \text{ , if } t = A \\ S_A(t) = \max\{w * S_A(t') \mid t' \in children(t)\} \text{ , if } t \neq A \end{cases} \tag{4}$$

where w is a weight factor for the edge in E_A connecting t and t'. Given two terms A and B, the semantic similarity between them is defined as

$$sim_{Wang} = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{\sum_{t \in T_A} S_A(t) + \sum_{t \in T_B} S_B(t)} \tag{5}$$

3.3 Semantic (Dis)similarity over Gene Products

There are several approaches proposed for measuring the semantic similarity for gene products. In this paper, we propose another method to define the semantic similarity over genes. Specifically, the semantic similarity between g_1 and g_2 is defined as:

$$\begin{aligned} sim(g_1, g_2) &= \max sim(t_i, t'_j) \text{ , if } l_1 = l_2 \\ sim(g_1, g_2) &= \min sim(t_i, t'_j) \text{ , if } l_1 \neq l_2 \end{aligned} \tag{6}$$

where l_1, l_2 are the class labels for g_1 and g_2 In the training set. Using the semantic similarities computed using Eq. (6), a semantic similarity matrix S_{sem} can be formed:

$$S_{sem} = \{sim(g_i, g_j)\}_{n \times n}, i, j \in \{1,...,n\} \tag{7}$$

Because the semantic similarity value has been normalized into [0, 1], a semantic distance matrix D_{sem} can be obtained using Eq. (8).

$$D_{sem} = I_{n \times n} - S_{sem} \tag{8}$$

3.4 Algorithm

The algorithm is shown in Fig. 1. The optimization problem in step 4 is defined as

$$\min_{A} \sum_{i>j} (D_{exp}(i,j) - D_{sem}(i,j))^2 \quad s.t. A \succeq 0 \tag{9}$$

Training process

1. Calculate the semantic similarities for genes in the training set using Wang's method and form the semantic similarity matrix S_{sem} using Eqs. (6) and (7);
2. Calculate the semantic distance matrix D_{sem} using Eq. (8);
3. Calculate the distance matrix D_{exp} for the gene products in the training set using Eqs. (2) and (3);
4. Find a distance metric $\|\cdot\|_A$ that minimizes the difference between D_{exp} and D_{sem};

Prediction process

5. Classify the target genes using a knn classifier and the learned distance metric.

Fig. 1. Distance metric learning with the semantic similarity information

The convex optimization problem in Eq. (9) is solved using the gradient descent method to obtain a full matrix A. We define the cost function in Eq. (10):

$$\begin{aligned} h(A) &= \sum_{i>j} (D_{exp}(i,j) - D_{sem}(i,j))^2 \\ &= \sum_{i>j} \left[(g_i - g_j)^T A (g_i - g_j) - D_{sem}(i,j) \right]^2 \triangleq \sum_{i>j} f_{ij}^{\,2}(A) \end{aligned} \tag{10}$$

The gradient of the function $h(A)$ is

$$\nabla h = 2 \sum_{i>j} \left[f_{ij}(A) \frac{\partial f_{ij}}{\partial A} \right], \quad \frac{\partial f_{ij}}{\partial A} = (g_i - g_j)(g_i - g_j)^T \tag{11}$$

The rationale behind the algorithm is that, if the functions of the training samples have been known, the semantic similarities obtained using Eq. (6) can correctly reflect the relationships between gene products. If a global distance metric that suitably maps the expression data to D_{sem} is learned in the training process, it will alleviate the effect of noise in the expression data. Under this assumption, when using the learned distance metric in the prediction process, the classification accuracy should be improved.

4 Experiments and Results

To evaluate the performance of our algorithm, it is tested on two datasets. In the experiments, we compared the classification accuracies of the standard knn classifier and the improved knn classifier using the learned distance metric.

4.1 Data Description and Experimental Setup

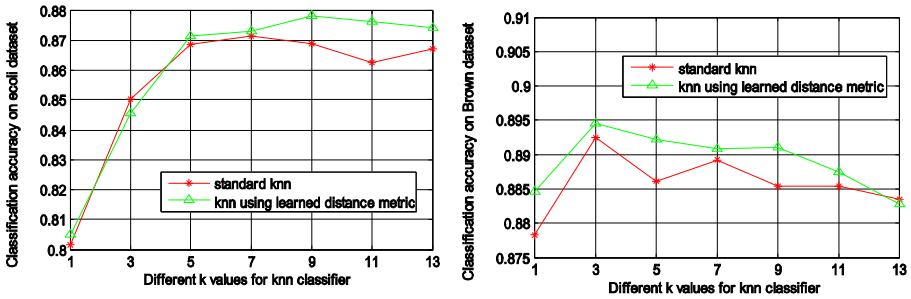
The first data set used in the experiments is the ecoli dataset from the UCI repository [1]. Annotations for gene products in the dataset were retrieved from the Uniprot database. After removing obsoleted genes in the Uniprot database, there are 309 genes left. In the experiments, only 5 classes (*cp*, *im*, *pp*, *imU*, and *om*) in which the numbers of instances are larger than 2 are used.

The second data set used is Brown's gene expression dataset (http://genome-ww.stanford.edu/clustering/Figure2.txt) [2]. The class labels can be obtained at http://compbio.soe.ucsc.edu/genex/targetMIPS.rdb. The genes are classified into 6 classes according to the MIPS function categories. Those genes that were not assigned to any of these classes and with multiple labels were eliminated. Annotations were retrieved from the SGD database. Those obsoleted genes in the SGD database were also removed. In the end, there are 224 genes left.

The semantic similarities for gene products in both datasets are computed using the *GOSemSim* package [13]. A 4-fold cross validation is performed on both datasets. We repeat the cross validation 20 times on each dataset and record the average classification accuracy for each k value.

4.2 Experimental Results

Fig. 2(a) shows the classification accuracies of the standard knn classifier and the improved knn classifier using the learned distance metric on the ecoli dataset. In this figure, the knn classifier using the learned distance metric outperforms the standard knn classifier except for the case of $k = 3$. When k is 11, the improved knn classifier outperforms the standard knn classifier by 1%. Fig. 2(b) shows the results of the experiments performed on the Brown's gene expression dataset. Again, the performance of the knn classifier using the learned distance metric is better than the standard knn classifier except for the case of $k = 13$. When k is 1, 5, and 9, the performance is improved by 0.6%.



(a) (b)

Fig. 2. Classification accuracies for the standard knn classifier and the improved knn classifier using the learned distance metric. (a) Classification accuracies on ecoli dataset; (b) Classification accuracies on Brown's gene expression dataset.

5 Conclusion

In this paper, we proposed a new method which utilizes the knowledge extracted from Gene Ontology to improve the gene function prediction accuracy by using the distance learning technique. In the training process, our method learns a global distance metric for the expression data under the supervision of the semantic similarity derived from GO. In the testing stage, the learned distance metric is used by the classifier to make decision. From the experiments, it can be seen that our method successfully improves the performance of the knn classifier, and provides a new way of integrating the GO knowledge into the classification problems in bioinformatics.

References

1. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/
2. Brown, M., Grundy, W., Lin, D., et al.: Knowledge-based Analysis of Microarray Gene Expression Data by Using Support Vector Machines. PNAS 97, 262–267 (2000)
3. Guyon, I., Weston, J., Barnhill, S., et al.: Gene Selection for Cancer Classification Using Support Vector Machines. Machine Learning 46, 389–422 (2002)
4. Hinton, G., Goldberger, J., Roweis, S., et al.: Neighborhood Components Analysis. In: Proc. NIPS, pp. 513–520 (2004)
5. Weinberger, K., Blitzer, J., Saul, L.: Distance Metric Learning for Large Margin Nearest Neighbor Classification. In: Proc. NIPS (2006)
6. Pandey, G., Myers, C.L., Kuma, V.: Incorporating Functional Inter-relationships into Protein Function Prediction Algorithms. BMC Bioinformatics 10, 142–164 (2009)
7. Tao, Y., Sam, L., Li, J., et al.: Information Theory Applied to The Sparse Gene Ontology Annotation Network to Predict Novel Gene Function. Bioinformatics 23, i529-i538 (2007)
8. Resnik, P.: Semantic Similarity in Taxonomy: An Information-based Measure and Its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research 11, 95–130 (1999)
9. The Gene Ontology Consortium: Gene Ontology: Tool for the Unification of Biology. Nature Genetics 25, 25–29 (2000)
10. Wang, J., Du, Z., Payattakool, R., et al.: A New Method to Measure the Semantic Similarity of GO Terms. Bioinformatics 23, 1274–1281 (2007)
11. Wu, H., Su, Z., Mao, F., et al.: Prediction of Functional Modules Based on Comparative Genome Analysis and Gene Ontology Application. Nucleic Acids Research 33, 2822–2837 (2005)
12. Xing, E., Ng, A., Jordan, M., et al.: Distance Metric Learning, with Application to Clustering with Side-information. In: Proc. NIPS, pp. 505–512 (2002)
13. Yu, G., Li, F., Qin, Y., et al.: GOSemSim: an R Package for Measuring Semantic Similarity Among GO Terms and Gene Products. Bioinformatics 26, 976–978 (2010)
14. Yu, H., Gao, L., Tu, K., et al.: Broadly Predicting Specific Gene Functions with Expression Similarity. Gene. 352, 75–81 (2005)