# Gene function prediction with knowledge from gene ontology

Ying Shen and Lin Zhang*

School of Software Engineering,
Tongji University,
Shanghai, China
Email: yingshen@tongji.edu.cn
Email: cslinzhang@tongji.edu.cn
*Corresponding author

**Abstract:** Gene function prediction is an important problem in bioinformatics. Due to the inherent noise existing in the gene expression data, the attempt to improve the prediction accuracy resorting to new classification techniques is limited. With the emergence of Gene Ontology (GO), extra knowledge about the gene products can be extracted from GO and facilitates solving the gene function prediction problem. In this paper, we propose a new method which utilises GO information to improve the classifiers' performance in gene function prediction. Specifically, our method learns a distance metric under the supervision of the GO knowledge using the distance learning technique. Compared with the traditional distance metrics, the learned one produces a better performance and consequently classification accuracy can be improved. The effectiveness of our proposed method has been corroborated by the extensive experimental results.

**Keywords:** gene ontology; semantic similarity; distance metric learning; gene function prediction; data mining; bioinformatics.

**Reference** to this paper should be made as follows: Shen, Y. and Zhang, L. (2015) 'Gene function prediction with knowledge from gene ontology', *Int. J. Data Mining and Bioinformatics*, Vol. 13, No. 1, pp.50–62.

**Biographical notes:** Ying Shen is an Assistant Professor in School of Software Engineering, Tongji University, Shanghai, PR China. She received her PhD degree from City University of Hong Kong in 2012. Her research interests include biological data mining, knowledge-based clustering, RNA motif recognition and RNA folding prediction.

Lin Zhang received the BS and MS degrees from the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, PR China, in 2003 and 2006, respectively. He received the PhD degree from the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, in 2011. From March 2011 to August 2011, he was a Research Assistant with the Department of Computing, the Hong Kong Polytechnic University. He is now an Assistant Professor in School of Software Engineering, Tongji University, Shanghai, PR China. His research interests include biometrics, pattern recognition, computer vision and perceptual image/video quality assessment, etc.

*This paper is a revised and expanded version of a paper entitled 'Improving classification accuracy using gene ontology information' presented at the 'International Conference on Intelligent Computing', Nanning, Guangxi Zhuang Autonomous Region, China, 28–31 July 2013.*

## 1 Introduction

Classification problems are very important in bioinformatics. For example, 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. Although the predictions made by computational algorithms are not as accurate as predictions made by human experts, the filtered results predicted by the computational methods greatly save the efforts for the biologists.

In order to improve the classification accuracy, many approaches have been proposed from the perspective of machine learning and pattern recognition (Furey et al., 2000; Guyon et al., 2002; Lee and Zhang, 2006; Nevins and Potti, 2007; Liu and Huang, 2008; Schweikert et al., 2009; Zheng et al., 2009; Cai et al., 2010; Leung and Hung, 2010; Zare et al., 2011; Zheng et al., 2011; Wang et al., 2012). Despite of the success achieved by these advanced techniques, 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. Intuitively, if additional information about the gene products is referred to in the prediction process, the classification accuracy should be improved regardless of the classification techniques used. Fortunately, the Gene Ontology (GO) (The Gene Ontology Consortium, 2000) provides us with such kind of information, which has been tentatively used for the gene function prediction in the last decade (Yu et al., 2005; Pandey et al., 2009).

GO characterises the functional properties of gene products using standardised terms. It contains three ontologies: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC). Based on GO, the semantic similarities are defined to quantitatively measure the relationships between two GO terms as well as two gene products. Several methods have been proposed to measure the semantic similarities over terms and gene products (Jiang and Conrath, 1997; Lin, 1998; Resnik, 1999; Pekar and Staab, 2002; Cheng et al., 2004; Wang et al., 2005; Schlicker et al., 2006; Wang et al., 2007), e.g., Resnik's method (Resnik, 1999) and Wang's method (Wang et al., 2007) for semantic similarity computation over terms and the 'Max' method (Wu et al., 2005) and 'Ave' method (Wang et al., 2005) for semantic similarity computation over genes, etc.

Compared with the expression data which may contain noise, the semantic similarity information is more reliable and reflects the true relationship between the terms and gene products. Considering its advantage, the semantic similarity information has been used as additional knowledge in the classification problems. The notion of the usefulness of the semantic similarity information for improving the classification accuracy is based on the assumption that the gene products with similar functions should also be similar in the biological experiments, i.e. the semantic similarity should be consistent with the similarity based on the expression data or other experimental data. Therefore, if the expression data contain noise which result in inconsistency between the two kinds of similarity, the similarity based on the gene expression data can be corrected under the supervision of the semantic similarity.

Several approaches have been proposed to make use of this semantic similarity information for the gene function prediction problems. Initially, researchers only used the semantic similarity information to predict the functions for genes. For example, a method proposed by Tao et al. (2007) first calculates the semantic similarity between the target gene and the training samples. Then the algorithm sorts the semantic similarity values and uses the k-Nearest Neighbour (KNN) classifier to predict functions for the target gene. Here comes the problem: because gene ontology is still under development and far from completeness, only some of the gene functions can be revealed by the experimental data and novel functions for some gene products may be masked by their known functions if the classifier only relies on the current semantic similarity information. When people realised that this kind of information is insufficient for gene function prediction, some improved methods combining both the semantic similarity and the experimental data are proposed (Yu et al., 2005; Pandey et al., 2009). In Pandey's method (Pandey et al., 2009), the similarity based on the expression data and the semantic similarity are weighted and together form the final combined similarity defined for two gene products. The likelihood of a gene *g* having a function represented by the term *t* is computed using the combined similarity. Term *t* with the largest likelihood will be assigned to *g* as its potential function. The attempt of making use of semantic similarity for gene function prediction is at an early stage and inadequate. Some essential problems such as the relationship between the semantic similarity and the gene expression similarity are still under discussion.

In this paper, we propose a novel method which integrates the semantic similarity information into the existing classification techniques. This method is inspired by the distance metric learning technique. 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 organised as follows. Section 2 provides some background knowledge about the global distance metric learning and a representative method in this field. Section 3 introduces our new algorithm which incorporates the semantic similarity information into the existing classification technique. Section 4 reports the experimental results. Finally, Section 5 concludes the paper with a summary.

## 2 Global distance metric learning

In this section, we will introduce some background knowledge about the distance metric learning. The first question is: why should we learn a distance metric? That is because the similarity/distance measure can significantly affect the classification results. For instance, in the KNN classifier, the distances between the test sample and the training samples will be first calculated and then the prediction is made based on the distances obtained in the first stage. Intuitively, the distance metric learned from the training data would be more suitable than a generic distance metric for solving a specific problem.

In global distance metric learning, we can learn a global distance metric using the samples in the training set. 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.

*Pairwise constraint* can be represented by two sets: the similarity constraint set $S$ and the dissimilarity constraint set $D$. Given a set of points $\{x_k \mid k = 1,\ldots,n\}$, $(x_i, x_j) \in S$, if two instances $x_i$ and $x_j$ are in the same class and $(x_i, x_j) \in D$, if they are in the different classes, where $i, j \in \{1,\ldots, n\}$.

According to the definition of pair wise constraint, the former problem can be further described as: Given two constraint sets $S$ and $D$, to find a distance metric that minimises the distance of samples in the same class and maximises the distance of samples in the different classes simultaneously. Researchers have shown that the learned distance metric can significantly enhance the classifier's accuracy than using a generic Euclidean distance metric (Hinton et al., 2004; Weinberger et al., 2006).

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. (2002) tries to solve this problem. It minimises the sum of distances between the samples in $S$ by solving a convex optimisation problem:

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

where $A$ is a positive semi-definite matrix used in the Mahalanobis distance:

$$d_A(x,y) = \| x - y \|_A = \sqrt{(x-y)^T A(x-y)} \tag{2}$$

The first constraint in equation (1) guarantees that $A$ does not collapse the dataset into a single point and the second constraint ensures that the learned matrix $A$ is positive semi-definite.

To solve the problem formulated in equation (1), solutions were provided by Xing et al. (2002) for two different cases of $A$. The first solution is the Newton-Raphson method for the case of an optimised diagonal matrix $A$. For the case of a full matrix $A$, Newton's method becomes prohibitively expensive. Therefore, Xing *et al.* used the gradient ascent method to solve the problem instead.

## 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. The algorithm is an extension of Xing's method. In this algorithm, the semantic similarities provide the constraints defined in the global distance metric learning problem. 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. Because the concept of 'similarity', 'dissimilarity' and 'distance' are equivalent, we will use them according to the context in the following sections.

In Section 3.1, we will first introduce the distance computation based on the gene expression data. In Sections 3.2 and 3.3, the semantic (dis)similarity computation for GO terms and gene products are presented. In Section 3.4, the details of the new algorithm proposed by us are described.

## *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:

$$d_{exp}(g_i, g_j) = \| g_i - g_j \|_A = \sqrt{(g_i - g_j)^T A (g_i - g_j)} \tag{3}$$

where $A$ is a positive semi-definite matrix.

When $A$ is a unit matrix, the Mahalanobis distance degenerates to the Euclidean distance.

Using the distances of all pairs of gene products, an $n \times n$ symmetric distance matrix $D_{exp}$ can be formed:

$$D_{exp} = \{d_{exp}(g_i, g_j)\}_{n \times n}, i, j \in \{i,\ldots,n\} \tag{4}$$

## *3.2 Semantic similarity over terms*

We adopt Wang's method (Wang et al., 2007) in our algorithm to compute the semantic similarity between terms.

In Wang's method, a GO term $A$ is represented as the tuple 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 connect the terms in $T_A$. The method defines a semantic value for term $A$ based on the contributions from all terms in $T_A$. The contribution $S$ of term $t$ in $T_A$ to term $A$ is defined as:

$$\begin{cases} S_A(t) = 1 if t = A \\ S_A(t) = \max\{w * S_A(t') \mid t' \in children(t)\} \ if t \neq A \end{cases} \tag{5}$$

where $w$ is a weight factor for the edge in $E_A$ connecting $t$ and its child $t'$. The authors suggested using $w = 0.8$ for the edges representing the 'is-a' relationship and $w = 0.6$ for the edges representing the 'part-of' relationship. Then the semantic value of $A$ is defined as the sum of contributions of terms in $T_A$:

$$SV(A) = \sum_{t \in T_A} S_A(t) \tag{6}$$

Given two terms $A$ and $B$ in form of DAG$_A$ and DAG$_B$, the semantic similarity between $A$ and $B$ is defined as:

$$sim_{Wang} = \frac{\sum_{t \in T_A \cap T_B} (S_A(t) + S_B(t))}{SV(A) + SV(B)} \tag{7}$$

## *3.3 Semantic (dis)similarity over gene products*

There are several approaches, e.g., the 'Max' method (Wu et al., 2005), proposed for measuring the semantic similarity for gene products. The semantic similarity for gene products is often based on the semantic similarity between their annotated terms. Suppose $\{t_i \mid i = 1,\dots,n\}$ and $\{t'_j \mid j = 1,\dots,m\}$ are annotations for two gene products $g_1$ and $g_2$. In the 'Max' method, the semantic similarity between $g_1$ and $g_2$ is defined as the maximum value of the semantic similarity between their annotations:

$$sim_{MAX}(g_1, g_2) = \max sim(t_i, t'_j) \tag{8}$$

where $t_i$ and $t'_j$ are annotations of $g_1$ and $g_2$, respectively, and $sim(t_i, t'_j)$ is the semantic similarity between $t_i$ and $t'_j$ computed by Wang's method.

If two genes $g_1$ and $g_2$ are annotated with a common term, their semantic similarity computed using the 'Max' method will be large. However, in the biological experiments, the functions of $g_1$ and $g_2$ may be quite different, although some of their annotations are the same. The 'Max' method may lead to inconsistency between the semantic similarity and the similarity based on the expression data due to the adoption of the incorrect relationship between gene products in such case.

To solve this problem, we propose another method to define the semantic similarity over genes. Specifically, we define the semantic similarity between $g_1$ and $g_2$ as the following:

$$\begin{aligned} sim(g_1, g_2) &= \max sim(t_i, t'_j)\ if l_1 = l_2 \\ sim(g_1, g_2) &= \min sim(t_i, t'_j)\ if l_1 \neq l_2 \end{aligned} \tag{9}$$

where $l_1$, $l_2$ are the class labels for $g_1$ and $g_2$ In the training set.

In equation (9), if two genes are in the same class, their semantic similarity is defined as the maximum value of the semantic similarity between their annotations; while if they are in different classes, their semantic similarity is defined as the minimum value of the semantic similarity between their annotations. Such definition is meaningful for the classification problems. If two genes are in the same class, i.e. they have similar functions, the terms assigned to them in the given problem should be similar; on the contrary, if they are in different classes, their annotations should be dissimilar.

Using the semantic similarities of all pairs of genes, an $n \times n$ semantic similarity matrix $S_{sem}$ can be formed:

$$S_{sem} = \{sim(g_i, g_j)\}_{n\times n}, i, j \in \{1, \dots, n\} \tag{10}$$

Because the semantic similarity value has been normalised into [0, 1], a semantic distance matrix $D_{sem}$ can be obtained using equation (11).

$$D_{sem} = I_{n\times n} - S_{sem} \tag{11}$$

## *3.4 Algorithms*

Our new algorithm is shown in Figure 1. In Step 4, the optimisation problem is defined as the following:

$$\begin{aligned} &\min_A \sum_{i>j} (D_{exp}(i,j) - D_{sem}(i,j))^2 \\ &s.t. A \succeq 0 \end{aligned} \tag{12}$$

**Figure 1** Distance metric learning with the semantic similarity information

**Training process**

1. Calculate the semantic similarities for gene products in the training set using Wang's method and form the semantic similarity matrix $S_{sem}$ using equations (9) and (10);
2. Calculate the semantic distance matrix $D_{sem}$ using equation (11);
3. Calculate the distance matrix $D_{exp}$ for the gene products in the training set using equations (3) and (4);
4. Find a Mahalanobis distance metric $\|\cdot\|_A$ that minimizes the difference between $D_{exp}$ and $D_{sem}$;

**Prediction process**

5. Classify the target gene products using a KNN classifier with the optimised distance metric obtained in Step 4.

The constraint in equation (12) guarantees that the matrix *A* is positive semi-definite.

The convex optimisation problem in equation (12) is solved using the gradient descent method to obtain a full matrix *A*. We define the cost function in equation (13).

$$\begin{aligned} h(A) &= \sum_{i>j} (D_{exp}(i,j) - D_{sem}(i,j))^2 \\ &= \sum_{i>j} \left[ (g_i - g_j)^T A (g_i - g_j) - D_{sem}(i,j) \right]^2 \\ &\triangleq \sum_{i>j} f_{ij}{}^2(A) \end{aligned} \tag{13}$$

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

$$\begin{aligned} \nabla h &= 2\sum_{i>j} \left[ f_{ij}(A) \frac{\partial f_{ij}}{\partial A} \right] \\ \frac{\partial f_{ij}}{\partial A} &= (g_i - g_j)(g_i - g_j)^T \end{aligned} \tag{14}$$

In equation (13), we use the sum of square errors between each pair of values in $D_{exp}$ and $D_{sem}$ to measure the difference between the two distance matrices.

Generally speaking, our new algorithm aims to learn a global distance metric that best maps the expression data to $D_{sem}$. The rationale behind the algorithm is that, if the functions of the training samples have been known, the semantic similarities obtained using equation (9) 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. Therefore, it renders the distances between the gene products in the training set and the testing set more representative of their correct quantitative relationship. Under this assumption, when using the learned distance metric in the prediction process, the classification accuracy should be improved. In this algorithm, on one hand, the property of the expression data is preserved. On the other hand, the semantic similarity has been integrated into the classification process. Therefore, our algorithm avoids the problem that the novel functions may be masked by the known functions in the classification process.

## 4 Experiments and results

To evaluate the performance of our algorithm, it is tested on two datasets. The first dataset is the *E. coli* dataset from the UCI repository (Asuncion and Newman, 2013) and the second dataset is Brown's gene expression data (Brown et al., 2000). In the experiments, we compared the classification accuracies of the standard KNN classifier and the improved KNN classifier using the learned distance metric. In Section 4.1, we will first introduce some details about the two datasets. In Section 4.2, we will show the experimental results for the two methods and offer some explanations.

### *4.1 Data description and experimental setup*

### *4.1.1 E. coli dataset*

The first data set is the *E. coli* dataset from the UCI repository (Asuncion and Newman, 2013). It consists of 336 proteins from the Uniprot database distributed in six classes (*cp* (cytoplasm), *im* (inner membrane without signal se-quence), *pp* (perisplasm), *imU* (inner membrane, un-cleavable signal sequence), *om* (outer membrane), *omL* (outer membrane lipoprotein), *imL* (inner membrane lipoprotein), *imS* (inner membrane, cleavable signal sequence)). The *E. coli* dataset is used for protein localisation site prediction. Annotations for gene products in the dataset were retrieved from the Uniprot database. We removed those genes obsolete in the Uniprot database. After this step, there are 309 genes left and the number of instances in each class is a bit less than the original number in UCI repository. The details of the *E. coli* dataset are shown in Table 1. In the experiments, we only used five classes (*cp*, *im*, *pp*, *imU* and *om*) in which the numbers of instances are larger than 2.

**Table 1** *E. coli* dataset from UCI repository

| *class name* | *# of instances* | *class name* | *# of instances* |
|---|---|---|---|
| cp | 131 | im | 76 |
| pp | 46 | imU | 34 |
| om | 19 | omL | 0 |
| imL | 1 | imS | 2 |
| # of genes | 309 | # of attributes | 7 |

### 4.1.2 Brown's gene expression dataset

The second data set is Brown's gene expression dataset (it can be downloaded from http://genome-ww.stanford.edu/clustering/Figure2.txt) (Brown et al., 2000), which contains the expression data for 2467 genes. The class labels can be obtained at http://compbio.soe.ucsc.edu/genex/targetMIPS.rdb. The genes are classified into six classes (*tca* (tricarboxylic-acid pathway), *resp* (respiration chain complexes), *ribo* (cytoplasmic ribosomal proteins), *proteas* (proteasome), *hist* (histones) and *hth* (Helix-turn-helix)) according to the MIPS function categories. We eliminated those genes that were not assigned to any of these classes and those with multiple labels. Annotations were retrieved from the SGD database. We also removed the genes obsolete in the SGD database. After these steps, there are 224 genes left with 79 attributes. The details of the Brown's dataset are shown in Table 2.

**Table 2** Brown's gene expression dataset

| *class name* | *# of instances* | *class name* | *# of instances* |
|---|---|---|---|
| tca | 14 | resp | 27 |
| ribo | 121 | proteas | 35 |
| hist | 11 | hth | 16 |
| # of genes | 224 | # of attributes | 79 |

The semantic similarities for gene products in both datasets are computed using the *GOSemSim* package (Yu et al., 2010). We performed fourfold cross-validation on both datasets. The value of $k$ for KNN classifier is chosen from the set of odd integers in $\{1,\ldots,13\}$. We repeated the cross-validation 20 times on each dataset and recorded the average classification accuracy for each $k$ value.

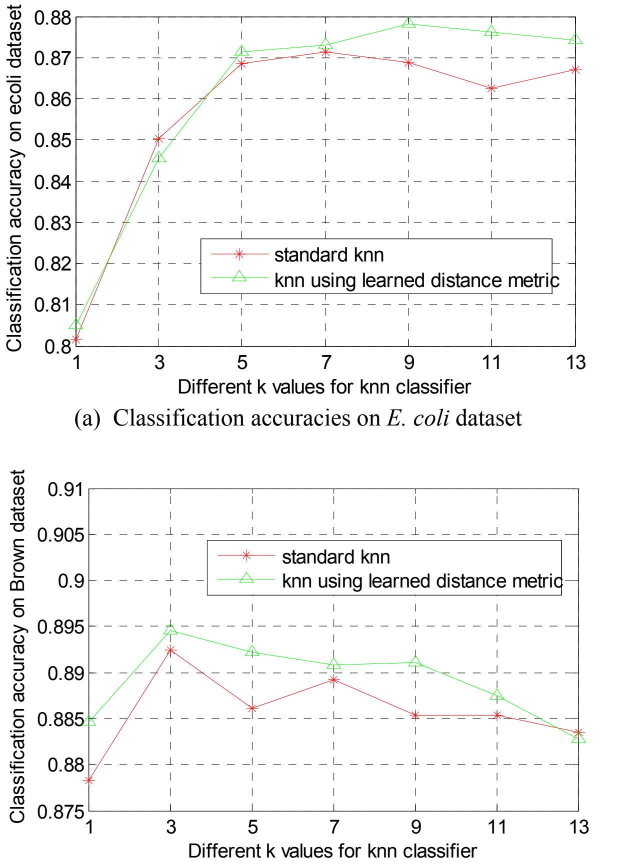
### 4.2 Experimental results

The algorithms were implemented using MATLAB software. The performance was evaluated in terms of classification accuracy. For each method, a confusion matrix C can be constructed based on predicted labels and actual labels. The entry of the confusion matrix $c_{ij}$ represents the number of genes belonging to class $i$ predicted to be of class $j$. Suppose there are $M$ classes in the test set and the classification accuracy for the evaluated method can be computed using equation (15) based on the confusion matrix.

$$\text{classification accuracy} = \frac{\sum_{i=1}^{M} c_{ii}}{\sum_{i=1}^{M}\sum_{j=1}^{M} c_{ij}} \tag{15}$$

Figure 2 (a) shows the classification accuracies of the standard KNN classifier and the improved KNN classifier using the learned distance metric on the *E. coli* dataset. In Figure 2 (a), 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%. Figure 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%.

**Figure 2** Classification accuracies for the standard KNN classifier and the improved KNN classifier using the learned distance metric (see online version for colours)

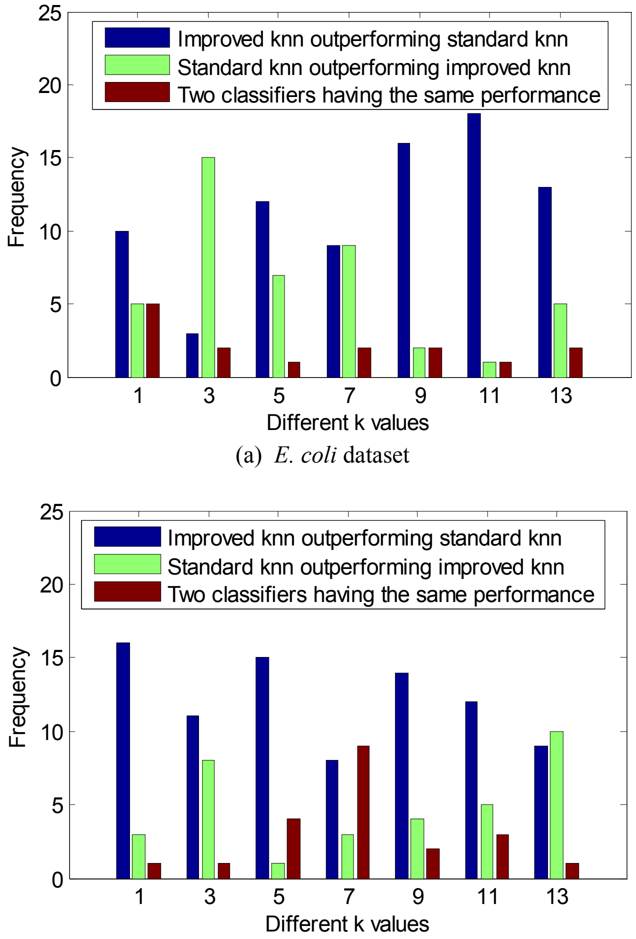


(a) Classification accuracies on *E. coli* dataset

(b) Classification accuracies on Brown's gene expression dataset

Figure 3 shows the frequencies of the three possible cases that could arise in the experiments, i.e. the improved KNN classifier using the learned distance metric outperforms the standard KNN classifier, the standard KNN classifier outperforms the improved KNN classifier and the two classifiers having the same performance. The frequency of a certain case is the number of cross-validation experiments corresponding to this case. In this figure, it can be seen that the performance of improved KNN classifier is better than the standard KNN for most experiments and the semantic similarity information is successfully incorporated into the classification process.

**Figure 3** The results of three different approaches on *E. coli* and Brown's dataset (see online version for colours)



(a) *E. coli* dataset

(b) Brown's gene expression dataset

Notes: Blue bar represents the number of cases in which the improved KNN classifier outperforms the standard KNN classifier; green bar represents the number of cases in which the standard KNN classifier outperforms the improved KNN classifier; and the red bar represents the number of cases in which the two classifiers have the same performance.

## 5 Conclusion

In this paper, we proposed a new method which utilises 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

Asuncion, A. and Newman, D.J. (2013) *UCI machine learning repository*. Available online at: http://archive.ics.uci.edu/ml/

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey T., Ares Jr., M. and Haussler, D. (2000) 'Knowledge-based analysis of microarray gene expression data by using support vector machines', *Proceedings of the National Academy of Sciences USA*, Vol. 97, No. 1, pp.262–267.

Cai, R., Hao, Z., Wen, W. and Huang, H. (2010) 'Kernel based gene expression pattern discovery and its application on cancer classification', *Neurocomputing*, Vol. 73, Nos. 13–15, pp.2562–2570.

Cheng, J., Cline, M., Martin, J., Finkelstein, D., Awad, T., Kulp, D. and Siani-Rose, M.A. (2004) 'A knowledge-based clustering algorithm driven by gene ontology', *Journal of Biopharmaceutical Statistics*, Vol. 14, No. 3, pp.687–700.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M. and Haussler, D. (2000) 'Support vector machine classification and validation of cancer tissue samples using microarray expression data', *Bioinformatics*, Vol. 16, No. 10, pp.906–914.

Guyon, I., Weston, J., Barnhill, S. and Vepnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, Vol. 46, Nos. 1–3,pp.389–422.

Hinton, G., Goldberger, J., Roweis, S. and Salakhutdinov, R. (2004) 'Neighborhood components analysis', *Proceedings of NIPS*, 13–18 December, Vancouver, BC, Canada, pp.513–520.

Jiang, J. and Conrath, D. (1997) 'Semantic similarity based on corpus statistics and lexical taxonomy', *Proceedings of International Conference on Research in Computational Linguistics*, pp.19–33.

Lee, J. and Zhang, C. (2006) 'Classification of gene-expression data: The manifold-based metric learning way', *Pattern Recognition*, Vol. 39, No. 12, pp.2450–2463.

Leung, Y. and Hung, Y. (2010) 'A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 7, No. 1, pp.108–117.

Lin, D. (1998) 'An information-theoretic definition of similarity', *Proceedings of International Conference on Machine Learning*, 24–27 July, Madison, WI, USA, pp.296–304.

Liu, K-H. and Huang, D.S. (2008) 'Cancer classification using rotation forest', *Computers in Biology and Medicine*, Vol. 38, No. 5, pp.601–610.

Nevins, J.R. and Potti, A. (2007) 'Mining gene expression profiles: expression signatures as cancer phenotypes', *Nature Review Genetics*, Vol. 8, No. 8, pp.601–609.

Pandey, G., Myers, C.L. and Kuma, V. (2009) 'Incorporating functional inter-relationships into protein function prediction algorithms', *BMC Bioinformatics*, Vol. 10, pp.142–164.

Pekar, V. and Staab, S. (2002) 'Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision', *Proceedings of the International Conference on Computational Linguistics*, 24 August–1 September, Taipei, Taiwan, pp.786–792 (2002)

Resnik, P. (1999) 'Semantic similarity in taxonomy: an information-based measure and its application to problems of ambiguity in natural language', *Journal of Artificial Intelligence Research*, Vol. 11, pp.95–130.

Schlicker, A., Domingues, F., Rahnenführer, J. and Lengauer, T. (2006) 'A new measure for functional similarity of gene products based on gene ontology', *BMC Bioinformatics*, Vol. 7, p.302.

Schweikert, G., Zien, A., Zeller, G., Behr, J., Dieterich, C., Ong, C.S., Philips, P., De Bona, F., Hartmann, L., Bohlen, A., Krüger, N., Sonnenburg, S. and Rätsch, G. (2009) 'mGene: accurate SVM-based gene finding with an application to nematode genomes', *Genome Research*, Vol. 19, No. 11, pp.2133–2143.

Tao, Y., Sam, L., Li, J., Friedman, C. and Lussier, Y.A. (2007) 'Information theory applied to the sparse gene ontology annotation network to predict novel gene function', *Bioinformatics*, Vol. 23, No. 13, pp.i529–i538.

The Gene Ontology Consortium (2000) 'Gene ontology: tool for the unification of biology', *Nature Genetics*, Vol. 25, No. 1, pp.25–29

Wang, H., Azuaje, F. and Bodenreider, O. (2005) 'An ontology-driven clustering method for supporting gene expression analysis, computer-based medical systems', *Proceedings of IEEE Symposium on Computer-Based Medical Systems*, 23–24 June, Dublin, Ireland, pp.389–394.

Wang, J., Du, Z., Payattakool, R., Yu, P. and Chen, C. (2007) 'A new method to measure the semantic similarity of GO terms', *Bioinformatics*, Vol. 23, No. 10, pp.1274–1281.

Wang, S-L., Zhu, Y., Jia, W. and Huang, D.S. (2012) 'Robust classification method of tumor subtype by using correlation filters', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 9, No. 2, pp.580–591.

Weinberger, K., Blitzer, J. and Saul, L. (2006) 'Distance metric learning for large margin nearest neighbor classification', *Proceedings of NIPS*, 4–7 December, Vancouver, BC, Canada.

Wu, H., Su, Z., Mao, F., Olman, V. and Xu, Y. (2005) 'Prediction of functional modules based on comparative genome analysis and gene ontology application', *Nucleic Acids Research*, Vol. 33, No. 9, pp.2822–2837.

Xing, E., Ng, A., Jordan, M. and Russell, S. (2002) 'Distance metric learning, with application to clustering with side-information', *Proceedings of NIPS*, 9–14 December, Vancouver, BC, Canada, pp.505–512.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y. and Wang, S. (2010) 'GOSemSim: an R package for measuring semantic similarity among GO terms and gene products', *Bioinformatics*, Vol. 26, No. 7, pp.976–978.

Yu, H., Gao, L., Tu, K. and Guo, Z. (2005) 'Broadly predicting specific gene functions with expression similarity', *Gene*, Vol. 352, pp.75–81.

Zare, H., Kaveh, M. and Khodursky, A. (2011) 'Inferring a transcriptional regulatory network from gene expression data using nonlinear manifold embedding', *PLOS ONE*, Vol. 6, No. 8, p.e21969.

Zheng, C-H., Huang, D.S., Zhang, L. and Kong, X-Z. (2009) 'Tumor clustering using non-negative matrix factorization with gene selection', *IEEE Transactions on Information Technology in Biomedicine*, Vol. 13, No. 4, pp.599–607.

Zheng, C-H., Zhang, L., Ng, V.T-Y., Shiu S.C-K. and Huang, D.S. (2011) 'Metasample-based sparse representation for tumor classification', *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, No. 5, pp.1273–1282.