Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Classification analysis of microarray data based on ontological engineering*

LI Guo-qi†, SHENG Huan-ye

(*Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*)

†E-mail: liguoqi@sjtu.edu.cn

Received Feb. 11, 2006; revision accepted Sept. 21, 2006

Abstract: Background knowledge is important for data mining, especially in complicated situation. Ontological engineering is the successor of knowledge engineering. The sharable knowledge bases built on ontology can be used to provide background knowledge to direct the process of data mining. This paper gives a common introduction to the method and presents a practical analysis example using SVM (support vector machine) as the classifier. Gene Ontology and the accompanying annotations compose a big knowledge base, on which many researches have been carried out. Microarray dataset is the output of DNA chip. With the help of Gene Ontology we present a more elaborate analysis on microarray data than former researchers. The method can also be used in other fields with similar scenario.

Key words: Ontological engineering, Data mining, Microarray, Support vector machine (SVM)
doi:10.1631/jzus.2007.A0638 **Document code:** A **CLC number:** TP391

INTRODUCTION

There is a Chinese saying that "Lessons learned from the past can guide one in the future". With the accumulation of knowledge, data mining is not an isolated mission. What we knew should be clear before we explore new knowledge. Information technology has been collaborating with traditional industries extensively and deeply, so data mining needs previous understanding of domain specific knowledge. The sharable knowledge bases built on ontology can be used to provide background knowledge automatically to direct the process of data mining. In the field of bioinformatics, Gene Ontology and the annotations compose a big knowledge base, on which many researches have been carried out. Microarray dataset is the output of DNA chip. With the assistance of Gene Ontology we present a more elaborate analysis on microarray data than former researches. In this section, we first present a brief introduction to ontology, ontological engineering and its application in biology, Gene Ontology. Then the background of microarray gene expression data analysis is described. Finally, a data mining experiment on microarray data based on Gene Ontology is designed and will be described in detail in the rest of the paper.

Ontological engineering and Gene Ontology

From the computational point of view, Ontologies are agreements on shared conceptualizations. Shared conceptualizations include conceptual frameworks for modeling domain knowledge, content-specific protocols for communication among inter-operating agents, and agreements on the representation of particular domain theories (Gruber, 1995). Ontology is used as an explication of knowledge or an organizer of metadata. There are two large differences between the roles of an ontology for knowledge bases and those for metadata: one is philosophical and the other is practical. The philosophical one is that an ontology, for knowledge bases, is a specification of the concep-

* Project (No. 20040248001) supported by the Ph.D. Programs Foundation of Ministry of Education of China

tualization of the target world; and the practical one, for metadata, is a set of computer-understandable vocabulary (Mizoguchi, 2003). In the first kind of cases, ontology works as a system of fundamental concepts, that is, a protocol specification of any knowledge base, explicating the conceptualization of the target world and providing a solid foundation, on which one can build sharable knowledge bases for wider usability than that of a conventional knowledge base. And in the other kind of cases, ontology was used as a tool for data retrieval and exchange between heterogeneous databases.

The Gene Ontology project provides a controlled vocabulary to describe gene and gene product attributes in any organism (Ashburner *et al*., 2000). It was organized into three categories, or have three ontologies: biological process, cellular component and molecular function. There are nearly 166k genes that have been annotated by Gene Ontology. An ontology comprises a set of well-defined terms with well-defined relationships. The structure itself reflects the current representation of biological knowledge as well as serving as a guide for organizing new data. Data can be annotated to varying levels depending on the amount and completeness of available information (Ashburner *et al*., 2000). The original intent of the Gene Ontology project was to construct a set of vocabularies comprising terms that we could share with a common understanding of the meaning of any term used, and that could support cross database queries. It soon became obvious that the combined set of annotations from the model organism groups would provide a useful resource for the entire scientific community. Therefore, in addition to developing the shared structured vocabularies, the Gene Ontology project is developing a database resource that provides access not only to the vocabularies, but also to annotation and query applications and to specialized datasets resulting from the use of the vocabularies in the annotation of genes and/or gene products (The Gene Ontology Consortium, 2001). It became a large knowledge base for molecular biology research. As described above, Gene Ontology is not a knowledge base itself but the databases annotated by Gene Ontology are knowledge bases. The technique of ontology makes it an open, specified and computer-understandable knowledge base which minimizes the gap between computational and biological researchers.

It seems that ontological engineering in biology field is more developed than any other, maybe except semantic web, because of the characteristics of biological research itself. Gene and protein sequence databases have voluminous data and continue to more than double in size every year (Roos, 2001; Benson *et al*., 2006). Facing the huge number of entities and their relationships, biologists resorted to the advanced knowledge management method, ontological engineering, to provide services. So, our example selected bioinformatics research issue. The method can also be sound in other fields with similar scenario.

Introduction of microarray gene expression data and their analysis

With the rapid development of genome-scale sequencing, many genomes have already been known. An essential and formidable task is to define the role of each gene and understand how the genome functions as a whole. As we know, in molecular level, DNA is transcribed into messenger ribonucleic acid or mRNA and then mRNA is translated to produce a protein. In fact, only a small part of segments in the genome exist in the code for genes. Most of the functional roles of other segments are still unknown. It is impossible to discover the mystery just from analyzing the sequence data. To study the relationships of molecular entities on system level, DNA chips were introduced that can simultaneously measure the expression levels of thousands of genes in cell (Allison, 2005). Microarray data are the output of DNA chips. In the microarray dataset, each column represents a gene and each row means a "frozen picture" of their expression level in a series of continuous conditions or samples with different characters. These microarray datasets typically have a large number of columns but a small number of rows. For example, many gene expression datasets may contain up to 10 000~100 000 columns but only 100~1 000 rows (Brown *et al*., 2000), because the number of genes is big and that of cell samples is usually limited.

Initial experiments (Eisen *et al*., 1998) suggest that genes of similar function yield similar expression patterns in microarray experiments. As data from such experiments accumulate, it will be essential to have accurate means for extracting biological significance and using the data to assign functions to genes

(Brown *et al*., 2000). The former researches on this field can be divided into three categories briefly. The earliest method is cluster analysis, which is an unsupervised fashion, learning in the absence of a teacher signal. This kind of methods begin with a definition of similarity, or a measure of distance between expression patterns, but with no prior knowledge of the true functional classes of the genes. Genes are then grouped by using a clustering algorithm such as hierarchical clustering (Eisen *et al*., 1998) or self-organizing maps (Tamayo *et al*., 1999). But the limitation of cluster analysis is obvious. Different kinds of analysis scenarios have to share the same metric of similarity such as Correlation or Euclidean distance. However, the choice of an appropriate distance metric is critical in order to reveal true underlying expression patterns beneath the samples (Phan *et al*., 2004). So researchers resorted to supervised method, such as SVM (support vector machine) (Brown *et al*., 2000) or other statistical methods to achieve more sensitive analysis. Supervised learning techniques use a training set to specify in advance which data should cluster together. As applied to gene expression data, a set of genes that have a common function and a separate set of genes that are known not to be members of the functional class is specified. These two sets of genes are combined to form a set of training examples in which the genes are labelled positively if they are in the functional class and negatively if they are known not to be in the functional class. Using this training set, a classifier would learn to discriminate between the members and non-members of a given functional class based on expression data. Having learned the expression features of the class, the classifier could recognize new genes as members or as nonmembers of the class based on their expression data.

Additional to the cluster and classification method, analyses based on Gene Ontology are also used to explore microarray datasets (Pavlidis *et al*., 2004). A researcher presented a method of clustering lists of genes mined from a microarray dataset using functional information from the Gene Ontology (Kennedy *et al*., 2004). The method uses relationships between terms in the ontology both to build clusters and to extract meaningful cluster descriptions. In essence, the method searches the background knowledge contained in the Gene Ontology and the annotation database to give an organized conclusion to the genes in the microarray in a cluster fashion.

There are other research issues successfully carried out by the aid of Gene Ontology, such as protein subcellular location (Chou and Cai, 2003). But their most active use is in the analysis of microarray data. The reason is the character of microarry data structure. Microarray datasets typically have a large number of columns but a small number of rows. So the relationships of entities in the analysis are more important than those in the analyses on traditional datasets. The problem then is urgent starvation for background knowledge. The ontological method in microarray data analysis is popular and effective. Different from most Gene Ontology researches which typically use artificial intelligence method, such as reasoning and deduction, we pay attention on how ontological engineering provides background knowledge and directs the process of data mining.

Finding the unknown function of genes in a microarray dataset with data mining

Genes of similar function are known to yield similar expression patterns in microarray experiments, with supervised learning techniques having been successfully used in the analysis of microarray datasets. So we can now utilize the background knowledge in the Gene Ontology, mainly using the biological process component, and its annotation for deeper and more accurate research into microarray datasets. The reason why the biological process ontology was selected lies in the character of the test dataset, which will be detailedly described in Section 2. We first search every gene biological process in the Gene Ontology and draw the subgraph of the ontology with node only related to the genes concerned. The subgraph can be seen as a taxonomic tree according to the biological issues of the genes. Every node is annotated with a biological process name and the genes set with the function. From the global background knowledge we can construct a set of classifiers to predict unknown functions of genes in the microarray dataset on different levels.

DATASET AND ITS BACKGROUND KNOWLEDGE IN GENE ONTOLOGY

To illustrate our method, we use a microarray dataset that had been used in many similar researches. The dataset has 79-element gene expression vectors for 2467 yeast genes. The data were generated from spotted arrays using samples collected at various time points during the diauxic shift, the mitotic cell division cycle, sporulation, temperature and reducing shocks (http://rana.stanford.eduyclustering) (Brown *et al*., 2000).

Before data analysis, normalization is an important step with microarray data. Conveniently, there are standard algorithms and software for solving the problem. Cluster3, an open source tool (available at http://rana.lbl.gov/EisenSoftware.htm), can be used to normalize microarray data. The algorithms are described in detail in the attached document of the software.

After normalization, input the genes into Gene Ontology and its annotation database, we got the subgraph of the ontology with node only related to the genes of interest. The information can also be represented as a gene-category matrix. In the matrix, each column means a microarray gene, which has also been annotated in the category of biological process of Gene Ontology with each row of delegates representing a function term in the hierarchy of Gene Ontology. We ignore the relationship between the nodes and view the result as a taxonomic tree in molecular function issue. The information ignored will be reconsidered in the biological analysis in the following research. Part of the subgraph is illustrated in Fig.1. We select four nodes marked with asterisk to carry out our analysis. In fact, only the four nodes and their upper level nodes are shown in Fig.1.

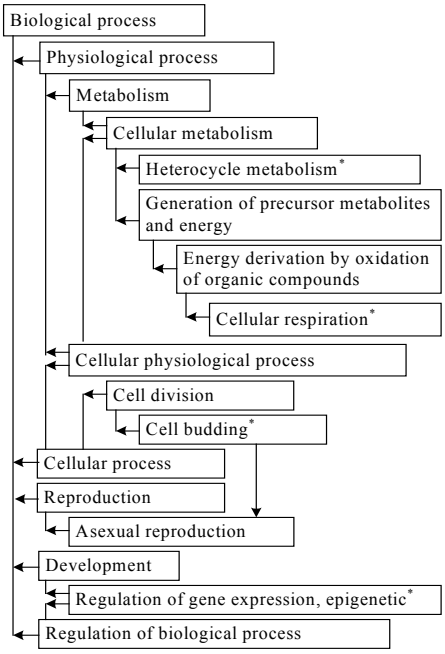

Fig.1 Part of subgraph of obtained gene ontology. The four nodes marked with asterisk are selected to carry out our analysis

Table 1 Categories and number of samples

Category ID	Process name	Number of samples
1	Cellular respiration	63
2	Heterocycle metabolism	60
3	Regulation of gene expression, epigenetic	60
4	Cell budding	63

CLASSIFICATION STRATEGIES

Four biological processes were selected. The biological process names and the number of genes annotated to each node are shown in Table 1. Every two classes are assembled as a group. Six independent binary classifiers were distributed to each group respectively. We divided the samples into two parts, one of which is used for training the classification and the other used for testing the accuracy of the corresponding classifier. Then we can answer such questions as whether we can predict that the gene participates in the biological process ID 1 or ID 2, if we knew it participated in one of them.

SVM is selected as the classifier in (Burges, 1998), as it has proven to have high performance in the classification of microarray data (Brown *et al*., 2000). Practically, we used LIBSVM, which is an integrated software for support vector classification, (C-SVC, nu-SVC), regression (epsilon-SVR, nu-SVR) and distribution estimation (one-class SVM) (Chang and Lin, 2001).

Before training the SVM, we must decide which kernel should be selected and then the penalty parameter C and kernel parameters are chosen. The RBF kernel $K(x,y)=\exp(-r\|x-y\|^2)$ had proven to have

many advantages (Burges, 1998; Chang and Lin, 2001) and is suitable for microarray data analysis (Brown *et al*., 2000). After selection of kernel, five-fold cross-validation was used to find the best C and γ. At last, the best parameters were used to train and test the classifiers by five-fold cross-validation. The result is shown in Table 2.

Table 2 The best parameters and classification accuracy with five-fold cross-validation

Classifier	Best C	Best γ	Average accuracy (%)
1-2	8.0	0.125	81.3008
1-3	3.0	0.125	91.0569
1-4	32.0	0.0078125	84.1270
2-3	32.0	0.0001220703125	70.8333
2-4	8.0	0.03125	81.3008
3-4	2.0	0.03125	74.7967

RESULT

The result shows that microarray data can be used to predict the function of genes in biological process. Data in Table 2 show that the classifier with category ID 1 has high accuracy. Because the biochemical experiment was carried out during a period of time with temperature changes, the regulation of cellular respiration exhibits their existence more notably. The accuracy of classifier 1-3 is higher than that of 1-2 or 1-4 because "cellular respiration" and "regulation of gene expression, epigenetic" are more unrelated than the other two couples. The result is reasonable because of its explanations by biological mechanism (Brown *et al*., 2000).

CONCLUSION

Data mining based on ontological engineering has many advantages. There have been many successful cases with the method. This paper gives a common introduction to it and presents a practical analysis example using SVM (support vector machine) as classifier. Our research mainly focused on microarray data analysis, as the application of ontological engineering is relatively mature in the field. The method can also be used in other fields with similar scenario. With the accumulation of knowledge and data, the application of ontological engineering and data mining based on it will be more popular. We will focus our following researches on this field.

References

Allison, D.B., 2005. DNA Microarrays and Related Genomic Techniques: Statistical Design, Analysis, and Interpretation of Experiments. Chapman & Hall/CRC, p.5-9.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherr, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al*., 2000. Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**:25-29. [doi:10.1038/75556]

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2006. GenBank. *Nucleic Acids Research*, **34**(Database issue):16-20. [doi:10.1093/nar/gkj157]

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.Jr, Haussler, D., 2000. Knowledge-based analysis of microarray gene expression data using support vector machines. *PNAS*, **97**(1):262-267. [doi:10.1073/pnas.97.1.262]

Burges, C., 1998. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**(2):121-167. [doi:10.1023/A:1009715923555]

Chang, C.C., Lin, C.J., 2001. LIBSVM: A Library for Support Vector Machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chou, K.C., Cai, Y.D., 2003. A new hybrid approach to predict subcellular localization of proteins by incorporating Gene Ontology. *Biochem. Biophys. Research Commun.*, **311**(3):743-747. [doi:10.1016/j.bbrc.2003.10.062]

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**(25):14863-14868. [doi:10.1073/pnas.95.25.14863]

Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Human-Computer Studies*, **43**(5-6):907-928. [doi:10.1006/ijhc.1995.081]

Kennedy, P.J., Simoff, S.J., Skillicorn, D.B., Catchpoole, D., 2004. Extracting and Explaining Biological Knowledge in Microarray Data. Pacific-Asia Conference on Knowledge Discovery and Data Mining, p.699-703.

Mizoguchi, R., 2003. Tutorial on ontological engineering—Part 1: introduction to ontological engineering. *New Generation Computing*, **21**(4):365-384.

Pavlidis, P., Qin, J., Arango, V., Mann, J.J., Sibille, E., 2004. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical Research*, **29**(6):1213-1222. [doi:10.1023/B:NERE.0000023608.29741.45]

Phan, J.H., Quo, C.F., Guo, K.J., Feng, W.M., Wang, G., Wang, M.D., 2004. Development of a Knowledge-based

Multi-scheme Cancer Microarray Data Analysis System. Proc. 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04), p.474-475.

Roos, D.S., 2001. Bioinformatics—trying to swim in a sea of data. *Science*, **291**:1260-1261. [doi:10.1126/science.291.5507.1260]

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E., Golub, T., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, **96**(6):2907-2912. [doi:10.1073/pnas.96.6.2907]

The Gene Ontology Consortium, 2001. Creating the Gene Ontology resource: design and implementation. *Genome Res.*, **11**(8):1425-1433. [doi:10.1101/gr.180801]