Gene Selection with Rough Sets for the Molecular Diagnosing of Tumor Based on Support Vector Machines

Shulin Wang[1,2], Huowang Chen[1], Renfa Li[2], and Dingxing Zhang[1]
1 School of Computer Science, National University of Defense Technology, Changsha, Hunan 410073, People's Republic of China
jt_slwang@hnu.cn
2 School of Computer and Communication, Hunan University, Changsha, Hunan 410082, People's Republic of China

ABSTRACT

The development of microarray technology has motivated interest of its use in clinical diagnosis of tumor and drug discovery. However the accurate classification of tumor by selecting the tumor-related genes from thousands of genes is a difficulty task due to the large number of redundant genes. Therefore, we propose a novel hybrid approach which combines rough set theory with support vector machines to further improve the classification performance of gene expression data. Our approach is assessed on two well-known tumor datasets, and experiments indicate that gene selection based on the rough set theory is effective because most of the selected genes are relevant to tumor using rough set attribute reduction, and support vector machines classifier has a better performance on the selected informative genes.

Keywords Gene expression profiles; DNA microarry; Support vector machines; Gene selection; Rough set theory

1: Introduction

The advent of DNA microarray technology provides biologists with the ability to measure the expression levels of thousands of genes in a single experiment. With the development of this technology, a large quantity of gene expression data from such experiments has been accumulating quickly, so a novel means should be explored to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease. However, in clinical application the first difficulty is how to select the tumor-related genes to be used as cancer biomarkers to conveniently diagnosis and treat tumor.

When we consider genes as features, we have to face the problem of feature selection. Gene selection can be seen as a typical combinatorial problem. Given a dataset described by a large number of genes, the goal is to find out the smallest subset that leads to the highest rate of correct classification. Generally speaking, existing methods for gene selection belong to three main families [3][4]: the filter approach, the wrapper approach and the embedded approach. The filter methods separate the gene selection procession from the classification process. But the wrapper approach relies on a classification algorithm that is used as a black box to evaluate each candidate subset of genes. In embedded methods, the process of selection is performed during the training of a specific learning machine.

However, due to the high dimensionality of gene expression profiles, the tumor-related gene selection is not an easy task. Rough set theory is a formal methodology that can be employed to reduce the dimensionality of dataset as a preprocessing step to training a learning system on the data. Rough set attribute reduction works by selecting the richest information attributes in a dataset without transforming the data. In this paper, we propose a hybrid classification approach which combines the filter approach with the wrapper approach. Concretely speaking, the proposed approach integrates gene ranking based on the revised feature score criterion and the attribute reduction of rough set theory with support vector machines classifier.

2: Related works

A great deal of research has been done in the

classification of gene expression data by utilizing unsupervised methods such as clustering and self-organizing maps. In recent years, supervised methods such as k-nearest neighbor (KNN) and support vector machines (SVM) have been broadly applied to gene expression profiles to classify tumor samples [4][5][6][7][8][9].

However, informative gene selection plays a key role in the classification problem of gene expression data, so gene selection for classification is an important aspect of data mining and a very active research topic. Guyon et al [7] proposed a gene selection approach utilizing support vector machines based on recursive feature elimination (RFE) by which the selected genes yield a better classification performance and are biologically relevant to tumor. Yuhang et al [12] developed a novel hybrid approach that combines gene ranking and clustering analysis. This approach applied feature filtering algorithms to select a set of top-ranked genes and then applied hierarchical clustering on these genes to generate a dendrogram which was used as the basis of marker gene selection.

Rough set theory has been developed quickly in recent years and has been successfully applied to gene expression profiles. Herman Midelfart et al [10] presented a general rough set approach for the classification of tumor samples. Bulashevska et al [20] applied rough set to extract informative rules. Jianwen Fang et al [19] utilized rough set approach to predict leukemia and to have found eight tumor-related genes and eight informative rules in the leukemia dataset. Those works show that rough set based learning combined with feature selection may become an important tool for microarray analysis.

3: The Classification Methods

3.1: Preprocessing of DNA microarray Data

DNA microarrays are composed of thousands of individual DNA sequences printed in a high density array on a glass microscope slide using a robotic array. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples are reverse-transcribed into cDNA, labeled using different fluorescent dyes mixed (red-fluorescent dye Cy5 and green-fluorescent dye Cy3). After the hybridization of these samples with the arrayed DNA probes, the slides are imaged using scanner that makes fluorescence measurements for each dye. The log ratio between the two intensities of each dye is used as the gene expression data: $gene\_expression = \log_2(Ratio)$, $Ratio = Int(Cy5)/Int(Cy3)$, where $Int(Cy5)$ and $Int(Cy3)$ are the intensities of red and green colors. Samples are generated under multiple conditions which may be a time series during a biological process or a collection of different tissue samples.

Let $G = \{g_1, \cdots g_n\}$ be a set of genes and $S = \{s_1, \cdots s_m\}$ be a set of samples. The corresponding gene expression matrix can be represented as $X = \{x_{i,j} \mid 1 \le i \le m, 1 \le j \le n\}$. The matrix X is composed of m row vectors $s_i \in R^n; i = 1,2,\cdots,m$, m is the number of samples, and n is the number of genes measured.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{bmatrix}$$

Where $x_{i,j}$ is the expression level value of sample s_i on gene g_j, and usually $n >> m$. Each vector s_i in the gene expression matrix may be thought of as a point in *n*-dimensional space. Each of the n columns consists of an *m*-element expression vector for a single gene.

Our task is to classify all samples into tumor samples and normal samples, which is a binary classification problem. A simple way to build a binary classifier is to construct a hyper-plane which separates tumor members from normal members in feature space. Suppose ω_T and ω_N be the two subsets of sample set S, satisfying $\omega_T \cap \omega_N = \phi, \omega_T \cup \omega_N = S$, which means that each vector ideally belongs to one and only one class ω_T or ω_N.

3.2: The model of classification algorithm

There are four steps in our classification algorithm that will be introduced below in details.

Step 1 For each gene g_i in G, we firstly calculate its score according to the revised feature score criterion (RFSC)[14], and then rank the genes according to their scores. On the basis of gene ranking, we simply take the top-ranked genes with the highest $F(g_i)$ scores as our selected gene subset G_{top}, satisfying $|G_{top}| << |G|$.

Step 2 Applying the attribute reduction of rough set theory to the top-ranked gene subset G_{top} to further select the gene subset G_r consisting of r genes as represents of G_{top}.

Step 3 Firstly, splitting the dataset into training dataset and testing dataset, and then applying SVM classifier to classify the training dataset described by the gene subset G_r to obtain a classification model.

Step 4 Using the model and SVM to predict the testing dataset.

3.3: Gene selection

Gene selection and dimensional reduction are necessary for performing the tumor classification with gene expression profiles. In measuring the classification information of genes, Golub et al [8] proposed a feature score criterion (FSC) as gene selection method. For each gene g_i in G, The FSC method firstly calculate the mean μ_i^+ (resp. μ_i^-) and standard deviation σ_i^+ (resp. σ_i^-) which correspond to the gene g_i of samples labeled +1(-1), respectively, and then calculate feature score with the formula $F(g_i)=\left|(\mu_i^+-\mu_i^-)/(\sigma_i^+ + \sigma_i^-)\right|$ for each $g_i \in G$, and rank the genes according to their scores. However, when the two expression means of a gene g_i in normal tissue and tumor are equal, there is a fault in this formula that this gene g_i is removed as noise from informative genes because of $F(g_i)=0$. Therefore, we apply another revised formula RFSC[14]:

$$F(g_i)=0.5\left|(\mu_i^+-\mu_i^-)/(\sigma_i^+ + \sigma_i^-)\right| + 0.5\ln((\sigma_i^{+2}+\sigma_i^{-2})/(2\sigma_i^+\sigma_i^-)) \quad (1)$$

to be used as our gene selection criterion. We simply take the top-ranked genes with the highest $F(g_i)$ scores as our gene subset G_{top}. Suppose $\left|G_{Top}\right| = p$, then we may obtain gene expression matrix $X_{m\times p}$.

3.4: Rough Set and Attribute Reduction Method [2]

Our learning problem is to predict the class of tumors. We may formalize this problem as a decision system which is defined as a quadruple: $S=<U,A,V,f>$, where universe $U=\{x_1,x_2,\cdots,x_n\}$ is a finite set of tumor or microarray samples; The set A is a finite set of attributes; the attributes in A are further classified into two disjoint subsets: condition attributes C for each gene and decision attributes D, corresponding to a clinical parameter, such that $A=C\cup D$ and $C\cap D=\phi$; $V=\bigcup_{a\in C}V_a$ is a set of gene expression values for each gene a and V_a is the domain of gene a; $f:U\times C\to V$ is an information function which assigns particular values from domains of attributes to objects such that $f(x_i,a)\in V_a$, for all $x_i\in U$ and $a\in C$. In our application, $D=\{d\}$ is a singleton set, where d denotes the classes of samples.

Given a decision system $DS=<U,A,V,f>$, let B be a subset of A, and let x_i and x_j be members of U, a relation $R(B)$, called an indiscernibility relation over B, is defined as follows:

$$R(B)=\{(x_i,x_j)\in U^2 \mid \forall a\in B, f(x_i,a)=f(x_j,a)\} \quad (2)$$

Let C be a set of condition attributes and $R(C)$ be an indiscernibility relation on U, an ordered pair $AS=<U,R(C)>$ is called an approximation space based on C.

Let $Y\subseteq U$ be a subset of objects representing a concept, and $R^*(C)=\{X_1,X_2,\cdots,X_n\}$ be the collection of equivalence classes induced by the relation $R(C)$. The lower approximation of a set Y in the approximation space AS denoted as $LOW_{R(C)}(Y)$, is defined as the union of those equivalence classes in the collection of $R^*(C)$ which are completely contained by the Y, $LOW_{R(C)}(Y)=\bigcup\{X\in R^*(C): X\subseteq Y\}$.

Let $R^*(D)=\{Y_1,Y_2,\cdots Y_m\}$ be the collection of equivalence classes of the relation $R(D)$. A positive region

$$POS_C(D)=\bigcup_{i=1,\cdots,m}\{LOW_{R(C)}(Y_i): Y_i\in R^*(D)\}$$

The positive region $POS_C(D)$ includes all samples of the equivalence classes of $R^*(C)$ in AS which can be certainly classified into classes of $R^*(D)$.

Attribute reduction techniques can eliminate redundant attributes and create a minimal subset of attributes called reduct for a decision system. Such minimal subset of attributes is an essential part of the decision system which can discern all samples discernible by the original table and cannot be reduced any more. Finding reducts is also expensive. An exhaustive search is obviously impossible, but heuristic search is also very time consuming.

Therefore, we use a feature selection approach to select genes with high discriminatory ability before finding reducts using rough set learning algorithm. Moreover, the gene expression values are real-valued, and must be discretized before gene selection [10].

3.5: Support Vector Machines

SVM is a relatively new type of statistic learning theory,

originally introduced by Vapnik and successively extended by a number of other researchers. SVM builds up a hyper-plane as the decision surface in such a way to maximize the margin of separation between positive and negative examples. Given a labeled set of m training samples $S = \{(x_i, y_i) \mid (x_i, y_i) \in R^n \times \{\pm 1\}, i = 1,2,\cdots m\}$, where $x_i \in R^n, y_i \in \{\pm 1\}$ is a label of sample x_i, and the discriminant hyper-plane is defined by:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i K(x_i, x) + b \qquad (3)$$

where $K(x_i, x)$ is a kernel function and the sign of $f(x)$ determines which class it belongs to. Constructing an optimal hyper-plane is equivalent to finding all the support vectors α_i and a bias b.

4: Experiments

4.1: Sample Datasets

We experiment with two dataset related to tumor. One is leukemia dataset [8]; another is colon cancer dataset [18]. Leukemia dataset is bone marrow samples that are taken from 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). It consists of 47 ALL samples and 25 AML samples. The dataset contains expression levels for 7129 human genes produced by Affymetrix high-density oligonucleotide microarrays. The scores in the dataset represent the intensity of gene expression after being re-scaled to make overall intensities for each chip equivalent. The dataset is available at web site http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi.

Colon cancer dataset involves comparing tumor and normal samples of the same tissue. The dataset consists of 62 samples of colon epithelial cells including 40 colon cancer samples and 22 normal samples. Gene expression level in these 62 samples was measured using high density oligonucleotide microarray. Among the 6000 genes detected in these microarrays, 2000 genes were selected based on the confidence in the measured expression level. The dataset is available at web site http://www.molbio.princeton.edu/colondata.

4.2: Experiment Methods

In our experiments, we firstly apply the rough set software RSES 2.2 (Downloaded from the web site http://logic.mimuw.edu.pl/~rses) to select genes that have a better discriminative ability. Then we use the SVM software LIBSVM [13] to classify the two tumor-related datasets. Training SVM requires specifying the type of kernel and the regularization parameter C. However, finding the best choices for the kernel and parameters can be challenging when applied to real datasets. Generally, the recommended kernel for nonlinear problems is the Gaussian radial basis kernel $K(x, y) = \exp(-\sigma \|x - y\|^2)$ that is also used in our experiments. We adopt the cross-validated (CV) accuracy to measure the classification performance of SVM classifier.

4.3: Results and analysis

Firstly, experiments are carried out using RFSC method to roughly select the top-ranked genes as represents of all genes, and then on the basis of the selected genes we employ rough set to find the tumor-related genes to be used as the input of SVM classifier. Table 1 shows the experiment results of two methods for leukemia dataset. The first column means the number of the roughly selected genes according to gene ranking; the second column means the CV accuracy obtained from SVM using the roughly selected genes; the third column indicates the selected gene subset, using RSES 2.2 software, whose CV accuracy is showed in the forth column.

Table 1. Recognition rate with gene selection methods for leukemia dataset

#Selected Genes	CV Accuracy	Selected Gene Subset Using Rough Set	CV Accuracy
Top 50	98.61%	{**X95735,** M83652, M23197}	95.83%
Top 100	98.61%	{**X95735,** M31523, M23197}	97.22%
Top 150	98.61%	{**X95735,** M31523, M23197}	97.22%
Top 200	98.61%	{**X95735,** M31523, M23197}	97.22%
Top 500	98.61%	{**X95735,** M31523, M23197}	97.22%
Top 1000	98.61%	{**X95735**, D87447, M31951}	95.83%
Top 1500	98.61%	{**X95735**, L33243, M31951}	95.83%
Top 3000	98.61%	{**X95735**, L32831, M31951}	95.83%
Top 6000	98.61%	{**X95735**, X68561, M31951}	95.83%

Further experiments show that the subset {X95735, M23197} has the same classification performance as the set {X95735, M23197, M31523} which achieves 97.22% CV accuracy. In fact, the genes X95735 and M23197 are relevant to leukemia. X95735 possesses LIM domain which is known to interact with leukemogenic bHLH proteins (TAL1, TAL2 and LYL1) [21]. M23197 has previously been identified as gene associated with myeloid leukemia and as "Coding for CD33, a differentiation antigen of myeloid progenitor cells" [22]. Fig.1 shows the scatter plot of the two genes. Along the ordinate axis are the expressional values of

gene M31523, and along the abscissa axis are the expression values of X95735. From this figure, we can see that the boundary between ALL and AML is very clear relatively.

Yuhang et al [12] utilize HykGene approach to obtain a gene subset {**X95735**, M27783, U41813, **M31523**, HG2562-HT2658, J05243, M17886, U43885, J02982, M10612, M17733, X99728} which can achieve the 100% CV accuracy using SVM classifier, but not all genes in this set are relevant to tumor. Another different gene candidate subset {M23197, **X95735**, **M31523**, U46499, M27891, L09209, M63138, HG1612-HT1612, M92287, M11722} can also achieve the 100% CV accuracy using the same classifier. This phenomena indicates that the gene subset that can achieve the highest CV accuracy is not solitary. Therefore, although the CV accuracy is the better way to indicate the performance, to some extent it is hard to evaluate the different gene selection methods which achieve the same CV accuracy, so evaluating experiment results should concern much medical knowledge. Compared with our results, {X95735, M31523} is the intersection of these selected gene subsets.

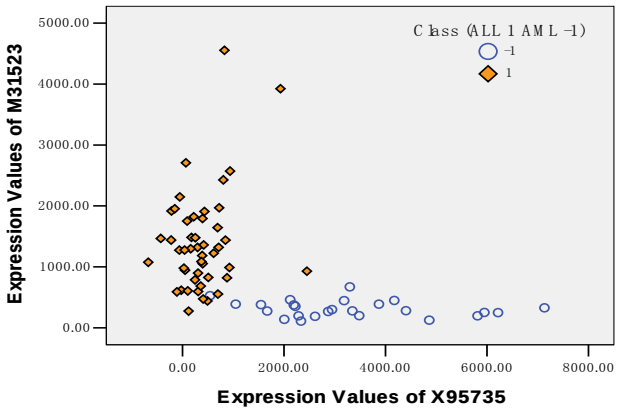

Fig.1 Scatter plot of two genes {X95735, M31523} in leukemia dataset.

Table 2. Gene selection for colon dataset and its CV accuracy of classification

#Selected Genes	CV Accuracy	Selected Gene Subset Using Rough Set	CV Accuracy
Top 50	88.71%	{M76378, U21090, **H08393**, R87126, R64115}	87.1%
Top 100	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%
Top 150	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%
Top 200	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%
Top 500	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%
Top 1000	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%
Top 1500	90.32%	{R36977, **H08393,** R87126, T62947}	87.1%

Table 2 shows the experiment results of two methods for colon dataset. The meanings of columns are similar to table 1. Further experiments show that the subset {H08393, R87126} can achieve the 88.71% CV accuracy that is higher than the gene subset {R36977, H08393, R87126, T62947}. H08393 and T62947 are two genes of colon cancer biomarkers that had been applied for United States patent whose number is 20050165556 in 2005. R36977 is not associated with colon cancer in previous literature, but is linked to either some forms of neoplasia or to the regulation of the cell cycle [24]. Fig.2 shows the scatter plot of the two genes. Along the ordinate axis are the expressional values of gene R87126, and along the abscissa axis are the expression values of gene H08393. The boundary between colon tumor and normal tissues is fuzzy relatively.

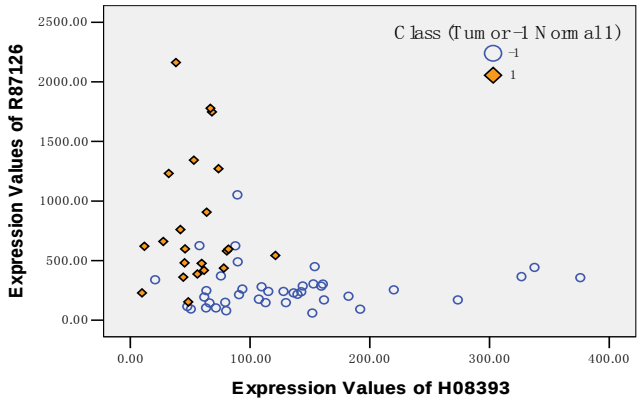

Fig.2 Scatter plot of two genes {H08393, R87126} in colon cancer dataset.

5: Conclusion and future work

Due to the gene redundancy in gene expression profiles, eliminating a large quantity of redundant genes from thousands of genes is a difficulty and important task for the tumor-related gene selection and tumor classification. In this paper, our main contribution is to introduce a novel hybrid approach which combines gene ranking based on RFSC and rough set attribute reduction to select biomarker genes for classification using SVM classifier. Experiments show that our hybrid method performs well in selecting biomarker genes related to tumor and in improving the performance of SVM classifier. The selected biomarkers are potential drug targets since they are relevant to the disease under study. We will further focus on developing the classification tool which will integrate various feature selection methods to help doctor to diagnose and predict cancer.

Acknowledgement

This research was funded by the National Natural Science Foundation of China under the grant No. 60233020 and Hunan Provincial Natural Science Foundation of China under the grant No. 04JJ6032.

References

1. Vapnik V. N.. Statistical learning theory. Springer, New York, 1998.
2. Pawlak Z.. Rough Sets: Theoretical Aspects of Reasoning about Data—Dordrecht, Kluwer Academic Publishers, 1991.
3. Guyon I. and Elisseeff A.. An introduction to variable and feature selection. Journal of Machine Learning Research, 2003, 3:1157-1182.
4. Edmundo Bonilla Huerta, Béatrice Duval, and Jin-Kao Hao. A hybrid GA/SVM approach for gene selection and classification of microarray data. EvoWorkshops, 2006, pp. 34-44.
5. Terrence S. Furey, Nello Cristianini, Nigel Duffy, David W. Bednarski, Michel Schummer, and David Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 2000, 16 (10):906-914.
6. Krzysztof Simek, Krzysztof Fujarewicz, Andrzej Swierniak, Marek Kimmel, Barbara Jarzab, Malgorzata Wiench, and Joanna Rzeszowska. Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. Engineering Applications of Artificial Intelligence, 2004, 17:417-427.
7. Guyon I., Weston J., Barnhill S., and Vapnik V.. Gene selection for cancer classification using support vector machines. Machine Learning, 2002, 46:389-422.
8. Golub T.R., Slonim D.K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., and Lander E.S.. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999, 286:531-537.
9. Wei Chu, Zoubin Ghahramani, Francesco Falciani, and David L. Wild. Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinformatics, 2005, 21(16):3385-3393.
10. Herman Midelfart, Jan Komorowski, Kristin Norsett, Fekadu Yadetie, Arne K. Sandvik, and Astrid Laegreid. Learning rough set classifiers from gene expressions and clinical Data. Fundamenta Informaticae, 2002, 53:155-183.
11. Sung-Bae Cho and Hong-Hee Won. Machine learning in DNA microarray analysis for cancer classification. Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics, 2003, 189-198.
12. Yuhang Wang, Fillia S. Makedon, James C. Ford, and Justin Pearlman. HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data. Bioinformatics, 2005, 21(8):1530-1537.
13. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines (2001), Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
14. Li Yingxin and Ruan Xiaogang. Feature selection for cancer classification based on Support Vector Machine. Journal of Computer Research and Development, 2005, 42(10):1796~1801.
15. Topon Kumar Paul and Hitoshi Iba. Extraction of informative genes from microarray data. Proceedings of the 2005 Conference on Genetic and Evolution Computation. Washington DC, USA, 2005, 453-406.
16. Daisuke Komura, Hiroshi Nakamura, and Shuichi Tsutsumi. Multidimensional support vector machines for visualization of gene expression data. Bioinformatics, 2005, 21(4):439-444.
17. Michael P.S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S. Furey, Manuel Ares Jr., and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 2000, 97(1):262-267.
18. Alon U., Barkai N., Notterman D.A., Gish K.,Ybarra S., Mack D., and Levine A.. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues by oligonucleotide arrays. Proc. Nat. Acad. Sci. USA, 1999, 96: 6745-6750.
19. Jianwen Fang and Jerzy W. Grzymala-Busse. Leukemia prediction from gene expression data—a rough set approach. 2006 Annual Kansas City Area Life Sciences Research Day, Kansas City, MO, 2006.
20. Bulashevska S., Dubitzky W., and Eils R. Mining gene expression data using rough set theory. In: Proceeding of Critical Assessment of Techniques for Microarray Data Analysis (CAMDA'00 Conference), Duke University, NC, US, 2000, pp 4-5.
21. Wadman I., Li J. X., Bash R. O., Forster A., Osada H., Rabbitts T.H., and Baer R.. Specific in-vivo association between the Bhlh and Lim proteins implicated in human T-cell leukemia. EMBO Journal, 1994, 13:4831-4839.
22. Simmons D. and Seed B.. Isolation of a cDNA encoding CD33, a differentiation antigen of myeloid progenitor cells. Journal of Immunology, 1988, 141(8):2797-2800.
23. Kamps MP, Murre C., Sun XH, and Baltimore D. A new homeobox gene contributes the DNA binding domain of the t(1;19) translocation protein in pre-B ALL. Cell, 1990, 60(4):547-555.
24. Han-yu Chuang, Huai-Kuang Tsai, and Yuan-fan Tsai. Ranking genes for discriminability on microarray data. Journal of Information Science and Engineering, 2003, 19:953-966.