Gene Selection and Cancer Classification: A Rough Sets Based Approach

Lijun Sun, Duoqian Miao, and Hongyun Zhang

Key Laboratory of Embedded System and Service Computing,
Ministry of Education,
Tongji University, Shanghai 201804, P.R.China
Department of Computer Science and Technology,
Tongji University, Shanghai, 201804, P.R.China
Sunlj1028@yahoo.com.cn, Miaoduoqian@163.com, Zhanghongyun583@sina.com

Abstract. Indentification of informative gene subsets responsible for discerning between available samples of gene expression data is an important task in bioinformatics. Reducts, from rough sets theory, corresponding to a minimal set of essential genes for discerning samples, is an efficient tool for gene selection. Due to the compuational complexty of the existing reduct algoritms, feature ranking is usually used to narrow down gene space as the first step and top ranked genes are selected . In this paper,we define a novel certierion based on the expression level difference btween classes and contribution to classification of the gene for scoring genes and present a algorithm for generating all possible reduct from informative genes.The algorithm takes the whole attribute sets into account and find short reduct with a significant reduction in computational complexity. An exploration of this approach on benchmark gene expression data sets demonstrates that this approach is successful for selecting high discriminative genes and the classification accuracy is impressive.

Keywords: gene selection, cancer classification, rough sets, reduct, feature ranking, bioinformatics, gene expression.

1 Introduction

Standard medical classification systems for cancer tumors are based on clinical observations and the microscopical appearances of the tumors, these systems fail to recognize the molecular characteristics of the cancer that often correspond to subtypes that need different treatment. Studying the expression levels of genes in tumor issues may reveal such subtypes and may also diagnose the disease before it manifest itself on a clinical level, tumor classification based on gene expression data analysis is becoming one of the most important research areas in bioinformatics.

Gene expression data often has thousands of genes while not more than a few dozens of tissue samples, with such a huge attribute space, it is almost certain that very accurate classification of tissue samples is difficult. Recent research

J.F. Peters et al. (Eds.): Transactions on Rough Sets XII, LNCS 6190, pp. 106–116, 2010.

has shown that the expression level of fewer than ten genes are often sufficient for accurate diagnosis of most cancers, even though the expression levels of a large number of genes are strongly correlated with the disease[1],[2],[3],[4]. In fact, the use of a much larger set of gene expression levels has been shown to have a deleterious effect on the diagnostic accuracy due to the phenomenon known as the curse of dimensionality. Thus Performing gene selection prior to classification will help to narrowing down the attribute number and improving classification accuracy. More importantly, by identifying a small subset of genes on which to base a diagnostic accuracy, we can gain possibly significant insights into the nature of disease and genetic mechanisms responsible for it. In addition, assays that require very few gene expression levels to be measured in order to make diagnosis are far more likely to be widely deployed in a clinical setting. How to select the most informative genes for cancer classification is becoming a very challenging task.

Rough set theory proposed by Pawlak provides a mathematic tool that can be used to find out all possible feature sets [5],[6].It works by gradually eliminating superfluous or redundant features to find a minimal set of essential features for classification. Reduct, corresponds to such a minimal subset, will be used to instead of all features in the learning process. Thus a feature selection method which is using rough set theory can be regarded as finding such a reduct with respect to the best classification [7].

Gene expression data analysis represents a fascinating and important application area for rough sets-based methods. Recently, researchers have focused their attention on gene subsets selection and cancer classification based on rough sets [8],[9]. Midelfart et al. used rough set-based learning methods implemented with ROSETTA involving GAs and dynamic reducts for gastric tumor classification; rule models for each of six clinical parameters are induced [10]. Valdes et al. investigated an approach using clustering in combination with Rough Sets and neural networks to select high discriminated genes[11] . Momin et al. developed a positive region based algorithm for generating reducts from gene expression data [12], results on benchmark gene expression datasets demonstrate more than 90% reduction of redundant genes. Banerjee et al. adopted an evolutionary rough feature selection algorithm for classifying microarray gene expression patterns. Since the data typically consist of a large number of redundant features, an initial redundancy reduction of the attributes is done to enable faster convergence. Thereafter rough set theory is employed to generate reducts in a multiobjective framework. The effectiveness of the algorithm is demonstrated on three cancer datasets, colon, lymphoma, and leukemia. In case of the leukemia data and lymphoma data, two genes are selected, whereas the colon data results in an eight-gene reduct size [13].

In this paper, we propose a rough set based method for identifying informative genes for cancer classification. We first develop a novel criterion for scoring genes, top ranked k genes are selected with this criterion, and then rough sets attribute reduction is employed to obtain all possible subsets that define the same partition with all gens, rule sets induced from all reducts used as the classifier for

labeling the new samples. The rest of the paper is organized as follows: Section 2 provides the basics of rough sets, then, our method is detailed in Section 3. Experimental results demonstrated on benchmark microarray data sets are listed, the discussions of these results are given in section 4. Finally, the conclusion is drawn in Section 5.

2 Rough Sets Preliminaries

In rough set theory, a decision table is denoted by $T = (U, C \cup D)$, where U is a nonempty finite set called universe, C is a nonempty finite set called condition attribute sets and $D = \{d\}$ is decision feature. Rows of the decision table correspond to objects, and columns correspond to attributes [7].

Let $B \subseteq C \cup D$, a binary relation $IND(B)$, called the indiscernibility relation, is defined as

$$IND(B) = \{(x, y) \in U \times U : \forall a \in B, a(x) = a(y)\} \tag{1}$$

$IND(B)$ is also an equivalence relationship on U. Let $U/IND(B)$ denotes the family of all equivalence classes of the relation $IND(B)$ (or classification of U), $U/IND(B)$ is also a definable partition of the universe induced by $IND(B)$.

A subset $X \subseteq U$ can be approximated by a pair of sets, called lower and upper approximation with respect to an attribute subset $B \subseteq C$, the $B-$ lower approximation of X is denoted as $\underline{B}(X) = \{Y \in \underline{U/IND(B)} : Y \subseteq X\}$ and the $B-$ upper approximation of X is denoted as $\overline{B}(X) = \{Y \in U/IND(B) : Y \cap X \neq \varphi\}$ respectively.

The C -positive region of D, denoted by $POS_C(D)$, is defined as:

$$POS_C(D) = \bigcup_{X \in U/IND(D)} \underline{C}(X) \tag{2}$$

For an attribute $c \in C$, if $POS_C(D) \neq POS_{C-\{c\}}(D)$, c is an indispensible feature, delete it from C will decrease the dicernibility between objects. Otherwise, if $POS_C(D) = POS_{C-\{c\}}(D)$, c is a dispensable feature, that is, c is redundancy and can be deleted from C without affecting the discernibility of objects. The subset of attributes $R \subseteq C$ is a reduct of attribute C with respect to D if

$$POS_R(D) = POS_C(D) \tag{3}$$

A reduct R of C is called a minimal reduct if $\forall Q \subset R$, $POS_Q(D) \neq POS_C(D)$,thus R represents the minimal set of non-redundant features which capable of discerning objects in a decision table, thus it will be used to instead of C in a rule discovery algorithm. The set of all indispensable features in C is core of C with respect to D, denoted by $CORE_D(C)$, we have

$$CORE_D(C) = \cap RED(C) \tag{4}$$

Where $RED(C)$ is the set of all reducts of C with respect to D.

Reducts have been nicely characterized in [17] by discernibility matrices and discernibility functions. If $T = (U, C \cup D)$ is a decision table, with $U = \{x_1, x_2, ..., x_n\}$, the disneribility matrix of T, denoted by M(T), mean a $n \times n$ matrix defined as:

$$m_{ij} = \{c \in C : c(x_i) \neq c(x_j) \wedge d(x_i) \neq d(x_j)\}, i = 1, 2, ..., n \tag{5}$$

Thus entry m_{ij} is the set of all attributes that classify object x_i and x_j into different decision class in $U/IND(D)$.The $CORE(C)$ can be defined as the set of all single element entries of the discernibility matrix, that is

$$CORE(C) = \{c \in C : mij = \{c\}\} \tag{6}$$

A discernibility function f is a Boolean function of m boolean variables $a_1, a_2, ..., a_m$ correspond to the attributes $c_1, c_2, ..., c_m$, defined as:

$$f = \wedge\{\vee(m_{ij}) : 1 \leq j < i \leq n, m_{ij} \neq \varphi\} \tag{7}$$

Where $\vee(m_{ij})$ is the disjunction of all the variables a with $c \in m_{ij}$.It is seen that $\{c_{i1}, c_{i2}, ..., c_{ip}\}$ is a reduction in the decision table T if and only if $\{c_{i1} \wedge c_{i2} \wedge ..... \wedge c_{ip}\}$ is a prime implicant of f.

In rough set community, most feature subset selection algorithm are attribute reduct-oriented, that is, finding minimum reducts of the conditional attributes of decision tables. Two main approaches to finding attribute reducts are recognized as discernibility function-based and attribute dependency-based [14],[15],[16],[17]. These algorithms, however, suffer from intensive computations of either discernibility functions for the former or positive region for the latter, which limit its application on large scale data sets.

3 Informative Genes

As mentioned above, for all genes measured by microarray, only a few of them play major role in the processes that underly the differences between the classes, the expression levels of many other genes maybe irrelevant to the distinction between tissue classes. The goal of this work is to extract those genes that demonstrate high discriminating capabilities between the classes of samples, these genes are called informative genes, an important characteristic of them is that the expression level in different classes has the remarkable difference, that is, the gene that demonstrate high difference in its expression levels in different classes is a good significant genes that is typically highly related with the disease of samples.

For a binary classification problem (assuming class1 vs. class2), when we consider a gene , if $U/IND(g) = \{X_1, X_2\}$ and all the samples in X_1 belong to class1 and all the samples in X_2 belong to class2, then gene g is most useful gene for classify new samples.

For that, we exam for each gene, all the values of each class, and get the number of equivalence classes.

a =# of equivalence classes of gene g in class1
b =# of equivalence classes of gene g in class2

Thus the most useful gene is the one that has the lowest a and b values and the case a =1 and b =1 gives the least noise, thus the measure $a \times b$ is a good indicator of how much a gene differentiates between two classes ,therefore we compute a V score for each gene

$$V = a \times b \tag{8}$$

In real applications, two or more genes may have same V values, therefore mutual information (MI) is used to distinguish such genes. Given the partition by D, $U/IND(D)$, of U, the mutual information based on the partition by $g \in C$, $U/IND(g)$, of U, is given by

$$\begin{aligned} &I(D,g) = H(D) - H(D|\{g\}) \\ &= \frac{1}{|U|} \sum_{X \in U/IND(D)} |X| \log \frac{|X|}{|U|} - \frac{1}{|U|} \sum_{X \in U/IND(D)} \sum_{Y \in U/IND(g)} |X \cap Y| \log \frac{|X \cap Y|}{|Y|} \end{aligned} \tag{9}$$

$I(D,g)$ quantifies the relevance of gene g for the classification task. Thus the criterion for scoring genes can be described below:

1. A gene with the lowest V value is the one that have the highest differences in its expression levels between two classes of samples.
2. If two or more genes have same V values, the one has maximal mutual information is more significant.

4 Our Proposed Methods for Gene Selection and Cancer Classification

Our learning problem is to select high discriminate genes from gene expression data for cancer classification, which define the same partition as the whole set of genes. We may formalize this problem as a decision table $T = (U, C \cup \{d\})$, where universe $U = \{x_1, x_2, ..., x_n\}$ is a set of tumors. The conditional attributes set $C = \{g_1, g_2, ..., g_m\}$ contains each gene, the decision attribute d corresponds to class label of each sample. Each attribute $g_i \in C$ is represented by a vector $g_i = \{v_{1,i}, v_{2,i}, ..., v_{n,i}\}$, $i = 1, 2, ...m$ where $v_{k,i}$ is the expression level of gene i at sample $k, k = 1, 2, ..., n$.

4.1 Discretization

The methods of rough sets can only deal with data sets having discrete attributes because real values attribute will lead to large number of equivalence classes, thus lead to large number of rules, thereby making rough sets classifiers inefficient. In our study, a simplest method is used to discretize the data set. It is applied for each continuous attribute independently, for each attribute, it sorts the continuous feature values and gets a cut $c = (V_a(x_i) + V_a(x_{i+1}))/2$ when the adjacent two samples x_i and x_{i+1} have different class labels, namely $d(x_i) \neq d(x_{i+1})$.

4.2 Gene Selection

A major obstacle for using rough sets to deal with gene expression data may be the large scale of gene expression data and the comparatively slow computational speed of rough sets algorithms. The computation of discernibility has a time complexity of $O(n^2)$, which is still much higher than many algorithms in bioinformatics. In order to solve this problem, each gene is measured for correlation with the class according to some criteria, top ranked genes are selected before we employee the existing attribute reduct algorithm on gene expression data sets. The literature discusses several method for scoring genes for relevance, Simple methods based on t-test and mutual information are proved to be effective [18],[19],[20].

Feature sets so obtained have certain redundancy because genes in similar pathways probably all have very similar score and therefore no additional information gain. If several pathways involved in perturbation but one has main influence it is Possible to describe this pathway with fewer genes, therefore Rough sets attribute reduction is used to minimize the feature sets.

Our proposed method uses the criterion described above in section 3 to score genes and starts with choosing genes having highest significance. From these chosen genes, possible subsets are formed and subsets having same classification capability with the entire gene set are chosen, the algorithm is formulated as the following:

Step 1: compute the significance of each gene
Step 2: sort the genes in descending order of significance
Step 3: select top ranked k genes which satisfied with $POS_{top}(D) = POS_C(D)$
Step 4: finding all possible reducts of the top ranked k genes

Where $POS_{top}(D)$ is the positive region of the top ranked k genes, thus each reduct of the top ranked k genes is also a reduct of all genes.

This algorithm finds the most promising k attributes from the whole data set based on the attribute significance and positive region coverage, the reducts are then generated from these promising attributes, thus there is significant reduction in computational complexity and time required for getting reducts as compared to getting reducts from the entire gene space. Moreover, this algorithm takes into account the positive region coverage of all the genes, hence no information lost.

4.3 Classification

Decision rules as knowledge representation can be used to express relationship between condition attributes (or the antecedent) and decision attribute (or the consequent). The rough set theory as a mathematical tool provides an efficient mechanism to induce decision rules from decision table. After attribute reduction, each row in the decision table corresponds to a decision rule; rule set induced from the reduct can be used as the classifier to predict the class label of a new sample. In order to simplify the decision rules, value reduction method proposed in [5] is adopted.

In our study, we mix all the rules we obtained from each possible reduct together to be the classifier. When rules conflict, stand voting is used to select the best one. Leave-one out cross-validation (LOOCV), which is a widely used process for gene classification, is employed to evaluate the performance of classification process. With LOOCV, each object in data set will in turn be the test set, and the left are training set. Thus each sample of the data set will be predicted once by classifier trained with the left samples. All iterations are then averaged to obtain an unbiased number of performance estimates.

5 Experimental Results and Discussions

In this study, we have focus on four sets of gene expression data as summarized in Table1.They are described as follows:

Table 1. Data sets

Data set	Number of genes	Number of Classes	Number of samples
Acute Leukemia	7129	2(AML vs.ALL)	72(47ALL:25AML)
Colon Cancer	2000	2(tumor vs. normal)	62(22tumor:40normalL)
Lung Cancer	12533	2(MPM vs. ACDA)	181(31MPM:150ACDA)
DLBCL	4026	2(germinal vs. activated)	47(24germinal:23activated)

1. The Acute Leukemia data set [21] consists of samples from two different types of acute leukemia, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). There are 47 ALL and 25 AML samples; each sample has expression patterns of 7129 genes measured.

2. The Colon Cancer data set [22] is a collection of 62 gene expression measurement from biopsy samples generated using high density oligonucleotide microarrays. There are 22 normal and 40 tumor samples; each sample has expression patterns of 2000 genes measured.

3. The Lung Cancer data set [23] provides a collection of 181 gene expression samples, 150 for ACDA, 31 for PMP; each sample has expression patterns of 12533 genes measured.

4. The DLBCL data set [24] provides expression measurement of 47 samples, 24 for germinal, 31 for activated; each sample has expression patterns of 4026 genes measured.

For each data set, we select top ranked k genes(k =1,2,3,4,,5,6,7,8,9,10) to form the subset, and search for all possible reducts of all genes from them , rule sets generated from each reduct are mixed together to be the classifier for predicting the class label of new samples.

Table2 lists the top 10 genes we selected in each dataset. Most of them are also identified by the different other methods in the published literatures. For example, Krishnapuram et.al. selects 20 genes in leukemia data set and colon

Table 2. top 10 genes selected by our proposed method in each data set

No. of genes	Acute Leukemia	Colon Cancer	Lung Cancer	DLBCL
1	M23197	M76378	37157_at	GENE3330X
2	X95735	X12671	37954_at	GENE3261X
3	M28791	X56597	36533_at	GENE3327X
4	D88422_at	M22382	33328_at	GENE3967X
5	M84526_at	R87126	37716_at	GENE3329X
6	M92287	M16937	39640_at	GENE3256X
7	U46499	H28711	515_s_at	GENE3332X
8	M31523	U25138	1500_at	GENE3328X
9	M11722	D63874	38127_at	GENE3315X
10	L09209_s_at	M63391	179_at	GENE1252X

Table 3. the classification accuracy results with top ranked 1 to 10 genes

No. of genes	Acute Leukemia		Colon Cancer	
	Proposed criterion	MI	Proposed criterion	MI
1	98.6%	98.6%	93.5%	74.2%
2	100%	100%	95.2%	93.5%
3	100%	100%	98.4%	96.8%
4	100%	100%	98.4%	96.8%
5	100%	100%	98.4%	98.4%
6	100%	100%	98.4%	98.4%
7	100%	100%	98.4%	98.4%
8	100%	100%	98.4%	98.4%
9	100%	100%	98.4%	98.4%
10	100%	100%	98.4%	98.4%
No. of genes	Lung Cancer		DLBCL	
	Proposed criterion	MI	Proposed criterion	MI
1	-	86.2%	93.6%	78.7%
2	98.9%	90.6%	100%	91.5%
3	99.4%	92.3%	100%	97.9%
4	99.4%	95.6%	100%	100%
5	100%	96.7%	100%	100%
6	100%	96.7%	100%	100%
7	100%	96.7%	100%	100%
8	100%	98.3%	100%	100%
9	100%	98.3%	100%	100%
10	100%	98.3%	100%	100%

data set respectively and achieved 100% classification of all the samples [25], our selected gene M23197, M27891, M84526_at, M11722 from acute leukemia data set and M76378, R87126 from colon data set are among them. Gene M23197, M95735, M92287, L09209_s_at we selected from leukemia data set is also identified by Deb's work [26]. For DLBCL data set, Gene GENE3330X, GENE3328X are proved to be significant [18].These proves the effectiveness of our proposed method.

Table 3 summarizes the results, which demonstrate that our feature selection method can produces high significant features as the classification accuracy is very impressive. For acute leukemia data set, when using one gene X23197, all ALL samples can be correctly classified and the classification accuracy of AML samples is 96%, when using two or more genes, the overall classification accuracy 100% are achieved. For Lung cancer data set, the first gene selected is gene 37157_at, it can not form a reduct, when using 2 genes, ACDA samples can be fully classified(100%), and when using 5 or more genes, all the samples can be fully classified(100%). For colon data set, normal samples can be fully classified with the top ranked 2 genes, and we achieve the overall classification accuracy 98.4% when 3 or more genes are selected. For DLBCL data set, we achieve 93.6% and 100% classification accuracy respectively with the top ranked 2 genes. For comparison, experiment results when using MI to select the top ranked genes are also list. We can see that, for acute leukemia data set, both methods obtain the same results, for the other 3 data sets, our proposed method can obtain higher classification accuracy with fewer genes on each data set.

6 Conclusions

Gene expression data set has very unique characteristics which are very different from all the previous data used for classification. In order to achieve good classification performance, and obtain more useful insight about the biological related issues in cancer classification, gene selection should be well explored to both reduce the noise and avoid overfitting of classification algorithm.

This paper explores feature selection techniques based on rough set theory within the context of gene expression data for sample classification. We define a novel measurement for scoring genes according to the expression level differences between classes of a gene and its Mutual information, and then present a gene selection method based on rough sets theory, the method takes whole attributes set into account and extract all possible gene subsets from the top ranked informative genes that allow for sample classification with high accuracy. Experimental results on benchmark datasets indicate our method has successfully achieved its objectives: obtain high classification accuracy with a small number of high discriminative genes.

References

1. Frank, A.: A New Branch and Bound Feature Selection Algorithm. M.Sc. Thesis, submitted to Technion, Israel Institute of Technology (2002)
2. Xiong, M., Li, W., Zhao, J., Jin, L., Boerwinkle, E.: Feature (gene) Selection In Gene Expression-based Tumor Classication. Molecular Genetics and Metabolism 73, 239–247 (2001)
3. Wang, L.P., Feng, C., Xie, X.: Accurate Cancer Classifcation Using Expressions of Very Few Genes. EE/ACM Transactions on Computational Biology and Bioinformatics 4, 40–53 (2007)
4. Li, W., Yang, Y.: How Many Genes Are Needed For A Discriminant Microarray Data Analysis? In: Methods of Microarray Data Analysis. Kluwer academic Publisher, Norwell (2002)
5. Pawlak, Z.: Rough Set- Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dorderecht (1991)
6. Palawk, Z.: Rough Sets. International Journal of Computer and Information Science 11, 341–356 (1982)
7. Zhong, N., Dong, J., Ohsuga, S.: Using rough sets with heruristic for feature selection. Journal of Intelligent Information Systems 16, 119–214 (2001)
8. Mitra, S., Hayashi, Y.: Bioinformatics with Soft Computing. IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews 36, 616–635 (2006)
9. Hvidsten, T.R., Komorowski, J.: Rough Sets in Bioinformatics. In: Peters, J.F., Skowron, A., Marek, V.W., Orłowska, E., Słowiński, R., Ziarko, W.P. (eds.) Transactions on Rough Sets VII. LNCS, vol. 4400, pp. 225–243. Springer, Heidelberg (2007)
10. Midelfart, H., Komorowski, J., Nørsett, K., Yadetie, F., Sandvik, A.K., Lægreid, A.: Learning Rough Set Classifiers From Gene Expressions And Clinical Data. Fundamenta Inf. 53, 155–183 (2002)
11. Valdes, J.J., Barton, A.J.: Gene Discovery in Leukemia Revisited: A Computational Intelligence Perspective. In: Orchard, B., Yang, C., Ali, M. (eds.) IEA/AIE 2004. LNCS (LNAI), vol. 3029, pp. 118–127. Springer, Heidelberg (2004)
12. Momin, B.F., Mitra, S., Datta Gupta, R.: Reduct Generation and Classifcation of Gene Expression Data. In: Proceeding of First International Conference on Hybrid Information Technology (ICHICT 2006), pp. 699–708. IEEE Press, New York (2006)
13. Banerjee, M., Mitra, S., Banka, H.: Evolutinary-Rough Feature Selection in Gene Expression Data. IEEE Transaction on Systems, Man, and Cyberneticd, Part C: Application and Reviews 37, 622–632 (2007)
14. Wang, J., Wang, J.: Reduction Algorithms Based on Discernibly Matrix: The Ordered Attributes Method. Journal of Computer Science and Technology 16, 489–504 (2002)
15. Miao, D.Q., Hu, G.R.: A Heuristic Algorithm for Reduction of Knowledge. Journal of Computer Research and Development 36, 681–684 (1999)
16. Shen, Q., Chouchoulas, A.: A modular approach to generating fuzzy rules with reduced attributes for monitoring of complex systems. Engineering Applications of Artificial Intellegence 12, 263–278 (2000)
17. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Intelligent decision Support. Handbook of Applications and Advances of the Rough Sets Theory. Kluwer Academic, Dordrecht (1992)

18. Wang, Y., Tetko, I.V., Hall, M.A., Frank, E., Facius, A., Mayer, K.F.X., Mewes, H.W.: Gene Selection from Microarray Data for Cancer Classification-A Machine Learning Approach. Computational Biology and Chemistry 29, 37–46 (2005)
19. Zhou, W.G., Zhou, C.G., Liu, G.X., Wang, Y.: Artificial Intelligence Applications and Innovations. In: Proceeding of IFIP Intenational Federation for Information, pp. 492–499. Springer, Heidelberg (2006)
20. Ding, C., Peng, H.C.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. Journal of Bioinformatics and Computational Biology 3, 185–205 (2003)
21. 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., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286, 531–537 (1999)
22. Alon, U., Barkai, N., Notterman, D.A.: Broad Patterns of Gene Expression Revealed By Clustering Analysis of Tumor And Normal Colon Tissues Probed By Oligonucleotide Arrays. PNASUSA 96, 6745–6750 (1999)
23. Armstrong, S.A.: MLL Translocations Specify A Distinct Gene Distinguishes A Expression Profile That Unique Leukemia. Nature Genetics 30, 41–47 (2002)
24. Alizadeh, A.A., et al.: Distict types of diffuse large B-cell lymphoma identified by gene expressionprofiling. Nature 403, 503–511 (2000)
25. Krishnapuram, B., et al.: Joint classifier and feature selection optimization for Cancer diagnosis using gene expression Data. In: Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology, pp. 167–175. ACM, New York (2003)
26. Deb, K., Reddy, A.R.: Reliable Classifcation of Two Class Cancer Data Using Evolutionary Algorithms. BioSystems 72, 111–129 (2003)