나이브 베이스 분류기를 이용한 유전발현 데이타기반 암 분류를 위한 순위기반 다중클래스 유전자 선택

(Rank-based Multiclass Gene Selection for Cancer Classification with Naive Bayes Classifiers based on Gene Expression Profiles)

홍 진 혁 † 조 성 배 ††
(Jin-Hyuk Hong) (Sung-Bae Cho)

요 약 최근 활발히 연구가 진행 중인 유전발현 데이타를 이용한 다중클래스 암 분류는 DNA 마이크로어레이로부터 획득된 대규모의 유전자 정보를 분석하여 암의 종류를 판단한다. 수집된 유전발현 데이타에는 대상 암과 관련이 없는 유전자도 포함되어 있기 때문에 높은 성능의 분류 결과를 얻기 위해서 유용한 유전자를 선택하는 것이 필요하다. 기존의 순위기반 유전자 선택은 이진클래스를 대상으로 고안되었고 이상표식 유전자(Ideal marker gene)를 이용하기 때문에 다중클래스 암 분류에 직접 적용하기에는 한계가 있다. 본 논문에서는 이상표식 유전자를 사용하지 않고 유전발현 수준의 분포를 직접 분석하는 순위기반 다중클래스 유전자 선택 기법을 제안한다. 유전발현 수준을 이산화하고 학습데이타로부터 빈도를 계산하여 클래스 간 분별력을 측정한 후, 선택된 유전자를 이용하여 나이브 베이즈 분류기를 사용해 다중 암 분류를 수행한다. 제안하는 방법을 다수의 다중클래스 암 분류 데이타에 적용하여 기존 유전자 선택 방법에 비해 우수함을 확인하였다.

키워드 : 유전발현 데이타, 다중부류 암 분류, 유전자 선택

Abstract Multiclass cancer classification has been actively investigated based on gene expression profiles, where it determines the type of cancer by analyzing the large amount of gene expression data collected by the DNA microarray technology. Since gene expression data include many genes not related to a target cancer, it is required to select informative genes in order to obtain highly accurate classification. Conventional rank-based gene selection methods often use ideal marker genes basically devised for binary classification, so it is difficult to directly apply them to multiclass classification. In this paper, we propose a novel method for multiclass gene selection, which does not use ideal marker genes but directly analyzes the distribution of gene expression. It measures the class-discriminability by discretizing gene expression levels into several regions and analysing the frequency of training samples for each region, and then classifies samples by using the naive Bayes classifier. We have demonstrated the usefulness of the proposed method for various representative benchmark datasets of multiclass cancer classification.

Key words : gene expression profiles, multiclass cancer classification, gene selection

· 본 연구는 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 지원사업의 연구결과로 수행되었음(IITA-2008-(C1090-0801-0046))

† 학생회원 : 연세대학교 컴퓨터과학과
hjinh@sclab.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sbcho@cs.yonsei.ac.kr

논문접수 : 2007년 12월 6일
심사완료 : 2008년 5월 12일

정보과학회논문지: 시스템 및 이론 제35권 제8호(2008.8)

1. 서 론

마이크로어레이 기술로 수집되는 유전발현 데이타는 질병과 관련된 다양한 정보를 가지고 있지만, 적은 샘플 수에 비해 대규모의 유전자로 구성된다. 따라서 연관이 있는 유전자를 선택하는 것은 정확한 질병 분류에 필수적이다[1]. 유전자 선택은 크게 필터와 래퍼 방법으로 나뉘는데, 필터 방법은 어떤 기준에 따라 유전자의 가치를 개별적으로 측정하는 반면, 래퍼 방법은 분류기와 연계된 우수한 유전자 집합을 찾는다. 래퍼 방법은 복잡한 유전자와 분류기 사이의 관계를 활용할 수 있으나 매우 많은 연산량이 필요하고 샘플이 적은 문제에는 적합하지 않다. 반면에 순위기반 유전자 선택 기법이라고도 불리는 필터 방법은 보통 적은 연산량에 비해 양호한 성능을 보인다.

기존의 순위기반 유전자 선택 기법은 이진 분류에 많이 사용되었으며, 사전에 설계된 이상표식 유전자와 유사한 유전자만을 선택한다. 특히 이상표식 유전자가 이진클래스로 구성되어 다중클래스 암 분류에 직접 적용하기에는 적합하지 않다[1]. 다중클래스 암 분류에서는 보통 다중클래스를 다수의 이진클래스 문제로 나눈 후 해결하는 방식을 취하기 때문에 유용한 유전자가 이진클래스 형태의 이상표식 유전자와 유사하지 않다는 이유로 제외될 수 있다.

본 논문에서는 다중클래스 암 분류에 적합한 유전자 선택 기법을 제안한다. 이상표식 유전자를 사용하지 않고 유전자로부터 직접 다른 클래스와의 구별력을 측정한다. 먼저 유전자를 세분화하여 각 영역별 샘플의 빈도를 계산하고 클래스구별력과 영역밀집도에 따라 유전자의 중요도를 측정한다. 중요도가 높은 유전자를 선택하고 나이브 베이즈(NB) 분류기를 이용하여 다중클래스 암 분류를 수행한다.

2. 배 경

2.1 유전발현 데이타를 이용한 다중클래스 암 분류

다중클래스 암 분류는 보통 세 개 이상의 암으로 구성된 데이타를 다루며, n개의 학습 데이타 $S = \{(x_1,y_1), \ldots (x_n,y_n)\}$ ($x_i \in X$: i번째 학습 샘플; $y_i \in Y=\{1,2,\cdots,k\}$: 대응하는 클래스 레이블)가 주어질 때, 각 샘플에 대응하는 레이블로 매핑하는 함수 $F: X \rightarrow Y$를 구해야 한다. 표 1은 다중클래스 암 분류의 대표적 연구들을 보여주는데, 보통 결정 트리나 k-최근접 이웃 등의 방법을 사용하여 직접 다중클래스를 다루거나 다수의 이진 분류 문제로 변환하여 SVM 등을 적용한다.

2.2 순위기반 유전자 선택

기존 순위기반 유전자 선택 기법은 사전에 정의된 이상표식 유전자와 유사한 유전자를 선택한다. 이상표식 유전자는 기본적으로 이진 분류에 적합하도록 정의되기 때문에 먼저 다중클래스 분류 문제를 One-Versus-Rest (OVR) 전략 등을 이용하여 복수의 이진 분류 문제로 분해해야 한다. 클래스 레이블 $y_i \in Y=\{1,2,\cdots,m\}$($m$: 클래스 수)에 대해서 n개의 학습 샘플이 주어지면 다음과 같이 길이가 n인 이상표식 유전자 집합 $K = \{K_1^+, K_1^-, K_2^+, K_2^-, \cdots, K_m^+, K_m^-\}$를 정의한다(단, j는 $1 \sim m$).

양의 이상표식 유전자 $K_j+ : (k_j^{(1)}, k_j^{(2)}, \cdots, k_j^{(n)})$

$$\begin{cases} k_j^{(i)} = 1, & \text{if } y_i = j, \\ k_j^{(i)} = 0, & \text{if } y_i \,!= j. \end{cases}$$

음의 이상표식 유전자 $K_j- : (k_j^{(1)}, k_j^{(2)}, \cdots, k_j^{(n)})$ (1)

$$\begin{cases} k_j^{(i)} = 0, & \text{if } y_i = j, \\ k_j^{(i)} = 1, & \text{if } y_i \,!= j. \end{cases}$$

$e_j^{(i)}$를 j번째 학습 샘플의 i번째 유전자의 발현 수준이라고 할 때, 학습 샘플에 대해 i번째 유전자 g_i는 다음과 같이 정의된다.

$$g_i = (e_1^{(i)}, e_2^{(i)}, \cdots, e_n^{(i)}), \quad (2)$$

그림 1과 같이 설계된 이상표식 유전자와 각 유전자의 유사도를 계산하여 유전자의 순위를 매긴다[8]. 이상표식 유전자와 유사한 양상을 보이는 유전자는 높은 순위를 가지며, 유사성이 떨어지는 유전자는 낮은 순위를

표 1 다중클래스 암 분류 관련연구

연구자	유전자 선택 기법	분류 기법	평가 데이타
Ramaswamy (2001) [2]	-	SVM	GCM
Lee (2003) [3]	BSS/WSS	Multicategory SVM	Leukemia data, SRBCT
Li (2004) [4]	IG, TR, GI, SM, MM, SV, t statistics	SVM, NB, kNN	Leukemia cancer data, GCM, SRBCT, NCI60
Statnikov (2005) [5]	BW, SN, one way ANOVA	SVM, kNN, NNs, Multicategory SVM	GCM, brain, leukemia, lung cancer data, SRBCT
Wang (2005) [1]	Relief F, IG, x^2-statistics	kNN, SVM, C4.5, NB	Leukemia cancer data
Yeung (2005) [6]	BSS/WSS, BMA	Logistic regression	Leukemia data, hereditary breast cancer data
Hong (2006) [7]	PC	SVM, NB	GCM

표 2 g_i 와 g_{ideal} 의 유사도를 측정하기 위한 방법

$$PC(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal} - \frac{\sum g_i \sum g_{ideal}}{N}}{\sqrt{\left(\sum g_i^2 - \frac{(\sum g_i)^2}{N}\right)\left(\sum g_{ideal}^2 - \frac{(\sum g_{ideal})^2}{N}\right)}}$$

$$SC(g_i, g_{ideal}) = 1 - \frac{6\sum (D_g - D_{ideal})^2}{N(N^2-1)}$$, (D_g, D_{ideal}은 g_i, g_{ideal}의 순위 행렬)

$$ED(g_i, g_{ideal}) = \sqrt{\sum (g_i - g_{ideal})^2}$$

$$CC(g_i, g_{ideal}) = \frac{\sum g_i g_{ideal}}{\sqrt{\sum g_i^2 \sum g_{ideal}^2}}$$

$$IG(g_i, c_j) = P(g_i \mid c_j) \log \frac{P(g_i \mid c_j)}{P(c_j) \cdot P(g_i)} + P(\bar{g}_i \mid c_j) \log \frac{P(\bar{g}_i \mid c_j)}{P(c_j) \cdot P(\bar{g}_i)}$$, (c_j : j 번째 클래스)

$$MI(g_i, c_j) = \log \frac{P(g_i, c_j)}{P(c_j) \cdot P(g_i)}$$

$$SN(g_i) = \frac{\mu_{c1}(g_i) - \mu_{c0}(g_i)}{\sigma_{c1}(g_i) + \sigma_{c0}(g_i)}$$

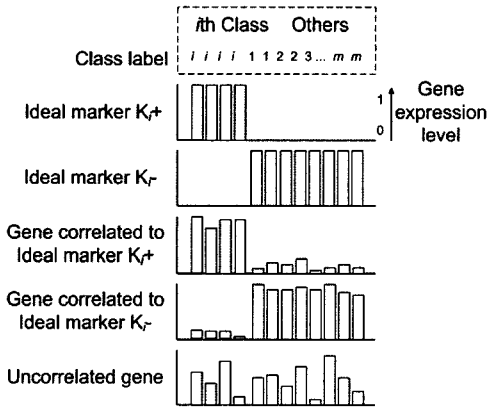

그림 1 기존 순위기반 유전자 선택 예

갖는다. 유사도는 표 2와 같이 피어슨 상관계수(PC), 스피어만 상관계수(SC), 유클리드 거리(ED), 코사인 계수(CC), 정보이득(IG), 상호정보(MI), 신호대잡음비(SN) 등의 다양한 방법을 사용하여 측정한다. 최근 유전자 선택에서는 피어슨 상관계수, 정보이득, 신호대잡음비 등이 많이 사용되지만, 유클리드 거리나 코사인 계수 등은 비교적 적은 연산으로 동작한다는 장점이 있다.

3. 제안하는 방법

3.1 다중클래스 암 분류를 위한 유전자 선택

본 논문에서 제안하는 순위기반 유전자 선택 기법은 그림 2와 같이 이산화, 빈도계산, 클래스구별력 및 영역밀집도 계산, 유전자 가치 및 순위 측정으로 구성된다.

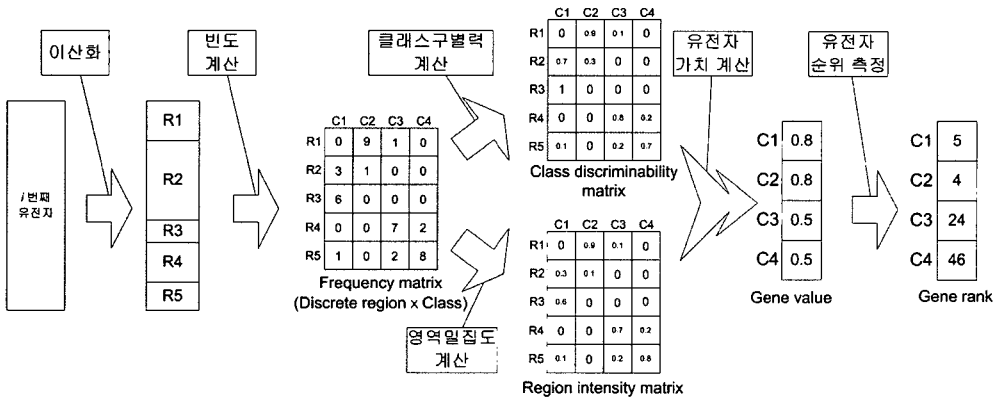

그림 2 제안하는 유전자 선택 방법

기존의 순위기반 유전자 선택 기법과 달리 제안하는 방법은 이상표식 유전자를 사용하지 않고 유전발현 수준을 직접 분석하여 유전자의 중요도를 측정한다. 먼저 수식 (3)과 같이 학습 데이타를 기반으로 t번째 유전자의 유전발현 수준을 d(=5)개의 영역으로 구분하고, 각 클래스에 대한 영역별 학습 샘플 발생 빈도를 측정한다. f_j^i는 클래스 레이블이 i이고 유전발현 수준이 j번째 영역에 속하는 학습 샘플의 수이다.

$$g_t = \begin{bmatrix} f_1^1 & f_1^2 & \cdots & f_1^m \\ f_2^1 & f_2^2 & \cdots & f_2^m \\ \vdots & \vdots & \ddots & \vdots \\ f_d^1 & f_d^2 & \cdots & f_d^m \end{bmatrix} \tag{3}$$

각 유전자의 학습 데이타에 대한 클래스/영역 빈도가 측정되면, 분류에 유용한 특성인 클래스구별력(Class discriminability; CD)과 영역밀집도(Region intensity; RI)를 계산한다. 이들 값은 수식 (4)와 같이 구분된 영역과 클래스에 대해 각각 구해진다. cd_j^i는 j번째 영역의 클래스 i에 대한 클래스구별력이고, ri_j^i는 j번째 영역의 클래스 i에 대한 영역밀집도이다.

$$CD(g_t) = \begin{bmatrix} cd_1^1 & cd_1^2 & \cdots & cd_1^m \\ cd_2^1 & cd_2^2 & \cdots & cd_2^m \\ \vdots & \vdots & \ddots & \vdots \\ cd_d^1 & cd_d^2 & \cdots & cd_d^m \end{bmatrix},$$

$$RI(g_t) = \begin{bmatrix} ri_1^1 & ri_1^2 & \cdots & ri_1^m \\ ri_2^1 & ri_2^2 & \cdots & ri_2^m \\ \vdots & \vdots & \ddots & \vdots \\ ri_d^1 & ri_d^2 & \cdots & ri_d^m \end{bmatrix}, \tag{4}$$

$$cd_d^c = \frac{f_d^c}{\sum_{k=1}^{m} f_d^k}, \quad ri_d^c = \frac{f_d^c}{\sum_{k=1}^{d} f_k^c}.$$

클래스구별력과 영역밀집도를 바탕으로 유전자가 클래스별로 얼마나 분류에 도움이 되는지를 식 (5)와 같이 계산한다. 특히 샘플이 고루 분포할 경우에 클래스구별력이 떨어지기 때문에 간단한 변환함수 $E(x)$를 정의하여 유전자의 중요도 계산에 반영하였다. $E(x)$는 샘플이 각 클래스에 고루 분포할 경우 가장 작은 값을 가지며, 한쪽으로 치우칠 경우 높은 값을 갖는다. 중요도 $c^{(t)}{}_i$는 t번째 유전자가 i번째 클래스를 얼마나 잘 분류하는지를 나타낸다. 유전자는 클래스 수만큼의 중요도를 가지며 클래스별 중요도에 따라 정렬된다. 분류에 선택되는 유전자는 클래스별 중요도 순위에 따라 그림 3과 같이 선택된다. 각 클래스를 잘 구분해주는 유전자를 선택하기 위해 클래스별로 다수의 유전자가 선택되도록 한다. 그림 3의 R은 최종적으로 선택된 유전자 리스트를 의미하고, o_i는 i번째 클래스의 중요도 c_i에 따라 정렬된 유전자 순위로, $o_i[j]$는 i번째 클래스를 분류하는데 j번째로 유용한 유전자를 의미한다.

```
입력: o_i[] (i번째 클래스에 대한 유전자 순위 리스트)
출력: R[] (선택 유전자 리스트)

gene = 0;
for (i=0; i<NUM_CLASS; i++) {
    cDiscriminability = 0.0;
    for (j=0; j<NUM_SELECTED_GENES/NUM_CLASS; j++) {
        if (cDiscriminability > THRESHOLD)
            break;
        t = R[gene++] = o_i[j];
        cDiscriminability += c^(t)_i;
        // c^(t)_i: goodness value of t^th gene for i^th class
}}
```

그림 3 유전자 순위 알고리즘

$$c_i^{(t)} = \sum_{j=1}^{d} (E(cd_j^i) \times ri_j^i),$$

$$E(x) = \begin{cases} 1 - m \times x, & \text{if } x < \dfrac{1}{m} \\ -\dfrac{1}{m-1} + \dfrac{m}{m-1} \times x, & \text{if } x \geq \dfrac{1}{m}. \end{cases} \tag{5}$$

3.2 NB 분류기를 이용한 다중 암 분류

NB 분류기는 샘플로부터 관측된 값과 미리 설계된 변수들의 사전 확률분포와, 특징과 클래스 사이의 조건부 확률분포를 바탕으로 각 클래스의 사후 확률을 계산한다[9]. 확률분포는 nT개의 학습 데이타로부터 계산되는데, 변수 A의 i번째 상태가 A_i이고 count(A_i)는 변수 A가 i번째 상태를 가지는 경우의 빈도를 나타낼 때, 사전 확률 P(A_i)는 식 (6)과 같이 계산된다.

$$P(A_i) = \frac{count(A_i)}{n_T} \tag{6}$$

만약 변수 A가 B를 부모 노드로 가지면, 조건부 확률 P($A_i|B_j$)는 식 (7)과 같이 계산된다.

$$P(A_i \mid B_j) = \frac{count(A_i, B_j)}{count(B_j)} \tag{7}$$

Bayes 이론에 따라 n개의 특징값이 증거로 주어질 때 각 클래스의 사후확률은 식 (8)과 같이 계산된다.

$$P(C \mid F_1, \ldots, F_n) = \frac{P(C)P(F_1, \ldots, F_n \mid C)}{P(F_1, \ldots, F_n)} \tag{8}$$

식 (8)의 분모는 클래스의 사후확률 계산에서 항상 동일하기 때문에 분자만을 고려한다면 클래스의 사후확

률은 특징들 사이의 독립성 가정에 따라 다음과 같이 표현되며, 가장 높은 값을 가지는 클래스로 샘플이 분류된다.

$$\begin{aligned} & P(C)P(F_1,...,F_n \mid C) \\ & = P(C)P(F_1 \mid C)P(F_2 \mid C)...P(F_n \mid C) \\ & = P(C)\prod_{i=1}^{n} P(F_i \mid C) \end{aligned} \quad (9)$$

4. 실험 및 결과

4.1 실험 환경

제안하는 방법을 평가하기 위해서 표 3에서와 같이 기존 논문에서 학습 데이타와 테스트 데이타를 구분한 대표적 다중클래스 유전발현 데이타인 GCM[2], Leukemia[10], NCI60[11]와 SRBCT[12] 데이타와 표 4에서와 같이 GEMS(http://www.gems-system.org/)에서의 9가지 다중클래스 유전발현 데이타를 사용하였다. 이들은 적은 수의 샘플에 비해 매우 많은 수의 유전자로 구성되어 있다. 본 논문에서는 클래스 수의 10배에 해당하는 개수의 유전자를 선택하였으며, 유전발현 수준은 모두 0에서 1사이로 정규화하여 실험을 수행하였다. 표 3의 데이타는 학습 데이타와 테스트 데이타를 초기 환경에 맞추어 실험하였고, 표 4의 데이타에 대해서는 5-집단 교차검증(5-fold cross validation)을 수행하여 결과를 획득하였다.

표 3 평가 데이타 설명

데이타	GCM	Leukemia	NCI	SRBCT
유전자 수	16,063	12,582	5,244	2,308
클래스 수	14	3	8	4
학습 데이타 수	144	57	43	63
테스트 데이타 수	54	15	18	20
선택 유전자 수	140	30	80	40

표 4 GEMS의 9가지 데이타 설명

데이타	유전자 수	샘플 수	클래스 수	선택 유전자 수
Leukemia 2	11,225	72	3	30
Leukemia 1	5,327	72	3	30
SRBCT	2,308	83	4	40
Lung	12,600	203	5	50
Brain 2	10,367	50	4	40
Brain 1	5,920	90	5	50
9 tumors	5,726	60	9	90
11 tumors	12,533	174	11	110
14 tumors	15,009	308	26	260

4.2 결과분석

각 데이타에 대해 모든 특징 선택 방법이 아주 적은 수의 특징을 사용하여 대부분 학습 데이타를 거의 완벽하게 분류하는 NB 분류기를 획득하였다. 표 5는 테스트 데이타에 대한 분류율을 보여주는데, 제안하는 방법이 대체로 다른 특징 선택 기법에 비해 높은 분류율을 보여주었으며, 평균 79.3%의 분류율로 기존의 방법보다 높은 분류성능을 획득하였다.

전반적으로 기존 방법에서 Leukemia 데이타에 대해 선택된 유전자의 발현 수준은 이상표식 유전자와 유사한 유전자를 뽑아 비슷한 양상을 보였다. 표 6은 각 유전자 선택 기준별 중복된 유전자의 수를 보여준다. PC, ED, CC, SN 등은 비슷한 유전자가 많이 뽑혔으며, SC와 제안하는 방법은 전반적으로 다른 특징 선택 방법과 다른 양상을 보였다. 성능이 저조한 IG와 MI의 경우는 다른 방법들이 뽑은 유전자를 거의 선택하지 않았다. 유전자(#2769)는 IG를 제외한 모든 방법에서 선택되어 Leukemia 암 분류에 매우 유용하였고, 유전자(#11296), 유전자(#1118) 등은 5가지 방법에서 선택되었다.

표 6 동일 유전자의 선택 빈도

	SC	ED	CC	IG	MI	SN	PP
PC	5	10	13	0	1	18	4
SC		2	3	0	2	4	7
ED			10	0	0	9	3
CC				0	0	0	4
IG					3	1	0
MI						0	0
SN							4

그림 4는 GEMS의 9가지 암 데이타에 대한 분류 결과로, 거의 모든 경우에서 제안하는 방법이 기존 유전자 선택 기법에 비해 높은 분류 성능을 얻었으며, 평균 4~10% 이상의 성능 향상을 확인하였다.

5. 결 론

다중클래스 분류는 패턴인식에서 매우 도전적인 과제

표 5 테스트 데이타에 대한 분류율

데이타 (%)	PC	SC	ED	CC	IG	MI	SN	PP
GCM	48	46	33	48	35	44	52	50
Leukemia	100	93	100	100	93	80	100	100
NCI	50	56	39	72	56	50	61	72
SRBCT	100	100	75	95	65	85	95	95
Avg	74.5	73.8	61.8	78.8	62.3	64.9	77	79.3

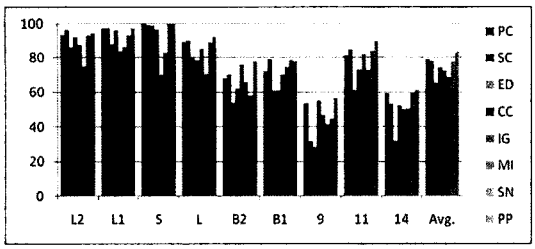

그림 4 GEMS 데이타에 대한 분류 성능

로 기존의 순위기반 특징 선택 방법을 직접 적용하기에는 한계가 있다. 본 논문에서는 이상표식 유전자를 사용하지 않고 유전자의 발현 수준을 직접 분석하는 방법을 제안하였고, 생물정보학의 대표적인 다중클래스 암 분류 데이타를 대상으로 다중클래스 암 분류에 적용하여 병렬적인 OVR 방식으로 기존의 특징 선택을 적용한 방법보다 높은 성능을 획득하였다. 향후에는 보다 다양한 다중클래스 데이타에 적용할 것이다.

참 고 문 헌

[1] Y. Wang, F. Makedon, J. Ford and J. Pearlman, "HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, Vol. 21, No.8, pp. 1530-1537, 2005.

[2] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander and T. Golub, "Multiclass cancer diagnosis using tumor gene expression signatures," *Proc. National Academy of Science*, Vol.98, No.26, pp. 15149-15154, 2001.

[3] Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, Vol.19, No.9, pp. 1132-1139, 2003.

[4] T. Li, C. Zhang and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, Vol.20, No.15, pp. 2429-2437, 2004.

[5] A. Statnikov, C. Aliferis, L. Tsamardinos, D. Hardin and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, Vol.21, No.5, pp. 631-643, 2005.

[6] K.-Y. Yeung, R. Bumgarner and A. Raftery, "Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, Vol.21, No.10, pp. 2394-2402, 2005.

[7] J.-H. Hong, and S.-B. Cho, "Multi-class cancer classification with OVR-support vector machines selected by naive Bayes classifier," *Lecture Notes in Computer Sciences*, Vol.4234, pp. 155-164, 2006.

[8] S.-B. Cho and J.-W. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proceedings of the IEEE*, Vol.90, No.11, pp. 1744-1753, 2002.

[9] J. Liu, B. Li and T. Dillon, "An improved naïve Bayesian classifier technique coupled with a novel input solution method," *IEEE Trans. Systems, Man, and Cybernetics-Part C: Applications and Reviews*, Vol.31, No.2, pp. 249-256, 2001.

[10] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, M. den Boer, M. Minden, S. Sallan, E. Lander, T. Golub, and S. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genetics*, Vol.30, No.1, pp. 41-47, 2002.

[11] D. Ross, U. Scherf, M. Eisen, C. Perou, P. Spellman, V. Iyer, S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein, and P. Brown, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics*, Vol.24, No.3, pp. 227-234, 2000.

[12] J. Khan, J. Wei, M. Ringnér, L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, and P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, Vol.7, No.6, pp. 673-679, 2001.

홍 진 혁

2002년 연세대학교 기계전자공학부 정보산업전공 졸업. 2002년~2004년 연세대학교 컴퓨터과학과 석사. 2004년~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 지능형 에이전트, 패턴인식, 바이오인포메틱스

조 성 배

1988년 연세대학교 전산과학과(학사). 1990년 한국과학기술원 전산학과(석사). 1993년 한국과학기술원 전산학과(박사). 1993년~1995년 일본 ATR 인간정보통신연구소 객원 연구원. 1998년 호주 Univ. of New South Wales 초청연구원. 1995년~현재 연세대학교 컴퓨터과학과 정교수. 관심분야는 신경망, 패턴인식, 지능정보처리