# CANCER CLASSIFICATION USING SINGLE GENES

XIAOSHENG WANG[1] OSAMU GOTOH[1,2]
david@genome.ist.i.kyoto-u.ac.jp o.gotoh@i.kyoto-u.ac.jp

[1] *Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan*
[2] *National Institute of Advanced Industrial Science and Technology, Computational Biology Research Center, Tokyo 135-0064, Japan*

We present a method for the classification of cancer based on gene expression profiles using single genes. We select the genes with high class-discrimination capability according to their depended degree by the classes. We then build classifiers based on the decision rules induced by single genes selected. We test our single-gene classification method on three publicly available cancerous gene expression datasets. In a majority of cases, we gain relatively accurate classification outcomes by just utilizing one gene. Some genes highly correlated with the pathogenesis of cancer are identified. Our feature selection and classification approaches are both based on rough sets, a machine learning method. In comparison with other methods, our method is simple, effective and robust. We conclude that, if gene selection is implemented reasonably, accurate molecular classification of cancer can be achieved with very simple predictive models based on gene expression profiles.

*Keywords*: gene expression profiles; cancer classification; biomarkers; rough sets; decision rules

## 1. Introduction

Conventionally morphological diagnosis of tumor is not always effective as revealed by frequent occurrences of misdiagnoses. Recent molecular biological studies have elucidated that cancer was a disease involving dynamic changes in the genome. Moreover, the rapid advances in microarray technology have made it possible to simultaneously measure the expression levels of tens of thousands of genes in a single experiment. This technology has much facilitated the detection of cancerous molecular markers [13]. Accordingly, the use of biomarkers might be an alternative approach to the diagnosis of tumors. To this end, abundant explorations have been conducted to carry out cancer diagnosis, prognosis or prediction based on DNA microarray data since the pioneering work of Golub et al. [5].

Generally speaking, this field includes two key procedures: gene selection and classifier construction. The gene selection is particularly crucial in this topic as the number of genes irrelevant to classification may be huge, and hence, accurate prediction can be achieved only by performing gene selection reasonably, that is, identifying most informative genes from a large number of candidates. Once such genes are chosen, the creation of classifiers on the basis of the genes is another undertaking. If we survey the established investigations in this field, we will find that almost all the accurate classification results are obtained based on more than two genes. A very few investigators have attempted to address the problem by using gene pairs [6, 7]. However,

multi-gene models suffer from the disadvantage that it is not easy to assess which gene is more important in the models, because they are run on the basis of a group of genes. As a result, the significant biomarkers of related cancers are hard to be detected. In addition, multi-gene models are prone to impart the difficulty in understanding the models themselves.

Recently, we proposed a rough sets based soft computing method to conduct cancer classification using single or double genes [19]. In this article, we reevaluate the method by exploring the classification of cancer on the basis of single genes with three distinct datasets. We want to underscore that sufficiently accurate classification can be achieved, and important biomarkers can be found with ease using single-gene models. In addition, differing from the previous study, we estimate the classification accuracy rate by testing on independent samples, which is believed to be more unbiased than the cross validation.

## 2. Materials and Methods

### 2.1. *Datasets*

We use three datasets: leukemia, lung cancer and prostate cancer, which are available from the website: http://datam.i2r.a-star.edu.sg/datasets/krbd/. The gene number, class, training and test samples number contained in the three datasets are listed in Table 1.

Table 1. Summary of the three gene expression datasets.

| Dataset | # Genes | Class | # Training samples | # Test samples |
|---|---|---|---|---|
| leukemia | 7129 | ALL / AML | 38 (27 / 11) | 34 (20 / 14) |
| lung cancer | 12533 | MPM / ADCA | 32 (16 / 16) | 149 (15 / 134) |
| prostate cancer | 12600 | Tumor / Normal | 102 (52 / 50) | 34 (25 / 9) |

### 2.2. *Rough Sets*

In practice, we often want to dissect a collection of data to learn about their implications via the data with already known meaning. Yet, usually the significance of the analyzed data cannot be precisely explained as they incorporate vague components. In rough sets, the definite parts are described with the concept of positive region.

**Definition 1** Let $U$ be a universe of discourse, $X \subseteq U$, and $R$ is an equivalence relation on $U$. $U/R$ represents the set of the equivalence class of $U$ induced by $R$. The *positive region* of $X$ on $R$ in $U$, is defined as $pos(R,X) = \bigcup \{Y \in U/R \mid Y \subseteq X\}$.

The data studied by rough sets are mainly organized in the form of decision tables. One decision table can be represented as $S = (U, A=C \bigcup D)$, where $U$ is the set of samples, $C$ the condition attribute set and $D$ the decision attribute set. We can represent every cancer classification related microarray data with the decision table like Table 2. In the decision table, there are $m$ samples and $n$ genes. Every sample is assigned to one class label. Each gene is a condition attribute and each class is a decision attribute. $g(x, y)$ signifies the expression level of gene $y$ in sample $x$.

Table 2. Microarray data decision table.

| Samples | Condition attributes (genes) | | | | Decision attributes (classes) |
|---|---|---|---|---|---|
| | Gene 1 | Gene 2 | … | Gene *n* | Class label |
| 1 | *g(1, 1)* | *g(1, 2)* | ... | *g(1, n)* | *Class (1)* |
| 2 | *g(2, 1)* | *g(2, 2)* | ... | *g(2, n)* | *Class (2)* |
| … | ... | ... | ... | ... | ... |
| *m* | *g(m, 1)* | *g(m, 2)* | ... | *g(m, n)* | *Class (m)* |

Without loss of generality, we assume $D$ is a single-element set, and call $D$ the *decision attribute*. In $S$, we define $I_a$ as the function that maps a member of $U$ to its value on the attribute $a$ ($a \in A$), and an equivalence relation $R(A')$ induced by the attribute subset $A' \subseteq A$ as: for $s_1, s_2 \in U$, $s_1 R(A') s_2$ if and only if $I_a(s_1)=I_a(s_2)$ for each $a \in A'$. In rough sets, the *depended degree* of condition subset $P$ by the decision attribute $D$, written $\gamma_P(D)$, is defined as $\gamma_P(D)=\frac{|\mathrm{POS}_P(D)|}{|U|}$, where $|\mathrm{POS}_P(D)| = |\bigcup_{X \in U/R(D)} pos(P, X)|$ denotes the size of the union of the positive region of each equivalence class in $U/R(D)$ on $P$ in $U$, and $|U|$ signifies the size of $U$ (set of samples). The greater $\gamma_P(D)$ value often indicates the stronger classification power of $P$ [10, 7]. But it is not always the case, especially for the microarray data. In [19], we define *α depended degree* $\gamma_P(D,\alpha)=\frac{|\mathrm{POS}_P(D,\alpha)|}{|U|}$, where $0 \le \alpha \le 1$, $|\mathrm{POS}_P(D,\alpha)| = |\bigcup_{X \in U/R(D)} pos(P,X,\alpha)|$ and $pos(P,X,\alpha)=\bigcup \{Y \in U/P \mid |Y \cap X|/|Y| \ge \alpha\}$.

Inducing *decision rules* from decision tables is one of the main tasks in rough sets. One decision rule in the form of "$A \Rightarrow B$" indicates that "if $A$, then $B$", where $A$ is the description on condition attributes and $B$ the description on decision attributes. The *confidence* of $A \Rightarrow B$ is $\frac{\mathrm{support}(A \wedge B)}{\mathrm{support}(A)}$, where $\mathrm{support}(A)$ denotes the proportion of the samples satisfying $A$ and $\mathrm{support}(A \wedge B)$ the proportion of the samples satisfying $A$ and $B$ simultaneously. It reflects the reliable degree of one rule.

### 2.3. *Data Preprocessing, Gene Selection and Classification*

Before the learning algorithm is carried out, we discretize each original training set decision table by the entropy-based discretization method, proposed in [3]. We implement the discretization in the Weka package [17]. Every continuous-valued attribute is discretized into the attribute with no more than 3 different values. In addition, because there are microarray intensity discrepancies between the training set and the test set in the prostate cancer dataset caused by two different experiments, we normalize both the training and the test data. Each original expression level $g(x, y)$ is normalized to

$\frac{g(x,y)-(\max g(\bullet,y)+\min g(\bullet,y))/2}{(\max g(\bullet,y)-\min g(\bullet,y))/2}$, where $\max g(\bullet,y)$ and $\min g(\bullet,y)$ represent the maximum and the minimum expression levels of gene $y$ in all samples, respectively. After the normalization, all the gene expression levels are limited in interval [-1, 1]. For the other datasets, to avoid unnecessary loss of information, we do not conduct the process since the training and the test sets are from the same experiments.

We select informative genes based on $\alpha$ depended degree. Once $\alpha$ value is determined, we only choose the genes with $\gamma_P(D,\alpha)=1$ to build rule classifiers. Suppose $g$ is one of the selected genes and $U$ the sample set. $U/R(g)=\{c_1(g), c_2(g),\ldots, c_n(g)\}$ represents the set of the sample equivalence class induced by $R(g)$. Two samples $s_1$ and $s_2$ belong to the same equivalence class of $U/R(g)$ if and only if they have the same value on $g$. In addition, we represent the set of the sample equivalence class induced by $R(D)$ as $U/R(D)=\{d_1(D), d_2(D),\ldots, d_m(D)\}$, where $D$ is the class (decision) attribute. Likewise, two samples $s_1$ and $s_2$ belong to the same equivalence class of $U/R(D)$ if and only if they have the same value on $D$. For each $c_i(g)$ (i=1, 2,…, n), if there exists some $d_j(D)$ ($j\in\{1, 2,\ldots, m\}$), satisfying $|c_i(g)\cap d_j(D)|/|c_i(g)|\geq\alpha$, then we generate the classification (decision) rule: $\Phi(c_i(g))\Rightarrow\Phi(d_j(D))$, where $\Phi(c_i(g))$ is the description of sample subset $c_i(g)$ by $g$ value and $\Phi(d_j(D))$ is the description of sample subset $d_j(D)$ by the class value. It is noted that the confidence of every classification rule produced by the way is no less than $\alpha$ [19]. Thus, we can ensure sufficient reliability of the derived classification rules by setting high threshold of $\alpha$ value.

After data preprocessing, we carry out gene selection and classifier construction. Initially, we select high class-discrimination genes by the measure of $\alpha$ depended degree. We begin with $\alpha$=1, then gradually decrease $\alpha$ value. If we are lucky enough, we may stop the selection step at the point of very high $\alpha$ value. Otherwise, more search steps are needed. In the worst case, we will stop attempts at the point of $\alpha$=0.7, which is the lower bound. The genes with $\gamma_P(D,\alpha)=1$ are picked out. Next, we create the classifiers based on the decision rules induced by the selected genes, and apply the classifiers for independent test sets to validate the classification performance.

## 3. Results

### 3.1. *Classification Results*

In the leukemia dataset, when $\alpha$=1, gene #4847 is identified; when $\alpha$=0.95, gene #4847, #1926 and #1882 are identified; when $\alpha$=0.90, 25 genes are identified. With the decreasing of $\alpha$, more and more genes are identified. Given $\alpha\geq 0.90$, 25 genes are finally marked. Among the 25 genes, 8 genes have the classification accuracy no less than 85%, of which gene #1882 and #1834 have 94% accuracy, and gene #4847 and #760 possess 91% accuracy. We denote the expression level of gene x by g(x). Two decision rules induced by gene #4847 are: if g(#4847) > 994, then AML; if g(#4847) ≤ 994, then ALL.

Both rules have 100% confidence. By the rules, we obtain 91% classification accuracy in the test set. Likewise, gene #1882 induces two rules: if g(#1882) > 1419.5, then AML; if g(#1882) ≤ 1419.5, then ALL, which have 100% and 96% confidence respectively. 94% classification accuracy is achieved in the test set by the rules. Table 3 summarizes the information on the 8 genes with no less than 85% classification accuracy.

Table 3. Genes with high classification accuracy in the leukemia dataset.

| No.[a] | Accession | #Correctly-classified samples[b] | Classification accuracy (%)[c] | $\alpha$ |
|---|---|---|---|---|
| 4847 | X95735_at | 31 (18 / 13) | 91 (90 / **93**) | 1 |
| 1882 | M27891_at | 32 (19 / 13) | **94** (95 / **93**) | 0.95 |
| 760 | D88422_at | 31 (20 / 11) | 91 (**100** / 79) | 0.9 |
| 1834 | M23197_at | 32 (20 / 12) | **94** (**100** / 86) | 0.9 |
| 2402 | M96326_rna1_at | 29 (18 / 11) | 85 (90 / 79) | 0.9 |
| 4373 | X62320_at | 30 (19 / 11) | 88 (95 / 79) | 0.9 |
| 6376 | M83652_s_at | 30 (19 / 11) | 88 (95 / 79) | 0.9 |
| 6855 | M31523_at | 30 (20 / 10) | 88 (**100** / 71) | 0.9 |

[a] The order number of attributes (genes) in the decision table. [b] The number of correctly-classified samples in total and with respect to every class (presented in parentheses). [c] The classification accuracy in whole and in every class (presented in parentheses).

Table 4. Genes with high classification accuracy in the lung cancer dataset.

| No. | Accession | # Correctly-classified samples | Classification accuracy (%) | $\alpha$ |
|---|---|---|---|---|
| 2549 | 32551_at | 134 (14 / 120) | 90 ( 93 / 90 ) | 1 |
| 3250 | 33245_at | 137 (14 / 123) | 92 ( 93 / 92 ) | 1 |
| 3844 | 33833_at | 139 (13 / 126) | 93 ( 87 / 94 ) | 1 |
| 6571 | 36533_at | 141 (13 / 128) | 95 ( 87 / 96 ) | 1 |
| 7249 | 37205_at | 135 (12 / 123) | 91 ( 80 / 92 ) | 1 |
| 7765 | 37716_at | 145 (11 / 134) | 97 ( 73 / **100** ) | 1 |
| 9863 | 39795_at | 135 (14 / 121) | 91 ( 93 / 90 ) | 1 |
| 11015 | 40936_at | 140 (12 / 128) | 94 ( 80 / 96 ) | 1 |
| 12114 | 575_s_at | 141 (14 / 127) | 95 ( 93 / 95 ) | 1 |
| 541 | 1500_at | 145 (13 / 132) | 97 ( 87 / 99 ) | 0.9 |
| 633 | 1585_at | 138 (13 / 125) | 93 ( 87 / 93 ) | 0.9 |
| 869 | 179_at | 137 (14 / 123) | 93 ( 93 / 92 ) | 0.9 |
| 2421 | 32424_at | 145 (11 / 134) | 97 ( 73 / **100** ) | 0.9 |
| 3333 | 33327_at | 143 (14 / 129) | 96 ( 93 / 96 ) | 0.9 |
| 3916 | 33904_at | 138 (14 / 124) | 93 ( 93 / 93 ) | 0.9 |
| 4336 | 34320_at | 144 (14 / 130) | 97 ( 93 / 97 ) | 0.9 |
| 5301 | 35276_at | 145 (14 / 131) | 97 ( 93 / 98 ) | 0.9 |
| 7200 | 37157_at | 146 (12 / 134) | **98** ( 80 / **100** ) | 0.9 |
| 8005 | 37954_at | 140 (14 / 126) | 94 ( 93 / 94 ) | 0.9 |
| 8537 | 38482_at | 139 (15 / 124) | 93 (**100** / 93 ) | 0.9 |
| 9474 | 39409_at | 138 (13 / 125) | 93 ( 87 / 93 ) | 0.9 |
| 9698 | 39631_at | 134 (13 / 121) | 90 ( 87 / 90 ) | 0.9 |
| 11841 | 41755_at | 139 (10 / 129) | 93 (67 / 96 ) | 0.9 |
| 11958 | 41871_at | 142 (10 / 132) | 93 (67 / 99 ) | 0.9 |
| 12200 | 661_at | 142 (12 / 130) | 93 (80 / 97 ) | 0.9 |

In the lung cancer dataset, when $\alpha$=1 or 0.95, 16 genes are identified; when $\alpha$=0.90, 56 genes are identified. Among the 56 genes, gene #1223 has 98% classification accuracy with the classification rules: if g(#1223) > 490.5, then MPM; if g(#1223) ≤ 490.5, then ADCA. Both rules have 100% and 94% confidence, respectively. The genes with classification accuracy no less than 90% are presented in Table 4.

In the prostate cancer dataset, when $\alpha$=1, 0.95 or 0.9, no any gene is identified; when $\alpha$=0.85, only gene #1315 is identified; when $\alpha$=0.8, 11 genes are marked. Among the 11 genes, genes #827 and #1266 have the highest classification accuracy of 91%. The classification rules induced by them are: if g(#827) > -0.612121, then Normal (80% confidence); if g(#827) ≤ -0.612121, then Tumor (89% confidence); and if g(#1266) > -0.543166, then Normal (81% confidence); if g(#1266) ≤ -0.543166, then Tumor (93% confidence). The genes with classification accuracy no less than 70% are presented in Table 5.

Table 5. Genes with high classification accuracy in the prostate cancer dataset.

| No. | Accession | # Correctly-classified samples | Classification accuracy (%) | $\alpha$ |
|---|---|---|---|---|
| 5757 | 36491_at | 30 (23 / 7) | 88 (**92** / 78) | 0.8 |
| 7557 | 32243_g_at | 31 (22 / 9) | **91** (88 / **100**) | 0.8 |
| 9050 | 38044_at | 29 (21 / 8) | 85 (84 / 88) | 0.8 |
| 10138 | 41288_at | 31 (22 / 9) | **91** (88 / **100**) | 0.8 |
| 10956 | 1767_s_at | 24 (22 / 2) | 71 (88 / 22) | 0.8 |
| 12148 | 575_s_at | 27 (18 / 9) | 79 (72 / **100**) | 0.8 |

### 3.2. *Comparison of Classification Results*

The leukemia dataset has been well studied by many researchers [4, 5, 7, 8, 10,12]. Although there are a few reports on the use of a single gene to distinguish the AML from the ALL, a majority of investigators conduct the classification with more than one gene, even tens or hundreds. In [8, 15, 18], the authors present the classification outcomes of 31 out of 34 samples correctly classified with one common gene: Zyxin. Yet, we correctly classify 32 samples using a single gene. Moreover, Zyxin (X95735_at) is also identified by our approach, by which we correctly classify 31 samples as well.

Regarding the three datasets, the best classification results reported in our and some other works are shown in Table 6, 7 and 8, respectively. These tables demonstrate that our single-gene classifiers perform comparatively well in these datasets. If using single genes, our accuracy is the highest among all the methods, and the other methods must use far more genes to reach or slightly surpass our accuracy.

Table 6. Comparison of the best classification accuracy for the leukemia dataset.

| Methods ( feature selection + classification)[d] | # Selected genes | # Correctly-classified samples (accuracy) |
|---|---|---|
| $\alpha$ depended degree + decision rules [this work] | **1** | **32 (94.1%)** |
| t-test, attribute reduction + decision rules [15] | 1 | 31 (91.2%) |
| rough sets, GAs + k-NN [1] | 9 | 31 (91.2%) |
| EPs [8] | 1 | 31 (91.2%) |
| discretization + decision trees [16] | 1038 | 31 (91.2%) |
| CBF + decision trees [18] | 1 | 31 (91.2%) |
| RCBT [2] | 10-40 | 31 (91.2%) |
| neighborhood analysis + weighted voting [5] | 50 | 29 (85.3%) |
| prediction strength + SVMs [4] | 25-1000 | 30-32(88.2%-94.1%) |

[d] The text before "+" states the feature selection method while that after it states the classification method. The absence of "+" means both methods can not be separated clearly.

Table 7. Comparison of best classification accuracy for the lung cancer dataset.

| Methods ( feature selection + classification) | # Selected genes | # Correctly-classified samples (accuracy) |
|---|---|---|
| $\alpha$ depended degree + decision rules [this work] | **1** | **146 (98%)** |
| attribute reduction + k-NN [12] | 2 | 146 (98%) |
| PCLs [9] | unknown | 146 (98%) |
| C4.5 [9] | 1 | 122 (81%) |
| Bagging [9] | unknown | 131 (88%) |
| Boosting [9] | unknown | 122 (81%) |
| discretization + decision trees [16] | 5365 | 139 (93%) |
| RCBT [2] | 10-40 | 146 (98%) |

Table 8. Comparison of best classification accuracy for the prostate cancer dataset.

| Methods ( feature selection + classification) | # Selected genes | # Correctly-classified samples (accuracy) |
|---|---|---|
| $\alpha$ depended degree + decision rules [this work] | **1** | **31 (91%)** |
| PCLs [9] | unknown | 33 (97%) |
| discretization + decision trees [16] | 3071 | 25 (73.53 %) |
| RCBT [2] | unknown | 33 (97%) |
| SVMs [2] | unknown | 27 (79.41%) |
| signal to noise ratios + k-NN [14][e] | 4 | 26 (77.2%) |
| | 16 | 29 (85.7%) |

[e] We compare their results from normalized dataset. For facilitating comparison, the correctly-classified sample numbers are calculated according to the total of 34 instead of 35 samples used in [14].

### 3.3. *Analysis of Results*

In each dataset, we identify some highly-discriminative genes. These genes might be able to provide insight into the pathogenesis of specific or general tumors. In our models, the rules in the form of "if g(x)>t, then y" indicate that gene x is upregulated in the

samples with class y. In contrast, the rule like "if g(x)≤t, then y" imply that gene x is downregulated in the samples with class y.

In terms of this standard, among the eight genes identified in the leukemia dataset, seven are upregulated and one is downregulated in AML. The seven upregulated genes include CST3, CD33, Zyxin, Azurocidin, PPF, Granulin and CYSTATIN A, and the downregulated one is TCF3. The first five of the seven genes also lie in the list of the 50 informative genes distinguishing ALL from AML chosen by Golub et al [5] and are marked as highly expressed genes in AML. Zyxin is the only chosen gene when $\alpha$=1. It is also frequently selected by other learning algorithms [4, 5, 8, 15, 18]. Some investigations reveal Zyxin might play an important role in leukemia pathogenesis. CD33 is one of the two genes with the best classification performance identified by us. Indeed, CD33 is an important biomarker of AML [5, 11]. CST3 is another gene with the highest classification accuracy. Although it has not been found to be associated with AML or ALL directly, it was chosen as one of the most discriminative genes distinguishing AML from ALL by some authors [5, 18]. Granulin is correlated with the pathogenesis of various tumors. CYSTATIN A is also reported to be correlated with the prognosis and diagnosis of cancer.

In the lung cancer dataset, the gene with the strongest classification ability is Calretinin (98% accuracy), which is identified when $\alpha$=0.9. Our rules show Calretinin has higher expression levels in MPM, which is consistent with the reports from [6]. In fact, the gene has been recognized as one of the most important biomarkers in the diagnosis of MPM and lung cancer [6]. The genes with the second best classification performance include WT1, MRC OX-2, HAS1, Claudin 4 et al., all with 97% classification accuracy. Many investigations have revealed WT1 linked with the malignancy of tumors. That is, high levels of WT1 have been connected with poor prognosis in various cancers. Our rules indicate that WT1 has elevated expression in MPM, which is a highly lethal malignancy relative to ADCA. Thus our rules are reasonable. Likewise, our rules imply that both MRC OX-2 and HAS1 are overexpressed in MPM, conforming to the past studies [6]. In contrast, Claudin 4 is underexpressed in MPM while overexpressed in ADCA in our rules, which imply that the overexpression of Claudin 4 means a better prognosis.

In the prostate cancer dataset, our rules indicate that Thymosin beta is overexpressed in tumors. Some investigations have revealed that it was associated with the pathology of several cancers. TGFβ is a multifunctional peptide that controls proliferation, differentiation, and other functions in many cell types. Hence, its dysregulation may be concerned with various cancer types. By our rules, TGFβ is underexpressed in tumors. That is also revealed in [14]. GA733 is the gene encoding a carcinoma-associated antigen. Many investigations have revealed its overexpression in diverse tumors. Without exception, our rules indicate GA733 is highly expressed in prostate tumors.

## 4. Discussion

The principal advantage of our single-gene models is that the predication procedures and results are understood with ease, because our models are based on rules and our rules

are built by single genes. Obviously, biologists and clinicians prefer "rules" to "non-rules". Further, they favor simple rules more than complicated rules. Whereas some rule-based models do well in prediction, they are not inclined to be adopted as their rules are created via many features (genes) so that it is quite difficult to understand the rules. In contrast, our single-gene derived rules are fairly simple and concise.

Our models are both simple and effective. The efficacy of our models has been proven through their application to several noted microarray datasets. The results manifest that not only accurate classification of cancer can be achieved, but also biologically important genes can be identified with the models. Moreover, our models are rather robust. We select informative genes based on $\alpha$ depended degree instead of depended degree originally proposed in rough sets. Indeed, if we use the conventional depended degree standard, many important genes will be neglected for their depended degrees are often fairly lower, even zero. But their lower depended degrees are frequently caused by a small number of exceptional instances. If we tolerate the exceptions, we will find that these genes are indeed significant in class discrimination. Hence, we develop $\alpha$ depended degree standard to address the problem. By the operation of $\alpha$ value, we are able to not only select authentically important genes, but also control the size of selected genes as well as the confidence of classification rules. Accordingly, we can adjust our models to meet different datasets.

One might doubt the utility of our models as he holds that cancerous pathogenesis is so complex that it must be connected with many genes instead of just one. Our findings do not contradict this argument in that our methods can mark many significant genes solely, each of which could be the candidate biomarker of cancer. Clearly, our methods might achieve the goal of finding the possible molecular markers of cancer more easily compared with multi-gene models because when good predication is obtained by multi-gene models, it is difficult to gauge which genes are essential in cancerous pathogenesis.

## References

[1] Banerjee, M., Mitra, S., and Banka, H., Evolutinary-rough feature selection in gene expression data, *IEEE Transaction on Systems, Man, and Cybernetics, Part C: Application and Reviews*, 37:622–632, 2007.

[2] Cong, G., Tan, K.-L., Tung, A., and Xu, X., Mining top-k covering rule groups for gene expression data, *Proc. ACM SIGMOD International Conference on Management of Data*, 670-681, 2005.

[3] Fayyad, U.-M., and Irani, K.-B., Multi-interval discretization of continuous-valued attributes for classification learning, *Proc. 13th International Joint Conference of Artificial Intelligence*, 1022-1027, 1993.

[4] Furey, T.-S., Cristianini, N., Duffy, N., Bednarski, D.-W., Schummer, M., and Haussler, D., Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, 16(10):906-914, 2000.

[5] 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., et al., Molecular

classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286(5439):531-537, 1999.

[6] Gordon, G.-J., Jensen, R.-V., Hsiao, L.-L., Gullans, S.-R., Blumenstock, J.-E., Ramaswamy, S., Richards, W.-G., Sugarbaker, D.-J., and Bueno, R., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma, *Cancer Res*, 62(17):4963-4967, 2002.

[7] Geman, D., d'Avignon, C., Naiman, D.-Q., and Winslow, R.-L., Classifying gene expression profiles from pairwise mRNA comparisons, *Stat Appl Genet Mol Biol* 3:Article19, 2004.

[8] Li, J., and Wong, L., Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns, *Bioinformatics*, 18(5):725-734, 2002.

[9] Li, J., and Wong, L., Using rules to analyse bio-medical data: a comparison between C4.5 and PCL, *Advances in Web-Age Information Management*, 254-265, 2003.

[10] Li, D., and Zhang, W., Gene selection using rough set theory, *Proc. 1st International Conference on Rough Sets and Knowledge Technology*, 778–785, 2006.

[11] Lamba, J.-K., Pounds, S., Cao, X., Downing J.-R., Campana, D., Ribeiro, R.-C., Pui, C.H., and Rubnitz, J.-E., Coding polymorphisms in CD33 and response to gemtuzumab ozogamicin in pediatric patients with AML: a pilot study. *Leukemia*, 23(2):402-404, 2009.

[12] Momin, B.-F., and Mitra, S., Reduct generation and classification of gene expression data, *Proc. 1st International Conference on Hybrid Information Technology*, 699-708, 2006.

[13] Schena, R., Shalon, D., Davis, R.-W., and Brown, P.-O., Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270(5235):467-470, 1995.

[14] Singh, D., Febbo, P.-G., Ross K, Jackson D.-G., Manola, J., Ladd, C., Tamayo, P., Renshaw A.-A., D'Amico A.-V., Richie J.-P., et al., Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203-209, 2002.

[15] Sun, L., Miao, D., and Zhang, H., Efficient gene selection with rough sets from gene expression data, *Proc. 3rd International Conference on Rough Sets and Knowledge Technology*, 164–171, 2008.

[16] Tan, A.-C., and Gilbert, D., Ensemble machine learning on gene expression data for cancer classification, *Appl Bioinformatics*, 2(3 Suppl):S75-83, 2003.

[17] Witten, I.-H., and Frank, E., *Data mining: practical machine learning tools and techniques (second edition)*, Morgan Kaufmann, 2005.

[18] Wang, Y., Tetko, I.-V., Hall, M.-A., Frank, E., Facius, A., Mayer, K.-F., and Mewes H.-W., Gene selection from microarray data for cancer classification--a machine learning approach, *Comput Biol Chem*, 29(1):37-46, 2005.

[19] Wang, X., and Gotoh, O., Microarray-Based Cancer Prediction Using Soft Computing Approach, *Cancer Informatics*, 7:123–139, 2009.