Integrative Gene Selection for Classification of Microarray Data

Ong Huey Fang (Corresponding author), Norwati Mustapha & Md. Nasir Sulaiman

Department of Computer Science, Faculty of Computer Science and Information Technology

University Putra Malaysia, 43400 Serdang, Selangor Darul Ehsan, Malaysia

Tel: 60-3-8946-8946-6585 E-mail: ong.huey.fang@gmail.com

Tel: 60-3-8946-8946-6585 E-mail: norwati@fsktm.upm.edu.my

Tel: 60-3-8946-8947-1702 E-mail: nasir@fsktm.upm.edu.my

Abstract

Microarray data classification is one of the major interests in health informatics that aims at discovering hidden patterns in gene expression profiles. The main challenge in building this classification system is the curse of dimensionality problem. Thus, there is a considerable amount of studies on gene selection method for building effective classification models. However, most of the approaches consider solely on gene expression values, and as a result, the selected genes might not be biologically meaningful. This paper presents an integrative gene selection for improving microarray data classification performance. The proposed approach employed the association analysis technique to integrate both gene expression and biological data in identifying informative genes. The experimental results show that the proposed gene selection outperformed the traditional method in terms of accuracy and number of selected genes.

Keywords: Association analysis, Classification, Gene selection, Integrative, Microarray

1. Introduction

DNA microarray is an ordered collection of genes, which usually printed onto a small glass slide. It is a high throughput technology that allows the expression levels up to thousands of genes to be assay simultaneously. The patterns of gene expression can be measured under different physiological conditions to provide important clues of gene functions. Therefore, DNA microarray has become an indispensable research tool for understanding the underlying genetic causes of many human diseases. The challenges now are finding ways to organize and analyse this high throughput data effectively.

Today, the analysis of gene expression data has been a topic of interest in the field of health informatics (Ng & Pei, 2007). Classification is one of the main types of microarray analysis that allows the discovery of hidden patterns in expression profiles and opens the possibility for more accurate disease diagnosis. For instance in cancer classification, microarray is used to classify tumor samples into different classes. Previous studies have shown that microarray data can be used to differentiate between normal and cancerous tissues (Alon, et al., 1999; Pau Ni, et al.), to classify multiclass cancer subtype (Rifkin, et al., 2003; Yeang, et al., 2001), and even to identify new cancer subtypes (Golub, et al., 1999). Having good cancer classification is crucial in order to give the most effective and cost saving treatments for patients (Soh, et al., 2007).

The major concern in classification of microarray data is the curse of dimensionality problem, where there is a large number of genes compared to small samples sizes (Piatetsky-Shapiro & Tamayo, 2003). This is because high dimensionality datasets will usually reduce the accuracy and speed of classification systems (Shang & Shen, 2006). To overcome this, a considerable amount of gene selection approaches have been proposed to identify differentially expressed genes for building effective classification models (Inza, et al., 2004; Saeys, et al., 2007; Shang & Shen, 2006). However, most of the methods are statistical analyses and based only on gene expression measurements. There are disadvantages by just considering expression data alone, as the values may not be measured accurately and the complexity of microarray experiments may cause discrepancy in data obtained. Moreover, statistical significance might not be able to directly translate to biological relevance (Liu, et al., 2005), causing non-informative genes being selected. Besides that, they give little information about the relationship among genes and provide few biological insights on the underlying mechanisms for target classes.

In recent years, integrative analysis has emerged as an attractive approach for mining microarray data. The term “integrative analysis” is described as the analysis of high throughput data in the context of available biological knowledge (Fellenberg, 2003). Therefore, more efforts are directed towards developing integrative classification systems that consider gene expression data along with additional functional annotations such as gene ontology and metabolic pathways. Yet, most of the existing integrative approaches grouped the genes based on expression values and biological information is only used as a subsequent process to the analysis of expression data

(Carmona-Saez, et al., 2006). The main drawback to this type of approach is that finding co-expressed genes alone are unable to reflect the underlying complex relationships among genes. Hence, in this study, we aim to group genes not only by considering their expression patterns but also their additional biological properties. Association analysis is one of the techniques that able to integrate different kinds of data into a single dataset for discovering interesting relationships. This paper presents an integrative gene selection for improving classification performance in terms of accuracy and number of selected genes. The proposed approach applied association analysis technique to integrate both microarray data and gene annotations in identifying sets of informative genes.

The remainder of this paper is organized as follows. Section 2 covers the existing works on integrative gene selection method for microarray data classification. Section 3 introduces the biological information that can be used for microarray analysis. Section 4 describes the proposed integrative gene selection based on association analysis. The experiments and results are discusses in Section 5. Finally, Section 6 concludes our study and discusses on some possible future directions.

2. Related works

One of the well-known gene selection methods is filter method. The main drawback of filter method is that it evaluates each gene individually without considering the interactions among them. Therefore it usually combined with other methods to compensate its disadvantages. The most common combination is hybrid model of filter and wrapper techniques as in Yang, et al. (2008). Although wrapper method is capable of finding optimal gene subset, but it is classifier dependent and has higher risk of over fitting the data (Saeys, et al., 2007). In another work, Wang, et al. (2005) proposed a hybrid approach that combines filter method and clustering analysis. However, opposed to the fact that a gene can interacts with several other genes, clustering techniques can only group each gene to one of the clusters. Besides that, though clustering is able to associate genes with similar properties into the same cluster, it is unable to show the relationship between genes within the cluster.

Currently there are few studies that apply integrative analysis approach for microarray data classification. But most of the suggested approaches evaluate genes individually with their expression value and incorporate external biological information as a subsequent step to verify the selected genes. For instance, Horng, et al., (2009) had developed an expert system that is able to identify a small set of gene markers for microarray data classification. With their method, genes that occur regularly in the decision tree play an important role in training models. The top ranked genes are then mapped to KEGG pathway to gain information of the marker genes. The chosen genes form the final set of genes that will be input for classification system. Though the classification system had shown better results compared to other non-integrative system such as HykGene (Wang, et al., 2005) and GEMS (Alexander, et al., 2005), but the assumption of genes with correlatable expression profiles also share similar biological properties may not always be true (Carmona-Saez, et al., 2006). In another works, Qi & Tang (2007) had proposed a GO based method that integrates gene ontology annotations to select informative genes. This method first calculates the individual discriminative power for each candidate gene by using traditional filtering methods such as the information gain. Then, the GO terms for these genes are retrieved from the public database. The results showed that the GO based method outperformed the tradition expression-only model and have a simple structure that can be conveniently constructed from other gene selection algorithms. However, the GO model did not outperformed in all cancer datasets.

Therefore, expression profiles are integrated with additional biological knowledge and processed together to reveal the significant relationships among genes. By doing so, we are able to identify a subset of genes that is the most informative to differentiate different phenotypes. One of the approaches suggested by Carmona-Saez, et al. (2006) is association rule discovery data mining technique to incorporate gene annotations with expression data in a single framework for uncovering the biological connections among functional annotations and expression patterns. The results showed that the uncovered biologically meaningful associations are supported by previous studies. The drawbacks of this study are huge amount of rule being generated and the manual process to examine the rules.

In addition to these studies, there are others studies that not only measure the genes relationships but also took into considerations the relationships between genes and their functional annotations categories**.** For example, Huang & Chow (2007) presented a classification model that is able to identify relevant gene functional categories to be integrated together with gene expression data. The approach considers four correlation values namely gene-gene, gene-response, category-response and category-category to select relevant gene functional categories. On the other hand, Trajkovski & Lavrač (2007) had proposed a method to generate new enriched gene sets that can capture the biology characteristics for given target class. This works combines the existing gene sets with GO term determined by Gene Set Enrichment Analysis (Subramanian, et al., 2005) added with gene-gene interaction data from Entrez database. Enrichment scores are computed for each of the newly generated gene sets. The experimental results showed that the method can find descriptions for enriched gene sets which highlight the underlying biology that is responsible for distinguishing target classes. Although there are tools that can be used to automatically retrieve relevant biological categories, but the process of integrating different types of biological data have to be done separately and manually. The procedures of integrating the datasets and running classifications are usually done on different systems and often required a lot of user interventions.

3. Biological information

Today, with the fast growing of information technology most of the biological information can be obtained freely from web-based database resources. Biological discoveries and works made by scientists can be recorded in electronic format and share through open and community-maintained knowledge bases. Example of information available is such as DNA sequences, gene and protein interactions, functional ontologies and metabolic pathways. This information can significantly complement any data analysis and improve its results (Bellazzi & Zupan, 2007). Table 1 shows a list of molecular biology databases available on the web.

Each molecular biology database provides different types of biological information at different levels, such as at the gene level, gene location, clone level, protein level and functional level. For instance, the Affymetrix database provides the gene information in GeneChip arrays, the GeneBank database provides the sequence information of all publicly available DNA and protein sequences, and the Gene Ontology (GO) is a controlled biological vocabulary that can be used to annotate genes for all species. In this paper, the biological information from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database is incorporated into the gene selection process. One of its database is known as the KEGG Pathway that provides the metabolic pathways information for metabolism, genetic information processing, environmental information processing, cellular processes, human diseases and drug development. In order to access this biological information, a unique identifier (ID) is needed to retrieve its corresponding database entries. One of the ways to map the genes in microarray data to particular pathways is by using the KEGG API. It is the SOAP/WSDL interface for KEGG, which enable users to write their own programs to access and utilize the resource (Kanehisa, et al., 2006).

4. An integrative gene selection based on association analysis

Gene selection method is introduced to reduce the number of selected genes by removing irrelevant, redundant, or noisy data (Li, et al., 2007). In microarray data analysis, gene selection or also known as feature selection is an indispensable task to identify differentially expressed genes and to remove irrelevant genes in high dimensional datasets. Integrative gene selection involves the merging of different kind of data from multiple sources into a single file that is appropriate for mining. In this paper, we proposed a gene selection method based on association analysis to find interesting relationships among genes for selecting informative genes.

4.1 Association analysis

Association analysis is a data mining methodology for discovering interesting relationships hidden in large datasets. The uncovered relationships may reveal interesting connections among attributes, which can be represented in the form of association rules or sets of frequent itemsets (Tan, et al., 2005). Such information is helpful for developing understanding and decision making in application domains. For example in market basket analysis, association analysis is often used to identify items that are frequently purchased together. In microarray data analysis, association analysis can be used to identify group of genes that are likely to co-occur in target samples (Han & Kamber, 2001). Frequent itemset mining is one of the approaches used to find groups of genes that have related functionality. In this paper, we proposed a new way of constructing transactional data to extract interesting associations among discriminative genes based on their biological properties. The selected discriminative frequent itemsets will then be used to construct better classification model.

Finding frequent itemsets is one of the major steps in association analysis to identify interesting sets of items. Relevant gene subsets are selected with the assumption that genes sharing similar biological properties (annotations) also work together in certain cell conditions. Filter method can be applied on preprocessing phase to ensure only discriminative genes will be considered when generating the frequent itemsets. Figure 1 shows the general pseudo-code for the Apriori algorithm (Han & Kamber, 2001) to generate frequent itemsets.

4.2 The construction of transactional data

Depending on type of associations to be extracted, a microarray dataset is transformed into a transactional data format before it can be further analysed using association analysis. The transactional data is constructed from the microarray data and annotation data. From the microarray data, a list of genes is extracted, where their discriminative scores are not equal to zero. Then, their annotations are retrieved from the biological database. In this study, we aim at extracting interesting relationships between genes based on their biological properties. Therefore a transaction database is created based on gene annotations, where the genes are treated as "items" and the annotations as "transactions". Figure 2 illustrates the construction of a transactional data by using the KEGG pathway annotations. The transactional data can be further transformed to include genes discriminative scores as shown in Table 2. By doing this, it helps to retain the information from gene expression profiles and further sorts the genes within the transactional data for generating frequent itemsets and in identifying representative genes. Discriminative score shows the importance of each gene with respect to class labels and can be computed using any traditional filtering method such as the information gain. Given a set of samples *S* in *k* classes, the entropy (or information-content) is defined as:

$$Ent\ (S) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

where p_i is the probability of class *i* in *S*. Then, let selects an interval boundary *T* for gene X_i, which partition the samples into subsets S_1 and S_2. The entropy resulting from the partitioning is:

$$Ent\ (X_i,\ T;\ S) = \frac{|S_1|}{S} Ent(S_1) + \frac{|S_2|}{S} Ent(S_2)$$

where $Ent\ (S_j),\ (j = 1,\ 2)$ corresponds to the entropy for a subset of *S* according to the definition:

$$Ent\ (S_j) = -\sum_{i=1}^{k} P(C_i, S_j) \log_2(P(C_i, S_j))$$

where $P(C_i,\ S_j)$ is the probability of samples in S_j that have class C_i. Finally, the information gain value (Witten, and Frank, 2005) for gene X_i can be calculated as follows:

$$InfoGain\ (X_i,\ T;\ S) = Ent\ (S) - Ent\ (X_i,\ T;\ S)$$

4.3 Gene ranking

Traditional gene ranking method is used to rank frequent itemsets based on their interestingness and to form potential gene subsets. The ranking mechanism takes the discriminative scores for genes in an itemset and then calculates an interestingness score for that itemset. The interestingness for a frequent itemset is defined as the average value of discriminative scores for all the genes in them. The generated frequent itemsets are ranked according to their interestingness scores in descending order. The most top-ranked frequent itemsets is considered as the most informative itemset, and its representative gene is considered the most informative genes. A representative gene in an itemset is the item with the highest discriminative score. Gene subsets are then generated from the representative genes of the top-ranked frequent itemsets, and only the most informative gene subset will be inputted for building the classification model. The best gene subset is the set of genes that can achieve the highest predictive accuracy with less number of genes. Following is the high level description of the proposed approach for ranking frequent itemsets and finding the optimal gene subset.

1. Calculate the discriminative scores for all genes in microarray dataset.
2. Retrieve annotations for these genes. Genes without annotation will be ignored.
3. Create a transactional data with annotation as "transactions" and genes as "items". Gene with discriminative score equals to zero will be ignored when generating the transactional data and repetitive genes in a transaction will be removed.
4. Run association algorithm with specific minimum support to find frequent itemsets.
5. Calculate the interestingness of frequent itemsets that are averages of discriminative scores of genes contain in them.
6. Rank frequent itemsets by their interestingness and the itemset with the highest values is considered as the most informative frequent itemset.
7. Find the representative gene from each frequent itemset, which is the highest discriminative score gene from the itemset.

8. Output these representative genes from m top-ranked frequent itemsets and use them as gene subset to compute the leave-one-out cross validation (LOOCV) classification accuracy for microarray data.
9. Output the smallest set of representative genes corresponding to the best LOOCV.

5. Experiments

In order to evaluate the performance of the proposed approach, several experiments have been conducted. In the experiments, we used WEKA application (Witten & Frank, 2005) to run the data preprocessing and classifications. Two criteria to evaluate the effectiveness of the proposed approach are the number of selected genes and predicative accuracy. We aim to select the smallest number of genes that can achieve the highest predictive accuracy.

5.1 Datasets and preprocessing

Two publicly available microarray datasets of cancer research are used in the experiments. These include colon cancer (Alon, et al., 1999) and breast cancer (van 't Veer, et al., 2002) datasets. Table 3 shows the details of the datasets. The missing values in the datasets are replaced by the mean value of the gene. Then, the microarray data are standardized and discretized using entropy-based discretization method (Fayyad & Irani, 1993). The information gain method was used to calculate discriminative scores for all the genes in the microarray. Genes with discriminative scores equal to zero are removed.

With the remaining genes, we retrieved their biological information from KEGG pathway database (Kanehisa, et al., 2006). Each KEGG pathway annotation has a unique KEGG identifier (KEGG ID) and contains information such as name, descriptions and gene they annotated. The annotations can be retrieved using gene identifier (gene ID) or gene symbol as defined in the related database. Genes without annotations will be ignored in the subsequent process. Thus, this paper has suggested another way of reducing the dimensionality by removing the non-informative genes, or in other words genes that did not have any biological information recorded. Table 4 shows the summary of gene annotations collected for both colon and breast cancer datasets. From the table, we can see that the number of genes is reduced for more than 95% compared to the original size of dataset.

5.2 Experiment setups

We implemented association analysis using Apriori algorithm (Agrawal, et al., 1993). The minimum support is set to 1% when generating the large frequent itemsets. The stopping criterion for searching the optimal genes subset is set to 100. Selected gene subset is input to the classification system. Three classifiers are imported from the WEKA packages (Witten & Frank, 2005) to evaluate the performance of the proposed approach. The classifiers include Naïve Bayes (NB), Support Vector Machine (SVM), and Logistic Regression (LR).

5.3 Experiment results and discussion

In this section we compare the performance of proposed approach with that traditional gene ranking method that considers expression values only. Two criteria to evaluate effectiveness of the approach are the number of selected genes and predicative accuracy. The best result is to select the smallest number of genes which can achieve the highest predictive accuracy. In expression-only methods, the genes are ranked by their information gain values, while for integrative approach, the genes subsets are ranked by the proposed algorithm. Among the top ranked gene subsets, we select the one that will produce the highest accuracy for each classifier used. The accuracy is obtained using the LOOCV. The results for both datasets are shown in Table 5 and Table 6.

In the result for colon cancer dataset, the proposed method outperformed the traditional methods with accuracy as high as 93.55% and smaller number of selected genes. The best subset contains only 2 genes: M26383 (Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds) and J02854 (Myosin regulatory light chain 2, smooth muscle isoform (human); contains element TAR1 repetitive element). Both genes involved in pathways which are able to discriminate between normal and colon cancer tissues. For instance, J02854 is muscle-related gene that reflects normal colon tissue had higher muscle content compare to colon cancer tissue (Fu & Fu-Liu, 2005), and M26383 is found related to pathway in cancer (Kanehisa, et al., 2006). For breast cancer dataset, the proposed method exceeds the expression-only method in the SVM classifier with accuracy 95.88%. Moreover, the number of selected genes is still less than 50.

In summary, the experimental results demonstrate that integrative based algorithm is indeed more effective than those approaches that did not incorporate any additional biological information. Moreover the experiments show that KEGG pathways are suitable to be integrated with microarray data for identifying gene markers for cancer classification purpose.

6. Conclusion

In this paper, we proposed an integrative gene selection for microarray data classification by applying association analysis techniques. The proposed approach is able to integrate both microarray data and additional biological data to identify informative genes. Association analysis technique is applied in the selection process to find the relationships among genes. The experimental results have shown that integration is a right strategy for improving classification system in term of number of selected genes and classification accuracy. Our method also provides a solution for eliminating redundancy by removing genes without annotations, and by grouping together genes with similar biological pathways. For future works, more well-defined association analysis technique will be introduced into the integrative classification system. Moreover, different types of biological information will be incorporated with gene expression data in order to build more effective and informative classification models.

References

Agrawal, R., Imieliński, T., & Swami, A. (1993). *Mining association rules between sets of items in large databases.* Paper presented at the Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C., United States.

Alexander, S., Ioannis, T., Yerbolat, D., & Constantin, F. A. (2005). GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data. *International journal of medical informatics, 74*(7), 491-503.

Alon, U., Barkai, N., Notterman, D. A., Gishdagger, K., Ybarradagger, S., Mackdagger, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences of the United States of America, 96*(12), 6745-6750.

Bellazzi, R., & Zupan, B. (2007). Towards knowledge-based gene expression data mining. *Journal of Biomedical Informatics, 40*(6), 787-802.

Carmona-Saez, P., Chagoyen, M., Rodriguez, A., Trelles, O., Carazo, J. M., & Pascual-Montano, A. (2006). Integrated analysis of gene expression by association rules discovery. *BMC Bioinformatics, 7*(1), 54.

Fayyad, & Irani. (1993). *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning.* Paper presented at the Proceedings of the International Joint Conference on Uncertainty in AI.

Fellenberg, M. (2003). Developing integrative bioinformatics systems. *BIOSILICO, 1*(5), 177-183.

Fu, L. M., & Fu-Liu, C. S. (2005). Evaluation of gene importance in microarray data based upon probability of selection. *BMC Bioinformatics, 6*(1), 67.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science, 286*(5439), 531-537.

Han, J., & Kamber, M. (2001). *Data Mining: Concepts and Techniques*: Morgan Kaufmann.

Horng, J. T., Wu, L. C., Liu, B. J., Kuo, J. L., Kuo, W. H., & Zhang, J. J. (2009). An expert system to classify microarray gene expression data using gene selection by decision tree. *Expert Systems with Applications, 36*(5), 9072-9081.

Huang, D., & Chow, T. W. S. (2007). Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer. *Bioinformatics, 23*(12), 1503-1510.

Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial Intelligence in Medicine, 31*(2), 91-103.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., et al. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucl. Acids Res., 34*(suppl_1), D354-357.

Li, Z. J., Zhang, L. J., & Chen, H. W. (2007). *Are filter methods very effective in gene selection of microarray data?* Paper presented at the IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW2007, Fremont, CA.

Liu, H., Dougherty, E. R., Dy, J. G., Torkkola, K., Tuv, E., Peng, H., et al. (2005). Evolving Feature Selection. *IEEE Intelligent Systems, 20*(6), 64-76.

Ng, R. T., & Pei, J. (2007). Introduction to the special issue on data mining for health informatics. *SIGKDD Explorations Newsletter, 9*(1), 1-2.

Pau Ni, I. B., Zakaria, Z., Muhammad, R., Abdullah, N., Ibrahim, N., Aina Emran, N., et al. (2010). Gene expression patterns distinguish breast carcinomas from normal breast tissues: The Malaysian context. *Pathology - Research and Practice, 206*(4), 223-228.

Piatetsky-Shapiro, G., & Tamayo, P. (2003). Microarray data mining: facing the challenges. *SIGKDD Explorations Newsletter, 5*(2), 1-5.

Qi, J., & Tang, J. (2007). *Integrating gene ontology into discriminative powers of genes for feature selection in microarray data.* Paper presented at the Proceedings of the 2007 ACM Symposium on Applied Computing, Seoul, Korea.

Rifkin, R., Mukherjee, S., Tamayo, P., Ramaswamy, S., Yeang, C. H., Angelo, M., et al. (2003). An Analytical Method for Multiclass Molecular Cancer Classification. *SIAM Review, 45*(4), 706-723.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Shang, C., & Shen, Q. (2006). Aiding classification of gene expression data with feature selection: a comparative study. *International Journal of Computational Intelligence Research, 1*(1), 68-76.

Soh, D., Dong, D., Guo, Y., & Wong, L. (2007). Enabling more sophisticated gene expression analysis for understanding diseases and optimizing treatments. *SIGKDD Explorations Newsletter, 9*(1), 3-13.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America, 102*(43), 15545-15550.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*: Addison Wesley.

Trajkovski, I., & Lavrač, N. (2007). *Interpreting Gene Expression Data by Searching for Enriched Gene Sets.* Paper presented at the Proceedings of the 11th Conference on Artificial Intelligence in Medicine, Amsterdam, Netherlands.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature, 415*(6871), 530-536.

Wang, Y., Makedon, F. S., Ford, J. C., & Pearlman, J. (2005). HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. *Bioinformatics, 21*(8), 1530-1537.

Witten, I., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*: Morgan Kaufmann.

Yang, C. S., Chuang, L. Y., Ke, C. H., & Yang, C. H. (2008). A Hybrid Feature Selection Method for Microarray Classification. *IAENG International Journal of Computer Science, 35*(3), 285-290.

Yeang, C. H., Ramaswamy, S., Tamayo, P., Mukherjee, S., Rifkin, R. M., Angelo, M., et al. (2001). Molecular classification of multiple tumor types. *Bioinformatics, 17*(suppl_1), S316-322.

Table 1. Molecular biology databases

Databases	**Valid Values**	**Descriptions**
Affymetrix	Affymetrix Probe Set ID	The Affymetrix table in the Gene Database contains probe set ID's for all currently available GeneChip arrays.
Entrez Gene	Entrez Gene ID	Curated sequence and descriptive information about genetic loci, and provides an integrated cross-referencing system between sources.
GeneBank	GeneBank Accession ID	Sequence database, an annotated collection of all publicly available DNA and protein sequences. Part of International Nucleotide Sequence Database Collaboration
Gene Ontology	GO ID	The GO consortium is developing structured, controlled vocabularies (ontologies) that describe gene products.
KEGG	KEGG ID	Metabolic and regulatory pathways.
UniGene	UniGene Cluster ID	Contains sets of non-redundant gene-oriented sequence clusters. It is created through automatic partitioning of GenBank sequences, and each UniGene cluster represents a unique gene.

Table 2. Transactional data with gene discriminative scores

path:hsa03050	T54276 : 0.182					
path:hsa03040	M15841: 0.168	X12466 : 0.169	U30825 : 0.28	R84411 : 0.245		
path:hsa03010	T58861 : 0.216	T57619 : 0.245	T61609 : 0.169	T72879 : 0.154	X55715 : 0.184	U14971: 0.182
path:hsa04514	X53586 : 0.235					
path:hsa04670	J02854 : 0.315					
path:hsa04060	M26383 : 0.435					
path:hsa05200	X53586 : 0.235	T51023 : 0.169	M26383 : 0.435			
path:hsa04062	M26383 : 0.435					
path:hsa04640	X53586 : 0.235					
……………..	……………	……………	……………	……………	……………	……………

Table 3. Summary of collected microarray datasets

Datasets	Number of Samples	Number of Genes	Samples per class	
Colon Cancer	62	2000	Tumor	Normal
			40	22
Breast Cancer	97	24481	Relapse	Non-relapse
			46	51

Table 4. Summary of collected biological information from KEGG database

Dataset	Number of Retrieved KEGG Annotation IDs	Number of Genes Annotated *(discriminative score ≠ 0)*	Number of Gene Reduced	Percentage of Genes Reduced
Colon Cancer	61	46	1945	97.7%
Breast Cancer	142	195	24286	99.2%

Table 5. Results on colon cancer dataset

Classifier	Expression-Only (Information Gain)		KEGG *(Information Gain + Association Analysis)*	
	50 top-ranked genes	100 top-ranked genes	Accuracy	Number of genes
NB	90.32%	91.94%	93.55%	2
LR	77.42%	80.65%	93.55%	2
SVM	91.94%	88.71%	93.55%	2

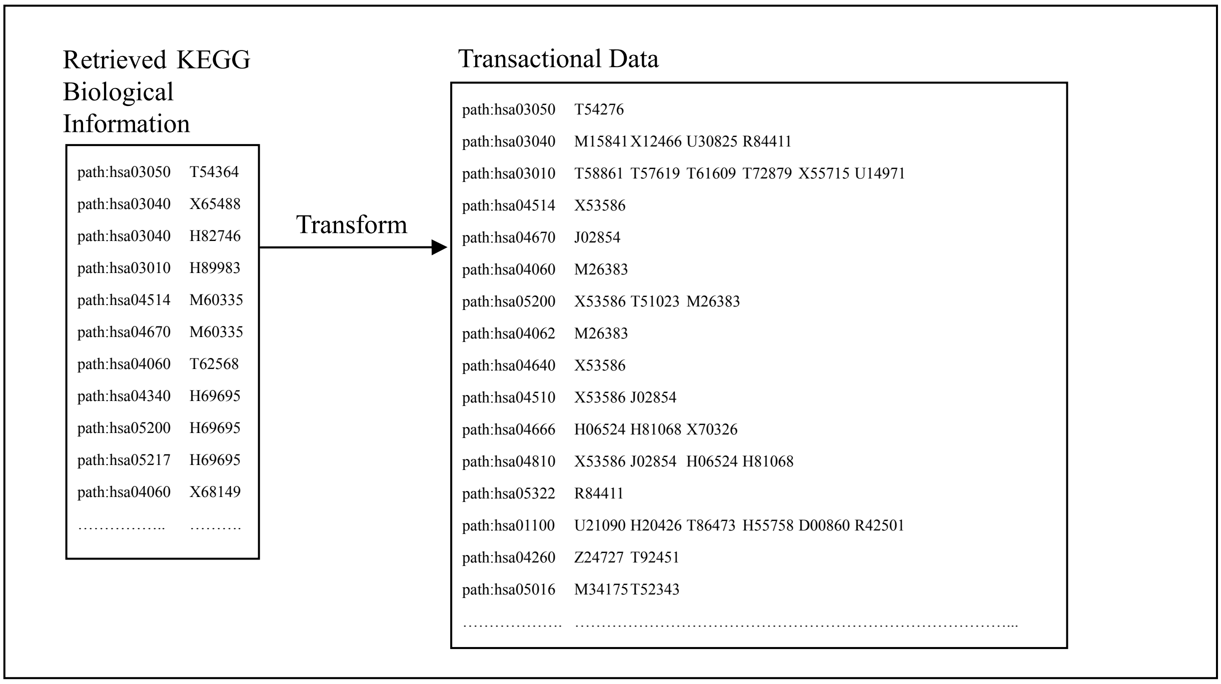
Table 6. Results on breast cancer dataset

Classifier	**Expression-Only (Information Gain)**		***KEGG (Information Gain + Association Analysis)***	
	50 top-ranked genes	100 top-ranked genes	Accuracy	Number of genes
NB	91.75%	91.75%	90.72%	38
LR	89.69%	95.88%	89.69%	97
SVM	92.78%	92.78%	95.88%	47

```
$C_k$: Candidate itemset of size k
$L_k$ : Frequent itemset of size k

$L_1$ = {frequent items};
for ($k$ = 1; $L_k$ !=∅; $k$++) do begin
        $C_{k+1}$ = candidates generated from $L_k$;
    for each transaction $t$ in database do increment the count of
    all candidates in $C_{k+1}$ that are contained in $t$
        $L_{k+1}$ = candidates in $C_{k+1}$ with min_support
        end
return $\cup_k L_k$;
```

Figure 1. The Apriori Algorithm for Finding Frequent Itemsets



path:hsa03050	T54364
path:hsa03040	X65488
path:hsa03040	H82746
path:hsa03010	H89983
path:hsa04514	M60335
path:hsa04670	M60335
path:hsa04060	T62568
path:hsa04340	H69695
path:hsa05200	H69695
path:hsa05217	H69695
path:hsa04060	X68149
……………..	……….

path:hsa03050	T54276
path:hsa03040	M15841 X12466 U30825 R84411
path:hsa03010	T58861 T57619 T61609 T72879 X55715 U14971
path:hsa04514	X53586
path:hsa04670	J02854
path:hsa04060	M26383
path:hsa05200	X53586 T51023 M26383
path:hsa04062	M26383
path:hsa04640	X53586
path:hsa04510	X53586 J02854
path:hsa04666	H06524 H81068 X70326
path:hsa04810	X53586 J02854 H06524 H81068
path:hsa05322	R84411
path:hsa01100	U21090 H20426 T86473 H55758 D00860 R42501
path:hsa04260	Z24727 T92451
path:hsa05016	M34175 T52343
……………….	………………………………………………………………………....

Figure 2. The construction of transactional data based on gene annotations