数理モデル化と問題解決 32−17
（2000. 11. 22）

遺伝子発現パターンからの細胞分類アルゴリズム

阿久津 達也　宮野 悟

東京大学医科学研究所ヒトゲノム解析センター

DNA マイクロアレイなどにより得られた遺伝子の発現量データを解析することにより、がん細胞などの分類を行なおうという研究が行なわれている。分類を行なう際には、利用可能な全遺伝子の発現量データを用いることは少なく、分類のために有用な情報が得られる遺伝子のみを選択して分類を行なうことが多い。本研究では、発現量を0か1に丸めることにより、この有用遺伝子の選択問題をブール関数の閾値関数の学習問題として定義し、その問題に対し、単純かつ有効な貪欲算法型アルゴリズムを示す。そして、公開されている白血病細胞の発現量の実データを用いて、このアルゴリズムを他の2種類のアルゴリズムと比較した結果について示す。その結果、残念ながらテストデータに対しては大きな差が得られなかったが、学習データに対しては提案するアルゴリズムがより良い分類精度を示すことがわかった。

Selecting Informative Genes for Cancer Classification Using Gene Expression Data

Tatsuya Akutsu　Satoru Miyano

Human Genome Center, Institute of Medical Science, University of Tokyo
4-6-1 Shirkanedai, Minato-ku, Tokyo 108-8639, Japan
{takutsu,miyano}@ims.u-tokyo.ac.jp

Recently, several methods have been proposed for classification of cancer cells based on gene expression monitoring by DNA microarrays. In these methods, not all genes were used for classification, but several tens of genes that were relevant to class distinction were selected and used. In this article, this selection problem is formalized using threshold functions for Boolean variables. A simple greedy algorithm is also proposed for the selection problem. This greedy algorithm was compared with two other algorithms using real gene expression data obtained from human acute leukemia patients by Golub *et al.* The results of comparison show that the greedy algorithm is as good as the other two algorithms for the test data set and is much better for the training data set.

1 Introduction

Accurate classification of tumor types is very important already-defined classes. Recently, a new approach [2, 4] based on global gene expression analysis using DNA microarrays [1] has been proposed. Although effectiveness of such an approach was already demonstrated, the proposed information processing methods [2, 4] were rather heuristic. Therefore, further studies should be done for making more accurate classification.

Golub *et al.* [2] divided cancer classification into two problems: *class discovery* and *class prediction*. Class discovery is to define previously unrecognized tumor subtypes, whereas class prediction is to assign particular tumor samples to already-defined classes. Although class discovery is more challenging, we consider class prediction here because class prediction seems to be more basic and thus information processing methods for class prediction should be established

earlier. In the previous methods [2, 4], predictions were made by means of the *weighted voting*. All genes were not used for weighted votes, but several tens of genes relevant to class distinction were selected and used. Use of selected genes seems better because of several reasons. For example, computational cost for determining parameters and cost for measurement of gene expression levels are much lower if the selected genes are used. Golub *et al.* called these selected genes *informative genes*. However, previous selection methods of informative genes [2, 4] were rather heuristic. Indeed, Golub *et al.* [2] wrote that *the choice to use 50 informative genes in the predictor was somewhat arbitrary*. Therefore, we focus on the selection problem.

In this article, we treat this selection problem as an inference problem of *threshold functions* for Boolean variables. We treat class prediction as a problem of deciding whether or not a given sample belongs to the target class. Note that class prediction with multiple classes can be treated by making class prediction for each class independently. We do not use real values because it is difficult to give an appropriate mathematical definition using real values and it is widely recognized that gene expression data obtained by DNA microarrays contain large noises. Instead, each value is simplified to either 1 (high expression level) or 0 (low expression level), where the method of simplification is omitted in this article.

Since threshold functions are useful, many studies have been done in the field of machine learning. Among them, the WINNOW algorithm [3] is famous. We applied WINNOW to selection of informative genes. However, the results were not satisfactory. Therefore, we developed a new algorithm. We compared the algorithm with WINNOW and a very simple algorithm, using gene expression data [2] obtained from human acute leukemia patients.

2 Definition of the Problem

Let $\{g_1, \ldots, g_n\}$ denote the set of genes. Let $\{s_1, \ldots, s_m\}$ denote the set of samples from patients. Assume that it is known whether each sample s_j belongs to the target cancer class. We let $class(s_j) = 1$ if s_j belongs to the class, otherwise we let $class(s_j) = 0$. Let $x_{i,j}$ be the expression level (either 0 or 1) of gene i for sample j. We formalize the selection problem using threshold functions for Boolean variables. We use *r-of-k threshold functions* [3]. An r-of-k threshold function $f(z_1, \ldots, z_n)$ is defined by selecting a set of k significant variables. The value of f is 1 whenever at least r of these k variables are 1. If the k selected variables are $z_{i_1}, \ldots, z_{i_k}$, then f is 1 exactly when $z_{i_1} + \ldots + z_{i_k} \geq r$. For example, consider a case of $n = 5$, $k = 3$, $r = 2$ and $i_1 = 1, i_2 = 2, i_3 = 5$. Then, $f(1,1,1,1,1) = 1$, $f(0,0,0,0,0) = 0$, $f(1,0,1,1,1) = 1$, $f(1,0,1,1,0) = 0$, and $f(1,1,0,0,0) = 1$.

We define the selection problem of informative genes as follows. Assume that expression data and k are given as an input. Then, the problem is to determine a set of k genes $\{g_{i_1}, \ldots, g_{i_k}\}$ which maximizes r under the condition that $x_{i_1,j} + \ldots + x_{i_k,j} \geq r$ if $class(s_j) = 1$, otherwise $(1 - x_{i_1,j}) + \ldots + (1 - x_{i_k,j}) \geq r$. The latter case means that at least r variables must be 0 if the corresponding sample does not belong to the target class. It is expected that predictions can be done more robustly if r is larger.

3 A Simple Greedy Algorithm

The inference of an r-of-k function consistent with training data is known to NP-hard. Therefore, development of heuristic algorithms is a reasonable choice. We developed a kind of greedy algorithm. This algorithm is denoted by GREEDY in this article.

GREEDY maintains non-negative real-valued weights $w_1, \ldots, w_m$, where the weights are not assigned to genes, but are assigned to samples. We say that gene g_i *covers sample* s_j if $class(s_j) =$

$x_{i,j}$. First, genes which do not cover most samples are removed and are not considered as candidates for informative genes. Precisely, any gene g_i such that $\#\{s_j | class(s_j) \neq x_{i,j}\} > \theta_0$ is removed, where θ_0 is a threshold and we are currently using $\theta_0 = 7 \sim 9$. Next, GREEDY selects informative genes iteratively (i.e., one gene is selected per iteration). For h-th iteration, gene g_{i_h} which maximizes the score is selected. The score is defined by $\sum_{j=1}^{m} \left(\beta^{\delta(class(s_j), x_{i_h,j}) \times (h - w_j)}\right)$, where $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. β is a constant defined based on the experience and we are currently using $\beta = 1.5$. Weight w_j is increased by 1 if the selected gene g_i contributes to the classification of sample s_j. Precisely, w_j is updated by $w_j \leftarrow w_j + 1$ if $x_{i_h,j} = class(s_j)$. Thus, w_j represents the number of genes (among $g_{i_1}, \ldots, g_{i_h}$) which cover s_j. GREEDY tries to cover each sample as many times as possible. The following is the description of the GREEDY algorithm. It is easy to see that GREEDY runs in $O(kmn)$ time.

1. Remove all g_i such that $err(g_j) > \theta_0$, where $err(g_i) = \#\{s_j | x_{i,j} \neq class(s_j)\}$.
2. Let $w_i \leftarrow 0$ for all $i = 1, \ldots, m$.
3. For $h = 1$ to k, execute STEP 4 and STEP 5.
4. Select g_{i_h} maximizing the score $\sum_{j=1}^{m} \left(\beta^{\delta(class(s_j), x_{i_h,j}) \times (h - w_j)}\right)$, where $g_{i_h} \notin \{g_{i_1}, \ldots, g_{i_{h-1}}\}$.
5. Let $w_j \leftarrow w_j + 1$ for all j such that $x_{i_h,j} = class(s_j)$.

4 Computational Results

We compared GREEDY with WINNOW and SIMPLE, where SIMPLE selects genes with k smallest $err(g_i)$ values. For implementation and comparison, we used a PC with a 700 MHz AMD Athron processor. In each case, the inference could be done within ten seconds.

We used the data set obtained from *acute leukemias* patients by Golub *et al* [2]. Acute leukemias are basically classified into two classes: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). They used two data sets: one for training and the other for test. The training data set (TR) consisted of 38 samples (27 ALL, 11 AML) and the test data set (TS) consisted of 34 samples (20 ALL, 14 AML). For each sample, expression levels for 6817 genes were measured by microarrays produced by Affymetrix.

We examined the following three cases. (A) TR and TS. (B) TR was used as the test data set and TS was used as the training data set. (C) It is known that ALL samples are further classified into T-cell ALL and B-cell ALL [2]. 27 ALL (19 B-cell ALL, 8 T-cell ALL) samples from TR were used as training data and 20 ALL (19 B-call ALL, 1 T-cell ALL) samples from TS were used as test data.

Since the rounded Boolean values were biased when the number of the training samples belonging to the target class was different from the number of the other training samples, we made the numbers to be equal by duplicating samples not belonging to the target class.

For each of SIMPLE, WINNOW and GREEDY, we examined four cases: $k = 20$, $k = 30$, $k = 40$ and $k = 50$. Recall that k is the number of informative genes to be selected.

First we measured the qualities of the sets of informative genes by means of r. In Table 1, the maximum r computed from each set of informative genes is shown. Recall that we defined the selection problem as a maximization problem on r. From this table, it is seen that GREEDY is much better than SIMPLE and WINNOW.

Next, we made computational experiments on predictions. Using the informative genes computed by each algorithm, we made predictions on both the samples in the training data set

Table 1: Qualities of the sets of informative genes measured by r. For each case, the maximum r computed from each set of informative genes is shown.

		$k = 20$	$k = 30$	$k = 40$	$k = 50$
(A)	SIMPLE	15	19	24	29
	WINNOW	14	20	24	29
	GREEDY	16	24	33	40
(B)	SIMPLE	13	17	22	29
	WINNOW	12	18	22	27
	GREEDY	17	25	34	43
(C)	SIMPLE	10	15	18	21
	WINNOW	9	15	20	27
	GREEDY	15	23	31	38

Table 2: Comparison of the selection algorithms. For each case, the number of samples that were assigned as uncertain is shown, where the number after the symbol '+' denotes the number of samples that were classified into the wrong class.

		TRAINING				TEST			
		$k = 20$	$k = 30$	$k = 40$	$k = 50$	$k = 20$	$k = 30$	$k = 40$	$k = 50$
(A)	SIMPLE	0	1	1	1	6+1	7	5+1	7
	WINNOW	0	0	1	1	5+1	7	6+1	7
	GREEDY	0	0	0	0	7+1	9	6+1	7
(B)	SIMPLE	2	2	2	2	5	6	7	9
	WINNOW	2	2	2	2	5	7	7	11
	GREEDY	0	0	0	0	5	7	7	8
(C)	SIMPLE	0	1	1	1	1	1	1	1
	WINNOW	1	1	1	1	1	1	1	1
	GREEDY	0	0	0	0	1	1	1	1

and the samples in the test data set, by means of the majority voting. Results of the predictions are shown in Table 2. In Table 2, the number of samples assigned as uncertain is shown for each case. The number of samples which were classified into the wrong class is also shown (after the symbol '+'). For example, 5+1 means that 5 samples were assigned as uncertain and 1 sample was classified into the wrong class. From this table, it is seen that GREEDY always made correct predictions for samples in the training data set. However, for the test data set, there was no significant difference among SIMPLE, WINNOW and GREEDY.

References

[1] J.L. DeRisi, V.R. Lyer and P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[2] T.R. Golub *et al*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531–537, 1999.

[3] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear-threshold algorithm, *Machine Learning*, 2:285–318, 1988.

[4] T. Tsunoda *et al*, Diagnosis system of drug sensitivity of cancer using cDNA microarray and multivariate statistical analysis, in *Currents in Computational Molecular Biology* (Universal Academy Press, Tokyo), 16–17, 2000.