韓國컴퓨터情報學會 論文誌
第15卷 第12號, 2010. 12.

2010-15-12-4-9

조건(암, 정상)에 따라 특이적 관계를 나타내는 유전자 쌍으로 구성된 유전자 모듈을 이용한 독립샘플의 클래스예측

정 현 이*, 윤 영 미**

Class prediction of an independent sample using a set of gene modules consisting of gene-pairs which were condition(Tumor, Normal) specific.

Hyeoniee Jeong*, Youngmi Yoon**

요 약

대용량(High-throughput) 형태로 얻어진 cDNA 마이크로어레이 데이터에 다양한 데이터 마이닝 기법을 적용하면 서로 다른 조직에서 추출한 유전자의 발현정도를 비교할 수 있고 정상세포와 암세포에서 발현량의 차이를 보이는 DEG(Differently Expression Gene) 유전자를 추출할 수 있다. 이들을 이용하여 병을 진단할 수 있을 뿐만 아니라, 암의 진행 단계(Cancer Stage)에 따른 치료 방법을 결정할 수 있다. 마이크로어레이를 기반으로 한 대부분의 암 분류자는 기계학습 기법을 이용하여 암 관련 유전자를 추출하여, 이들 유전자를 총체적으로 이용하여 독립샘플의 클래스(암, 정상)를 판정한다. 하지만 유전자의 발현량의 차이뿐만 아니라 유전자와 유전자의 상관관계의 변화가 질병 진단에 활용될 수 있다. 대부분의 질병은 단독 유전자의 변이에 의한 것이 아니라 유전자의 모듈로 이루어진 유전자조절네트워크의 변이에 의한 것이기 때문이다. 본 논문에서는 조건에 따라 특이적 관계를 나타내는 유전자 쌍을 식별하여, 이들 유전자 쌍을 이용한 유전자 분류 모듈을 생성한다. 분류 모듈을 이용한 암 분류 방법이 기존의 암 분류 방법보다 높은 정확도로 암과 정상 샘플을 분류함을 보여주고 있다. 분류 모듈을 구성하는 유전자의 수가 상대적으로 적으므로 임상키트로의 개발도 고려할 수 있다. 향후 분류 모듈에 속하는 유전자의 기능적 검증을, GO(Gene Ontology)를 활용함 으로서, 밝혀지지 않은 새로운 암 관련 유전자를 식별하고, 분류 모듈을 확대하여 암 특이적 유전자조절네트워크 구성에 활용할 계획이다.

Abstract

Using a variety of data-mining methods on high-throughput cDNA microarray data, the level of gene expression in two different tissues can be compared, and DEG(Differentially Expressed Gene) genes in between normal cell and tumor cell can be detected. Diagnosis can be made with these

• 제1저자 : 정현이 교신저자 : 윤영미
• 투고일 : 2010. 06. 29, 심사일 : 2010. 07. 19, 게재확정일 : 2010. 10. 15.
* 가천의과학대학교 IT학과 학생 ** 가천의과학대학교 정보공학부 교수
※ 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2010-0008639).

genes, and also treatment strategy can be determined according to the cancer stages. Existing cancer classification methods using machine learning select the marker genes which are differential expressed in normal and tumor samples, and build a classifier using those marker genes. However, in addition to the differences in gene expression levels, the difference in gene-gene correlations between two conditions could be a good marker in disease diagnosis. In this study, we identify gene pairs with a big correlation difference in two sets of samples, build gene classification modules using these gene pairs. This cancer classification method using gene modules achieves higher accuracy than current methods. The implementing clinical kit can be considered since the number of genes in classification module is small. For future study, Authors plan to identify novel cancer-related genes with functionality analysis on the genes in a classification module through GO(Gene Ontology) enrichment validation, and to extend the classification module into gene regulatory networks.

▸ Keyword : 데이터마이닝(datamining), 분류분석(classification), 지식기반 데이터마이닝(knowledge-based datamining), 마이크로어레이데이터분류분석(microarray data classification)

Ⅰ. 서 론

신체에서 정상세포의 기능은 매우 복잡하고도 치밀한 상호작용에 의해 유지되고 있다. 그러나 만약 세포 분열을 촉진시키는 암 유전자나 세포 분열을 억제 시키는 종양 억제 유전자에 이상이 생겨 평형이 깨지게 되면, 비정상적인 세포분열로 인해 암 발생이 일어날 수 있다. 하나의 암 유전자 또는 종양 억제 유전자의 변화가 단독으로 암을 일으키는 것은 아니며, 대부분의 암 발생은 유전자와 유전자의 상호작용의 변화에 의한 것이라 할 수 있다.

대용량(High-throughput) 형태로 얻어진 cDNA 마이크로어레이 데이터에 다양한 데이터 마이닝 기법을 적용하면 서로 다른 조직에서 추출한 유전자의 발현정도를 비교할 수 있고, 정상세포와 암세포에서 발현량의 차이를 보이는 DEG(Differently Expression Gene) 유전자를 추출할 수 있다. 이들을 이용하여 병을 진단할 수 있을 뿐만 아니라, 암의 진행 단계(Cancer Stage)에 따른 치료 방법을 결정할 수 있다 [1].

마이크로어레이를 기반으로 한 대부분의 암 분류자는 기계학습 기법을 이용하여 특정 클래스(암, 정상)에서 유의한 발현량의 차이를 보이는 유전자를 개별적으로 추출하거나, 이들 유전자를 총체적으로 이용하여 독립샘플의 암 여부를 판정한다 [2]. 하지만 대부분의 질병은 단독 유전자의 변이에 의한 것이 아니라 유전자의 모듈로 이루어진 유전자조절네트워크의 변이에 의한 것이므로 조건에 따른 유전자의 발현량의 차이뿐만 아니라 유전자와 유전자의 상관관계의 변화가 질병 진단의 주요한 지표가 될 수 있다[3]. 서포트 벡터머신(support vector machine) 과 부스팅 기법을 활용한 분류 방식은 유전자전체를 고차원 (Hyperplane)에 사상시키는 방법이기 때문에 유전자들 사이에 존재하는 관계를 파악하기 쉽지 않고, 그들이 복합작용에 의해 일어나는 역할 역시 쉽게 분석하기 힘들다[4]. 클래스 분류에 참여하는 유전자가 명확히 식별이 되지 않을 뿐만 아니라, 조건에 따른 유전자와 유전자의 상관관계 변화를 고려한 방법이 아니다.

본 논문에서는 암, 정상 조건에 따라 특이적 관계를 보이는 유전자 쌍을 식별 하여, 이 들 유전자 쌍을 연결하여 유전자 모듈을 만든 후, 이 모듈을 분류자로 이용하여 독립샘플의 클래스를 판별하는 것을 목적으로 한다. 선택된 유전자 분류 모듈은 그 모듈 안에 참여하는 유전자들을 식별할 수 있으므로 생물학적 해석이 용이하며 [5], 그 모듈을 구성하는 유전자들의 발현패턴이 암샘플 집단과 정상샘플 집단에서 명확하게 구분된다. 각 샘플 집단에서 특이적인 발현 패턴을 보이는 모듈은 그 모듈을 구성하고 있는 유전자들의 기능적 연관성을 표시한다[6].

마이크로어레이 실험은, 동일한 실험실에서 동일한 유전자 세트로 수행되어도 약간의 실험조건의 변이를 포함한 시스템적인 오류가 발생할 수 있으므로, 마이크로어레이 데이터의 분석알고리즘은 데이터에 내재된 노이즈를 고려 할 수 있어야 한다 [7]. 본 논문에서는 마이크로어레이에 나타나는 유전자 발현값의 오차범위와 데이터에 내재된 노이즈를 고려하기 위하여 두 개의 파라미터를 사용한다.

Ⅱ. 관련 연구

1. 마이크로어레이 (Microarray)

주어진 조건에서 유전자들의 발현정도를 정량적으로 표시

해주는 마이크로어레이는 세포의 상태를 표현하는 방법으로 가장 많이 사용되고 있다. DNA나 RNA, 단백질같이 전하를 띄고 있는 생체분자들을 전기적인 특성과 분자량 등을 이용하여 분리하고 필름에 감광하여 생체분자들의 양을 측정하는 방식을 확장한 DNA 마이크로어레이는 DNA가 상보적으로 결합하는 성질을 이용하는 것이다. 핀을 이용한 마이크로어레이 스팟팅이나 잉크젯의 원리를 응용한 기술로 만들어지고 있는 cDNA칩은 비교적 비용이 적게 들고 제작방식이 쉽다[8].

DNA 마이크로어레이는 한 번의 실험으로 많은 수의 유전자의 발현값을 동시에 측정할 수 있으므로 분자 생물학 실험에 있어 중요한 도구로 사용된다[7].

	tumor				normal			
	s_1	s_2	..	s_n	s_{n+1}	s_{n+2}	..	s_{n+m}
g_a	$g_{a,1}$	$g_{a,2}$	..	$g_{a,n}$	$g_{a,n+1}$	$g_{a,n+2}$	..	$g_{a,n+m}$
g_b	$g_{b,1}$	$g_{b,2}$	..	$g_{b,n}$	$g_{b,n+1}$	$g_{b,n+2}$	..	$g_{b,n+m}$
g_c	$g_{c,1}$	$g_{c,2}$	..	$g_{c,n}$	$g_{c,n+1}$	$g_{c,n+2}$	..	$g_{c,n+m}$
g_d	$g_{d,1}$	$g_{d,2}$	..	$g_{d,n}$	$g_{d,n+1}$	$g_{d,n+2}$	..	$g_{d,n+m}$
...	...	...	..	...	...	...	..	...
g_z	$g_{z,1}$	$g_{z,2}$	..	$g_{z,n}$	$g_{z,n+1}$	$g_{z,n+2}$	..	$g_{z,n+m}$

그림 1. 마이크로어레이 데이터
Fig. 1. Microarray Data

마이크로어레이 데이터는 유전자의 집합이 정상세포와 암세포와 같이 특정한 조건 하에서 얼마나 발현되는지를 수치화한 이차원 행렬 데이터이다. 유전자 발현 데이터를 담고 있는 마이크로어레이에서 세로축은 유전자에 대한 정보를 담고 있다. 그림 1에서와 같이 g_a, g_b로 표시한다. 가로축은 마이크로어레이가 만들어지게 된 시료와 시료가 얻어진 실험 조건에 대한 샘플정보를 가진다. 각각의 셀은 암 샘플 s_1의 유전자 g_a의 발현값을 나타낸다.

이 마이크로어레이는 유전자의 발현 정도를 측정하여, 유전자 기능 예측, 돌연변이나 다형성(polymorphism) 진단 및 질병관련 유전자 발굴 등의 연구 목적에 사용되며 의료진단, 신약개발 탐색 및 의약용 유전체 발굴 등에 활용 가능하다.

2. 피어슨 상관계수 (Pearson's Correlation coefficient)

상관분석은 두 변수(샘플/유전자) 사이에 선형적 관계의 존재여부를 찾는 것이다. 두 변수는 서로 독립적인 관계로부터 서로 상관된 관계일 수 있으며 두 변수간의 관계 강도를 상관관계라고 한다. 상관분석에서 구해지는 상관계수(correlation coefficient)는 두 변수 간의 연관성의 정도를 나타낸다. 두 변수가 직선적인 관계에 있다면 상관계수는 -1이나 1에 가까워지는데, -1에 가까워지는 경우는 음의 상관관계가 높은 경우이고 1에 가까워지는 경우는 양의 상관관계가 높은 경우이다. 두 변수의 상관관계를 보면 한 변수의 변화에 의해 다른 한 변수가 변하는 정도를 알아 볼 수 있다[9].

마이크로어레이에 피어슨 상관계수를 적용해서 샘플과 샘플의 상관관계와 유전자와 유전자의 상관관계를 알아낼 수 있다. 본 연구에서는 분류에 유용한 유전자 쌍을 찾기 위해서는 동일한 조건(암, 정상) 안에서의 유전자 벡터와 유전자 벡터 사이 상관계수를 사용한다. 또한 특이 유전자 쌍으로 생성된 유전자 모듈을 이용하여 클래스를 모르는 샘플 데이터를 판별할 때에도 상관계수가 쓰인다. 유전자 모듈을 구성하는 유전자들로 이루어진 샘플벡터 끼리의 상관계수를 이용하여 독립샘플의 클래스(암, 정상)를 판별한다.

피어슨 상관계수는 두 변수 혹은 벡터사이의 선형관계를 명확하게 설명한다. 두 벡터의 피어슨 상관계수는 다음과 같이 표현 된다[10].

$$PCC(X,Y)=\frac{\sum XY-\frac{\sum X\sum Y}{N}}{\sqrt{(\sum X^2-\frac{(\sum X)^2}{N})(\sum Y^2-\frac{(\sum Y)^2}{N})}},$$

X : 테스트 샘플에서 유전자 쌍을 구성하는 유전자 벡터,
Y : 유전자 분류 모듈을 이루는 샘플의 유전자 벡터.

특이적 규칙을 만족하는 유전자 쌍을 이용하여 유전자 모듈을 생성하기 위해 상관계수를 이용하는 경우, X와 Y는 유전자 쌍을 구성하는 유전자 벡터를 나타낸다. 유전자 벡터 X, Y의 상관계수는 PCC(X, Y)로 나타낸다. 클래스(암, 정상)를 모르는 테스트 샘플을 판별할 때 상관계수를 이용하는 경우, X는 테스트 샘플 벡터를 의미하고 Y는 클래스(암, 정상)를 판별하기 위해 만든 유전자 분류 모듈에 참여한 유전자로 이루어진 샘플 벡터를 의미한다.

3. 분류방식

랜덤 포레스트(Random Forest)는 처음 소개된 2001년 이후 다양한 분야에 적용되어 왔다[11]. 랜덤 포레스트는 무작위로 추출한 사례의 집합들을 이용하여 많은 수의 의사결정나무를 생성하고, 판별 클래스들을 가중 투표하여 최종 클래스를 결정하는 분류 기법이다[12]. 랜덤 포레스트는 적은 수의 임상 사례만으로도 일정 수준 이상의 정확성을 가지는 분

류기를 생성할 수 있다[13].

서포트 벡터 머신(Support Vector Machine, SVM)은 선형 분류(linear classifier)의 일종으로 분류기법으로 많이 사용되고 있는 방법이고, 특히 방대하고 다양한 데이터를 분류하는데 많이 사용되고 있다[14]. 마이크로어레이에서 두 클래스의 벡터들이 주어졌을 경우 커널함수를 이용하여 두 클래스의 샘플들의 사이 거리가 가장 구분이 크게 되는 분할평면(hyperplane)을 찾고 그 분할 평면을 기준으로 독립샘플을 분류한다[15]. 일반적으로 커널 함수의 특징에 따라 인식률의 차이가 있으므로 문제에 맞게 적절하게 선택해야 한다[16].

단순 베이지안 분류기는 사후확률의 베이즈 이론에 기반하고 있다. 주어진 데이터가 특정 클래스에 속할 확률과 같은 클래스의 소속확률을 예측한다. 베이지안 분류는 베이즈 이론에 기반하며 대규모 데이터에 적용되어도 높은 정확성과 속도를 보여준다[17].

의사결정나무(Decision Tree)는 클래스 레이블이 있는 트레이닝 데이터로부터 의사결정나무를 학습하는 것이다. 클래스가 알려져 있지 않은 샘플이 주어지면, 샘플의 속성 값들이 의사결정나무에 의해 검사된다[18]. 직관적 일 뿐만 아니라 학습과 분류의 단계가 간단하고 빠르다[19]. 일반적으로 정확도가 높지만, 데이터의 잡음이나 이상치를 처리하기 위한 가지치기(pruning) 과정으로 분류정확도를 더욱 향상시킬 수 있다.

k-근접이웃분류기는 대용량의 트레이닝 데이터가 주어졌을 때 사용되는, 장시간의 컴퓨터 계산시간을 요구하는 방법이다. 미지의 샘플이 주어졌을 때, k-근접이웃분류기는 미지의 샘플과 가장 가까운 k개의 트레이닝 샘플을 패턴공간에서 탐색한다[20]. 이러한 k개의 트레이닝 샘플들은 미지의 샘플에 대한 k개의 '근접이웃' 이 된다. 근접이웃 분류기에서 각 속성에 대한 근접도는 보통 같은 가중치를 부여하는 거리-기반 비교법인 유클리드거리 함수를 사용한다[21]. 따라서 연관성이 있는 속성들이 주어지면 좋지 않은 정확도를 보이게 된다.

III. 본 론

1. 시스템 개요

암세포와 정상세포에서 각각 얻은 유전자를 토대로 제작한 마이크로어레이를 이용하여 분류자를 찾아 테스트 데이터의 클래스를 판별하기 위해 총 세 단계의 과정이 진행된다. 본 알고리즘은 마이크로어레이 데이터의 암, 정상 샘플 각각의 경우에 대하여, 의미 있는 유전자 쌍을 식별하여, 이를 이용하여 유전자 분류 모듈를 구성한 후, 다수개의 분류모듈로 이루어진 분류자를 만들어서 새로운 데이터의 암, 정상 여부를 판별하는 방식으로 진행된다.

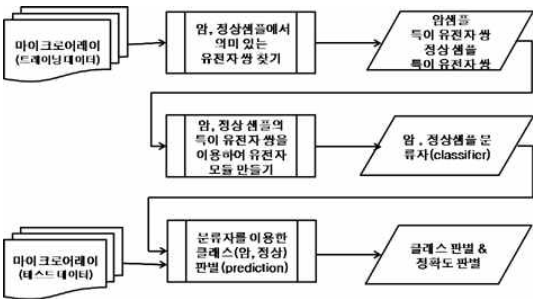

그림 2. 알고리즘 흐름도
Fig. 2. Algorithm Flow

가장 먼저 정확도가 높은 유전자 분류 모듈을 형성하기 위해 트레이닝 데이터에서 의미 있는 유전자 쌍을 찾아야 한다. 의미 있는 유전자 쌍이란 각기 암 샘플에서만 특이적 관계를 나타내는 유전자 쌍과, 정상 샘플에서만 특이적 관계를 나타내는 유전자 쌍을 말한다. 이들 유전자 쌍은 알고리즘에서 설정한 파라미터 값과 범위에서 암 샘플 혹은 정상 샘플에서 특정한 패턴을 보인다.

첫 단계에서 찾은 의미 있는 유전자 쌍을 이용하여 분류자를 구성하는 유전자 분류 모듈을 만들기 위해 유전자 쌍들을 연결한다. 본 논문에서는, 단지 개별 표지 유전자(marker gene)로 샘플의 클래스(암, 정상)를 판별하던 기존의 방식과 달리 유전자 모듈을 만들어 판별한다. 유전자 쌍과 다른 유전자 쌍에 존재하는 중복된 유전자를 이어서 유전자 모듈을 형성해 나간다. 이렇게 만든 유전자 모듈은 상관계수가 높은 순서대로 우선순위 큐 알고리즘에 의해 정렬되며 상관계수가 높은 상위 유전자 모듈부터 분류자를 이루는 유전자 분류 모듈로서의 채택이 가능하다. 유전자 모듈은 암 샘플을 판별할 수 있는 유전자 모듈과 정상 샘플을 판별할 수 있는 유전자 모듈로 나누어 생성되고, 분류자 역시 암 분류자와 정상 분류자가 분리돼서 사용된다. 암 분류자와 정상 분류자를 이루는 유전자 분류 모듈의 개수는 동일하게 유지한다.

마지막으로, 채택된 k개의 유전자 분류 모듈로 이루어진 암 분류자와 정상 분류자를 이용하여 클래스(암, 정상)를 모르는 샘플 데이터의 클래스를 판별한다. 암 샘플에 대하여 암 분류자에 포함되는 유전자만을 포함한 각 샘플 벡터와 동일한 유전자의 테스트 샘플벡터와 상관계수의 평균을 구한다. 정상 샘플에서도 값은 방식으로 계산하여 큰 상관계수평균을 갖는

클래스로 판정한다. 동시에 클래스(암, 정상)를 정확하게 판별한 비율을 정확도로 나타내어 알고리즘의 성능을 보인다.

2. 알고리즘

2.1 유전자 쌍 찾기

트레이닝 데이터에서 암과 정상 샘플을 유의미하게 구분하는 유전자 쌍을, 각 클래스(암,정상) 별로 구분하여 추출한다. 유의미한 유전자 쌍을 찾을 때 두 유전자 사이의 값이 의미있게 차이가 나도록 하기 위해 두 개의 파라미터를 사용 한다.

이 알고리즘에서 사용되는 파라미터 중의 하나인, n(noise)은 마이크로어레이 데이터 값이 내포할 수 있는 노이즈를 보정하기 위하여 사용된다. n을 파라미터로 설정하면 유전자 발현값(Value of Gene)을 나타내는 $V(g)$의 범위는 (a)와 같이 변환되고 $V(g)*(1-n)$은 유전자 발현값의 최솟값, $V(g)*(1+n)$은 최댓값이 된다. 이 범위는 마이크로어레이 데이터의 노이즈를 고려한, 특정한 클래스(암, 정상)를 판별하는데 도움을 줄 수 있는 유전자 발현값의 범위라고 할 수 있다. 낮은 n 은 두 유전자 발현값의 엄격한 차이를 의미하고, 반대로 높은 n 은 여유 있는 차이를 의미한다.

$$(a)\ V(g)*(1-n) < V(g) < V(g)*(1+n)$$
$$(b)\ V(g_a)*(1-n) > V(g_b)*(1+n),$$

$V(g)$: 유전자 'g' 의 발현값 (Expression Value of Gene)

n : 마이크로어레이에서의 유전자 발현값이 내포할 수 있는 노이즈를 보정하기 위한 파라미터

예를 들어 두 유전자 g_a와 g_b를 선택하여 비교했을 때, (b)처럼 g_a의 최솟값이 g_b의 최댓값보다 클 경우 유의미한 유전자 쌍으로 분류할 수 있다. 또한 나머지 파라미터, s(significance)는 유전자 발현값이 확실하게 차이나도록 하기 위한 값이다. 유전자 g_a의 발현값이 유전자 g_b의 발현값보다 큰 경우에 (c)와 같이 설정해주면서 $V(g_a)$와 $V(g_b)$가 확실하게 차이가 나도록 할 수 있다.

$$(c)\ V(g_a)*(1-n) > V(g_b)*(1+n)+s,$$

s : 유전자 g_a 와 유전자 g_b 의 발현값의 확연한 차이를 위한 파라미터

앞서 설명한 두 변수 n과 s를 이용하여 트레이닝 데이터의 모든 유전자를 대상으로 임의의 두 유전자 쌍의 샘플 당 유전자 발현값을 비교한다. 암 샘플은 암 샘플끼리, 정상샘플은 정상샘플끼리 비교하여 위 조건에 만족하는 유전자 쌍을 각각 분류한다.

파라미터 값에 의해 분류된 유전자 쌍이 어느 샘플을 판별하는데 유용하게 쓰이는지 구분한다. 암 샘플에서 유전자 쌍 (g_a-g_b)의 유전자 발현 값을 비교하여 n과 s에 의한 특성을 보이는 샘플의 비율, $tscore$을 계산한다. 이 비율은 암 샘플 내에서의 비율이다.

$$tscore = \frac{\text{암 샘플에서 } (c) \text{를 만족하는 유전자 쌍의 개수}}{\text{암 샘플에서 만들 수 있는 전체 유전자 쌍의 개수}}$$

$$nscore = \frac{\text{정상 샘플에서 } (c) \text{를 만족하는 유전자 쌍의 개수}}{\text{정상 샘플에서 만들 수 있는 전체 유전자 쌍의 개수}}$$

또한 정상 샘플에서도 그 특성에 만족하는 샘플의 비율, $nscore$을 계산한다. 이 비율로써 선택된 유전자 쌍이 암 샘플과 정상 샘플에서 어떤 차이를 보이고 있는지 비교할 수 있다.

표 1. tscore와 nscore에 의해 분류되는 유전자 쌍
Table 1. Gene pair by scoring of tscore and nscore

$tscore \geq 0.98$	$nscore \leq 0.40$	암 샘플 특이 유전자 쌍
$nscore \geq 0.98$	$tscore \leq 0.40$	정상 샘플 특이 유전자 쌍

기존 연구방법들 중에는 두 유전자의 크기관계식을 만족하는 유전자 쌍의 비율이 상대적으로 높은 클래스(암, 정상)를 선택하여 클래스 분류에 유용한 표지 유전자 쌍(marker gene pairs)을 찾는 방법이 있다[5, 22]. 본 연구에서는 크기 관계식을 만족하는 비율을 일정 비율이상으로 제한하였다. 암 샘플에서 두 유전자 쌍으로 계산한 $tscore$가 98% 이상이고 $nscore$가 40% 이하라면 이 유전자 쌍은 암 샘플을 구별하는 데 유용한 유전자 쌍이라고 구분한다. 반대로 정상 샘플에서 두 유전자 쌍으로 계산한 $nscore$가 98% 이상이고 $tscore$가 40% 이하라면 이 유전자 쌍은 정상 샘플을 구별하는 데 유용한 유전자 쌍이라고 구분한다.

암 샘플을 구별하는데 유용한 유전자 쌍은 암 샘플 분류모듈을 구성하는데 사용되고 정상 샘플을 구별하는데 유용한 유전자 쌍은 정상 샘플 분류모듈을 구성하는데 사용된다.

2.2 유전자 모듈로 이루어진 분류자의 생성

암 샘플과 정상 샘플을 구분하여 특이 유전자 쌍을 식별한 후 찾은 순서대로 유전자 쌍과 그에 맞는 유전자 발현값을 나열한다. 순서대로 정리된 유전자 쌍은 암 샘플과 정상 샘플을 구분하여 유전자 모듈을 만들 수 있다. 유전자 쌍의 개수를 2개 이상으로, 유전자 모듈을 이루는 유전자의 개수는 3개 이

상으로 유전자 모듈을 생성한다.

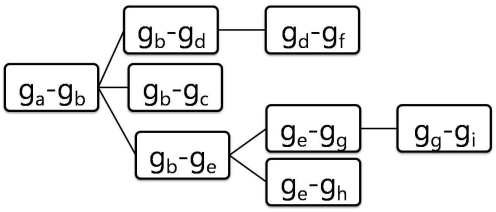

그림 3. 유전자 모듈 구성
Fig. 3. Constructing Gene Modules

그림 3처럼 ga-gb를 시작으로 유전자 모듈을 만들 수 있는 유전자 쌍들을 나열할 수 있다. 제시된 그림에서 만들 수 있는 유전자 모듈의 개수는 총 4개이다. 예를 들어 유전자 모듈의 형태는 [ga-gb-gd-gf]로, 최대로 연결될 수 있는 모든 형태의 유전자 모듈을 만들도록 한다.

연결된 유전자 쌍을 이어서 유전자 모듈을 생성 할 때, 암 클래스를 유의미하게 구분해주는 유전자 모듈과, 정상클래스를 유의미하게 구분해주는 유전자 모듈을 분리하여 생성 한다. 최소 유전자 개수를 만족하는 유전자 모듈을 식별한 후에, 유전자 모듈을 이루는 유전자끼리의 상관계수를 구한다. 마이크로어레이에서 동일 클래스의 유전자 벡터와 유전자 벡터사이의 피어슨 상관계수를 계산한다. 비교한 유전자 쌍의 개수만큼 나누어 유전자 모듈의 평균 상관계수를 구한다.

즉 유전자 모듈 [ga-gb-gd-gf]의 경우, 아래와 같이 평균 상관계수가 계산된다.

$$\frac{[PCC(g_a,g_b)+PCC(g_a,g_d)+PCC(g_a,g_f)+PCC(g_b,g_d)+PCC(g_b,g_f)+PCC(g_d,g_f)]}{6}$$

우선순위 큐를 사용하여 유전자 모듈의 상관계수가 큰 순서대로 삽입한다. 상관관계가 클수록 정확하고 분명한 판별이 가능하다고 볼 수 있다. 암 샘플에서 평균상관계수가 높은 상위 k개의 유전자 분류 모듈이 암 샘플의 분류자를 구성하고 정상샘플에서 평균상관계수가 높은 상위 k개의 유전자 분류 모듈이 정상샘플의 분류자를 구성한다. 상위 k개의 다양한 유전자 분류 모듈을 앙상블 형태로 사용하는 것이 단지 길이만 긴 하나의 모듈을 사용하는 것보다 분류의 정확도와 신뢰도를 높일 수 있다.

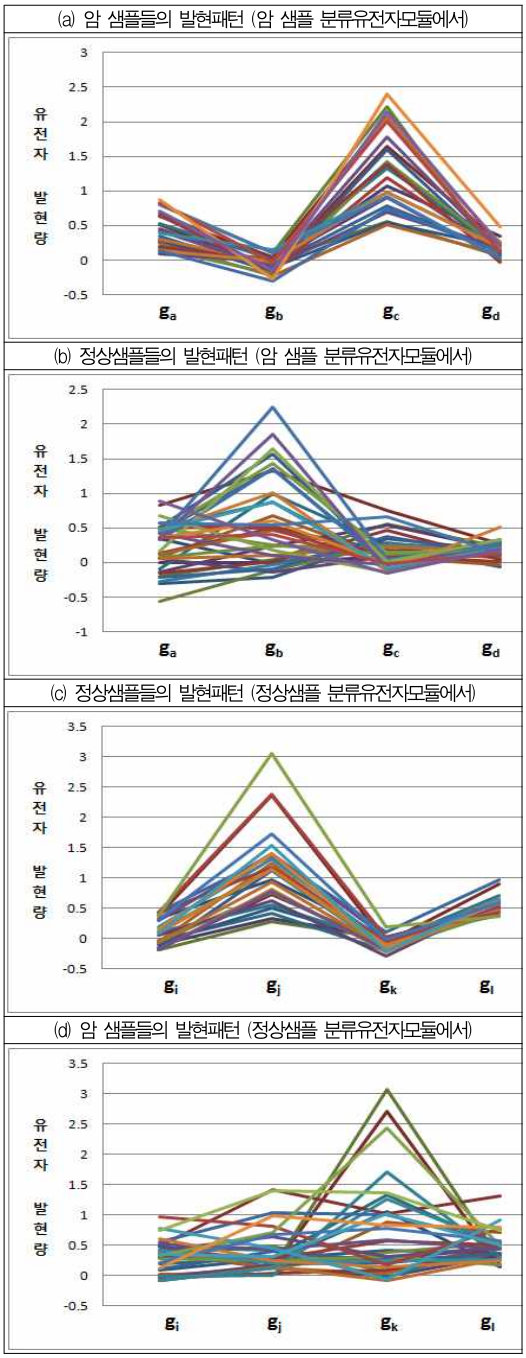

그림 4. 분류 유전자모듈에 따른 정상샘플들과 암샘플들의 발현패턴(각 샘플은 서로 다른 색으로 표시되어 있다)
Fig 4. Gene expression pattern of normal and tumor samples according to each classifying gene-module (Each sample is marked with a different color)

그림 4는 하나의 암 샘플 분류 유전자 모듈을 구성하는 유전자가 {g_a, g_b, g_c, g_d}이고, 하나의 정상 샘플 분류 유전자 모듈을 구성하는 유전자가 {g_i, g_j, g_k, g_l}일 때 각 유전자 분류 모듈의 발현 패턴이 암 샘플과 정상샘플에서 확연하게 차이가 나는 것을 보여준다. 그림 4에서 각 샘플은 서로 다른 색으로 표시되어 있다. 그림 4의 (a)에서와 같이 암 샘플 분류 유전자 모듈은 암 샘플에서 일정한 패턴을 보이나, 그림 4의 (b)에서와 같이, 정상샘플에 적용하면 패턴이 사라진다. 또한 그림 4의 (c)에서와 같이 정상 샘플 분류 유전자모듈은 정상샘플에서는 일정한 패턴을 보이나, 그림 4의 (d)에서와 같이 암 샘플에 적용하면 패턴이 사라진다. 이런 성질은 다음 단계인 판별에서 사용된다.

2.3 판별(Prediction)

2.1에서 유전자 쌍을 찾고 2.2에서 계산된 k개의 유전자 분류 모듈로 이루어진 분류자를 이용하여 독립된 샘플의 클래스를 구별할 수 있다. 암 샘플과 정상 샘플에 특이적으로 반응 하는 유전자 분류 모듈을 이루고 있는 유전자 벡터와 클래스를 모르는 독립 샘플데이터의 유전자벡터와의 상관관계를 이용하여 클래스를 구분할 수 있다.

그림 5에서 암 샘플을 분류하는 유전자 분류 모듈의 유전자 벡터 {g_a, g_b, g_c, g_d}를 TM(Tumor Module), 정상 샘플을 분류하는 유전자 분류 모듈의 유전자 벡터 {g_i, g_j, g_k, g_l}를 NM(Normal Module)라고 한다. TM는 총 n개의 암 샘플을, NM는 총 m개의 정상샘플을 가지고 있다. TM의 암 샘플 발현값 벡터는 T_i이고 NM의 정상 샘플 발현값 벡터는 N_i이다.

TM와 NM을 이용하여 클래스를 모르는 독립 샘플(Independent Sample)의 전체 유전자 벡터를 I, 독립 샘플의 TM 유전자 벡터를 I_{TM}, 독립 샘플의 NM 유전자 벡터를 I_{NM} 라 한다.

(a) 암 샘플 분류 유전자 모듈의 샘플과 독립 샘플

	T_1	T_2	T_3	…	T_n	I_{TM}
g_a	$g_{a,T1}$	$g_{a,T2}$	$g_{a,T3}$		$g_{a,Tn}$	$g_{a,I}$
g_b	$g_{b,T1}$	$g_{b,T2}$	$g_{b,T3}$		$g_{b,Tn}$	$g_{b,I}$
g_c	$g_{c,T1}$	$g_{c,T2}$	$g_{c,T3}$		$g_{c,Tn}$	$g_{c,I}$
g_d	$g_{d,T1}$	$g_{d,T2}$	$g_{d,T3}$		$g_{d,Tn}$	$g_{d,I}$

(b) 정상 샘플 분류 유전자 모듈의 샘플과 독립 샘플

	N_1	N_2	N_3	…	N_m	I_{NM}
g_i	$g_{i,N1}$	$g_{i,N2}$	$g_{i,N3}$		$g_{i,Nm}$	$g_{i,I}$
g_j	$g_{j,N1}$	$g_{j,N2}$	$g_{j,N3}$		$g_{j,Nm}$	$g_{j,I}$
g_k	$g_{k,N1}$	$g_{k,N2}$	$g_{k,N3}$		$g_{k,Nm}$	$g_{k,I}$
g_l	$g_{l,N1}$	$g_{l,N2}$	$g_{l,N3}$		$g_{l,Nm}$	$g_{l,I}$

그림 5. 암, 정상 클래스의 유전자모듈과 판별을 위한 독립 샘플
Fig 5. Gene module for Tumor, Normal class and an Independent sample for prediction

유전자 분류 모듈의 샘플 데이터들과 클래스를 모르는 독립 샘플 데이터의 상관관계를 알기 위해 피어슨의 상관계수를 이용해서 상관계수를 구한다. 파라미터에 의해 암 샘플과 정상 샘플에 특성을 가지는 유전자 분류 모듈을 찾을 때에는 유전자 발현값 벡터 사이의 상관계수를 구했으나 판별 과정에서는 샘플과 샘플 사이의 상관계수를 이용한다.

TM에서의 n개 암 샘플과 독립 샘플의 TM 유전자 벡터 데이터끼리의 상관계수(PCC(T_i, I_{TM}))의 평균, $Avg_{TM,I}$ 을 구한다. 또한 NM에서의 m개 정상 샘플과 독립 샘플의 NM 유전자 벡터 데이터끼리의 상관계수(PCC(N_i, I_{NM}))의 평균, $Avg_{NM,I}$도 구한다. 그 식은 다음과 같다.

$$Avg_{TM,I} = \frac{\sum_{i=1}^{n} PCC(T_i, I_{TM})}{n}$$

$$Avg_{NM,I} = \frac{\sum_{i=1}^{m} PCC(N_i, I_{NM})}{m}$$

표 2. 판별
Table 2. Prediction

$Avg_{TM,I}$	$Avg_{NM,I}$	판별	독립 샘플
높음	낮음	$Avg_{TM,I} > Avg_{NM,I}$	암 샘플
낮음	높음	$Avg_{TM,I} < Avg_{NM,I}$	정상 샘플

표 2처럼 $Avg_{TM,I}$가 $Avg_{NM,I}$보다 높으면 TM와 독립 샘플의 유전자 벡터가 상관관계가 높으므로 독립 샘플의 클래스는 암 샘플이라고 판별할 수 있다. 반면에 $Avg_{NM,I}$이 $Avg_{TM,I}$보다 높다면 NM와 독립 샘플의 유전자 벡터가 상관관계가 높으므로 독립 샘플의 클래스는 정상 샘플이라고 판별할 수 있다. 판별된 클래스와 원래 클래스를 비교하여 본 알고리즘의 정확도를 측정한다.

3. 실험결과

3.1 실험 설명 및 실험 환경

본 실험에서는 전립선 암 조직 12600개의 유전자 발현값을 측정한 마이크로어레이 데이터인 Singh[23], Welsh[24], LaTulippe[25]를 사용했다. 각 데이터의 암, 정상 샘플의 개수를 표 3에서 나타내었다.

표 3. 데이터의 암 샘플과 정상 샘플의 개수
Table 3. The number of data sample

	암 샘플 개수	정상 샘플 개수	전체 샘플의 개수
Singh	52	50	102
LaTulippe	23	3	26
Welsh	24	9	33

Welsh와 LaTulippe는 비교적 적은 수의 샘플을 가지고 있기 때문에 두 데이터를 통합하여 사용했다. Singh를 트레이닝 데이터로, Welsh와 LaTulippe를 합친, L_W 데이터를 테스트 데이터로 사용하였다. 특히 두 데이터를 통합할 때에는 Z-score를 사용하여 정규화(normalization)하였다.

실험에서 사용된 분류 알고리즘 (Decision tree, Naïve Bayesian, k-nearest neighbor, Random Forest, Support Vector Machine)은 Weka v.3.7.0의 구현을 이용했으며, 5가지 알고리즘 모두 Weka에서 제공되는 디폴트 파라미터를 사용했다[26]. 실험은 Genuine Intel(R) U7300 1.3GHz의 CPU와 4.00GB의 메모리, Window 7 Ultimate K 운영체제가 설치된 PC에서 수행했다.

3.2 변수를 위한 최적값 찾기

알고리즘에서 사용한 변수(parameter)들의 최적 값을 찾기 위하여 10-중첩 교차타당법(10-fold Cross Validation)을 실행 하였다.

데이터로는 Singh 데이터의 정상샘플과 암샘플 각 5개씩 추출하여 10개의 샘플을 한 세트로 묶은 데이터를 테스트 데이터로, 10개의 샘플 이외의 나머지 샘플 92개를 트레이닝 데이터로 사용하였다. 같은 방식으로 데이터를 분류하여 총 10쌍의 트레이닝 샘플과 테스트 샘플을 만들어서 실험을 하여 각각의 변수 값에서의 정확도를 평균으로 비교해 보았다.

표 4는 각 세트에서 유전자 쌍의 오차율을 줄여주는 n과 s의 따른 정확도 평균을 나타낸 것이다.

표 4. n과 s의 변화에 따른 정확도 비교
Table 4. Accuracy compared with changing n and s

s \ n	0.08	0.09	0.10	0.11	0.12
0.08	0.8	0.8333	0.8111	0.8	0.7889
0.09	0.8375	0.833	0.8556	0.8444	0.8
0.10	0.8125	0.8222	0.8	0.8	0.733
0.11	0.8111	0.8111	0.8222	0.8	0.79
0.12	0.85	0.8111	0.8	0.8	0.83

알고리즘의 중요변수인 n과 s를 0.08에서 0.12까지 범위를 지정하여 각 실험에서의 정확도를 비교해보았다. 각각의 n과 s 값에 대해 10번씩 실험을 하였고 표는 각 값들의 정확도 평균을 나타낸 것이다. n과 s가 작으면 작을수록 유전자 쌍을 찾을 수 있는 범위가 좁아져 적게 찾아지는 것을 확인할 수 있었다. 각 n과 s를 변경해가면서 실험을 한 결과 n과 s 모두 0.08부터 0.10 사이에서 큰 정확도를 보인다. 그 중에서도 n과 s가 작으면서 정확도가 큰 경우는 n이 0.08, s가 0.09일 때이다. 유전자 모듈의 유전자의 개수뿐만 아니라 클래스를 모르는 테스트 샘플을 판별할 때 쓰이는 유전자 모듈의 최적 개수를 알기 위해 n과 s의 값을 고정해놓고 분류자의 개수를 변경하면서 실험을 했다. 이 역시 Singh을 이용한 10-중첩 교차 타당법 실험 결과이다.

표 5에서 알 수 있듯 교차 타당법 실험결과 분류자를 구성하는 유전자 분류모듈의 개수를 2개로 했을 때 가장 높은 정확도를 보였고 최종적으로 n은 0.08, s는 0.09의 값, 분류자의 유전자 모듈의 개수는 2개로 실험했을 때 가장 좋은 정확도를 보였다.

표 5. 분류자를 구성하는 모듈 개수에 따른 중첩 교차 타당법의 정확도와 분류자생성비율 (n=0.08, s=0.09으로 고정)
[분류자 생성비율 = 분류자 생성횟수 / 전체 실험횟수]
Table 5. Accuracy comparison of cross validation varing the number of modules for classifier (n=0.08, s=0.09)

유전자 모듈 개수	1	2	3	4	5
Accuracy	0.8375	0.8444	0.8	0.9	0.775
분류자 생성비율	0.8	0.8	0.6	0.6	0.4

3.3 정확도 비교

두 클래스만이 주어진 경우, 긍정 투플(Positive tuple; 본 알고리즘에서 암 샘플)과 부정 투플(Negative tuple; 본 알고리즘에서 정상 샘플)로 나눌 수 있다. TP(true positive)는 분류기에 의해 올바르게 분류된 암 샘플의 수를 의미하고, TN(true negative)은 분류기에 의해 올바르게 분류된 정상샘플의 수를 의미한다. FP(false positive)는 분류기에 의해 잘못 분류된 정상샘플의 수를 의미하고, FN(false negative)은 분류기에 의해 잘못 분류된 암 샘플의 수를 의미한다.

이 실험에서는 성능평가기준의 지표를 총 5개로 지정하고 비교 및 분석했다. 분류기의 정확성을 나타내는 정확도(Accuracy), 분류기가 판별한 암 샘플의 클래스에 대한 정확성을 나타내는 TP-rate, 판별한 암 샘플의 클래스에 대한 부정확성을 나타

내는 FP-rate를 사용했다. 분류기가 암으로 판별한 샘플들 중에서 클래스를 정확히 판별한 샘플의 비율을 나타내는 Precision, 원래 데이터의 암 샘플 중 정확히 판별한 샘플의 비율인 Recall(TP-rate값과 같다)을 사용하였다[27]. 또한 테스트 데이터의 샘플이 판별된 클래스에 대한 신뢰성을 나타내는 F-measure를 사용하였다[28]. F-measure는 Recall과 Precision에 동등한 중요도를 부여하여 Recall과 Precision의 합으로 그 값의 두 배에 해당하는 값을 나누어 계산하는 평가방법 중 하나이다. 알고리즘에서 이 다섯 가지 척도는 알고리즘의 실용성과 효율성을 비교하는데 좋은 기준이 될 수 있다[28].

$$\text{정확도} = \frac{TN + TP}{TP + FP + FN + TN}$$

$$TP-rate = \frac{TP}{TP + FN}$$

$$FP-rate = \frac{FP}{TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$F-measure = \frac{2\text{Precision} \cdot Recall}{\text{Precision} + Recall} = \frac{2TP}{2TP + FP + FN}$$

표 6은 본 알고리즘(Gene Module)과 기존 분류기법에서의 다섯 가지 척도를 비교한 것이다. 암 샘플을 올바르게 분류한 비율을 나타내는 TP-rate가 다른 분류기에 비해 본 알고리즘이 매우 높은 것으로 나타났다.

표 6. 분류기법들의 성능 평가기준 비교
Table 6. Comparison of performance measures for classification methods

	TP-rate	FP-rate	Precision	F-measure
Gene Module	1	0.083	0.9787	0.9892
Random Forest	0.979	0.25	0.939	0.958
SVM	0.936	0.083	0.978	0.957
NB	0.957	0.083	0.978	0.968
Decision Tree	0.745	0.167	0.946	0.833
k-NN	0.957	0.083	0.978	0.968

아래 그림 6은 본 논문의 알고리즘 (Gene Module) 과 다른 분류 기법에 따라 측정된 정확도와 FP-rate를 나타낸 것이다.

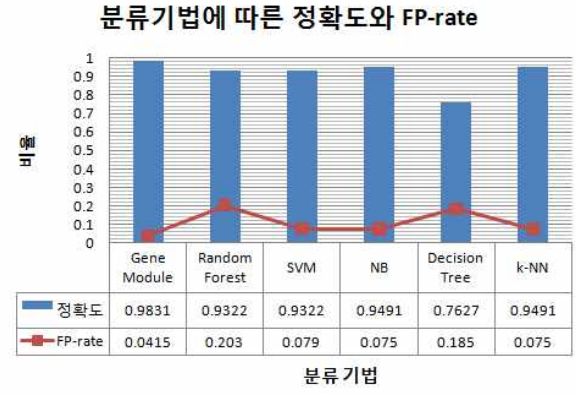

그림 6. 분류기법에 따른 정확도 비교
Fig 6. Accuracy Comparison of Classification Methods

분류 기법에 따른 평가 기준 중에서도 정확도는 분류 기법의 활용도를 비교하는데 지표가 될 수 있다. 반대로 FP-rate가 높으면 높을수록 클래스를 판별하는 정확도가 낮다고 할 수 있다. 본 논문의 알고리즘의 경우 다른 분류기법과 비교했을 때 매우 높은 정확도를 가지고 있음을 그래프에서 확인할 수 있다. 암샘플 분류의 부정확한 비율을 나타내는 FP-rate도 다른 분류기법보다 낮다.

IV. 결 론

본 논문에서는 1차적으로 트레이닝 마이크로어레이 데이터를 이용하여 암 샘플에 특이적으로 반응하는 유전자 쌍과 정상 샘플에 특이적으로 반응하는 유전자 쌍을 찾는다. 그리고 1단계에서 찾은 유전자 쌍을 이용하여 상관관계가 높은 유전자 모듈을 만든다. 유전자 모듈 역시 암 샘플과 정상 샘플을 구분하여 구성한다. 상위 유전자 모듈 k개를 분류자로 지정하여 클래스(암, 정상)를 모르는 샘플을 판별한다.

마이크로어레이의 발현데이터 값에는 시스템적으로 잡음이 내포될 수 있으므로 잡음에 견고하게 작동할 수 있는 알고리즘이 필요하다. 본 연구에서는 이 노이즈를 보정하기 위하여 두 개의 파라미터를 사용한다.

기존의 암분류 방법들이 암과 정상 샘플에서 차이가 나는 유전자를 특이 유전자로 식별하여 분류자로 사용하는데 반해 본 연구는 유전자와 유전자의 상관관계의 차이를 이용하여 분류모듈을 구성하여 분류자로 활용하였다. 분류모듈을 구성하

는 유전자의 수가 상대적으로 적으므로 임상키트로의 개발도 고려할 수 있다. 향후 GO(Gene Ontology)을 기반으로 분류 모듈의 기능적 검증을 수행하여, 밝혀지지 않은 새로운 암 관련 유전자를 식별하고 분류 모듈을 확대하여 암 특이적 유전자조절네트워크 구성에 활용할 계획이다.

참고문헌

[1] Daniela Dunkler, Michael Schemper and Georg Heinze, "Gene selection in Microarray survival studies under possibly non-proportional hazards.", BIOINFORMATICS, Vol.26, No.6, p.p. 784-790, 2010.

[2] Yuhang Wang, Fillia S. Makedon, James C. Ford, and Justin Pearlman, "HykGene: a hybrid approach for selecting maker genes for phenotype classification using microarray gene expression data", BIOINFORMATICS, Vol.21, No.8, pp. 1530-1537, 2005.

[3] 안재균, 윤영미, 신은지, 박상현, "유전자 발현값 상관관계 분석을 통한 암분류자 생성방법.", 제32회 한국정보처리학회 추계학술대회 논문집 제16권 2호, 769-770쪽, 2009년 11월

[4] 김선, 김수진, 장병탁, "마이크로어레이 기반 miRNA 모듈 분석을 위한 하이퍼망 분류기법", 정보과학회 논문지: 소프트웨어 및 응용, 제 35권, 제 6호, 347-356쪽, 2008년. 6월.

[5] Donald Geman, Christian d'Avignon, Daniel Q. Naiman, Raimond L. Winslow, "Classifying Gene Expression Profiles from Pairwise mRNA Comparisons.", Statistical Applications in Genetics and Molecular Biology, Vol.3, Issue.1, Article.19, 2004

[6] Junhee Seok, Amit Kaushal, Ronald W Davis, and Wenzhong Xiao, "Knowledge-based analysis of microarrays for the discovery of transcriptional regulation relationships.", BMC Bioinformatics, 2010. 1.

[7] 윤영미, 이종찬, 박상현, "두 단계 접근법을 통한 통합 마이크로어레이 데이터의 분류자 찾기", 정보과학회논문지: 데이터베이스 제34권, 제1호, 193-205쪽, 2007년 2월.

[8] 서울대학교 통계학과 생물정보통계연구실, "마이크로어레이 자료의 통계적 분석", 자유아카데미, 21-33쪽, 2005년.

[9] Guo Yu, Statistical issues in microarray data analysis: Array-to-array normalization, Empirical Bayes batch effect adjustment.

[10] Carla s. Moller-Levet, Catharine M. West, Crispin J. Miller, "Exploiting sample variability to enhance multivariate analysis of microarray data", BIOINFORMATICS, Vol.23, pp.2733-2740, 2007

[11] Breiman L, "Random Forest, Machine Learning", 45, p.p. 5-32, 2001 .

[12] 윤태균, 이관수, "의료진단 및 중요검사 항목 결정 지원 시스템을 위한 랜덤 포레스트 알고리즘 적용", 전기학회 논문지 제 57권, 제 6호, 1058-1062쪽, 2008년 6월.

[13] Ramon Diaz-Uriarte, Sara Alvarez de Andres, "Gene selection and classification of microarray data using random forest.", BMC Bioinformatics, p.p.1471-2105, 2006.

[14] Elias Zintzaras, Axel Kowald, "Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data.", ELSEVIER, Computers in Biology and Medicine 40, p.p.519-524, 2010.

[15] Rameswar Debnath, Takio Kurita, "An evolutionary approach for gene selection and classification of microarray data based on SVM error-bound theories.", ELSEVIER, BioSystems 100, p.p.39-46, 2010.

[16] 원홍희, 조성배, "암 분류를 위한 기계학습 분류기의 성능평가", 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호, 2002년 11월.

[17] 홍진혁, 조성배, "나이브 베이스 분류기를 이용한 유전발현 데이터기반 암 분류를 위한 순위기반 다중 클래스 유전자 선택", 정보과학회논문지: 시스템 및 이론, 제35권, 제8호, 372-377쪽, 2008년 8월.

[18] Gashier M., Giraud-Carrier C., Martinez T., "Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous.", Seventh International Conference on Machine Learning and Application, p.p.900-905, 2008.

[19] Jorng-Tzong Horng, Li-Cheng Wu, Baw-Juine Liu, Jun-Li Kuo, Wen-Horng Kuo, Jin-Jian

Zhang, "An expert system to classify microarray gene expression data using gene selection by decision tree.", Elsevier, Expert Systems with Applications, Vol.36, Issue.5, p.p.9072-9081, 2009.

[20] Chunrong Cheng, Kui Shen, Chi Song, Jianhua Luo, George C. Tseng, "Ratio adjustment and calibration scheme for gene-wise normalization to enhance microarray inter-study prediction.", Bioinformatics, p.p.1655-1661, 2009.

[21] Aleksey Fadeev, Oualid Missaoui, Hichem Frigui, "Ensemble Possibilistric k-NN for Functional Clustering of Gene Expression Profiles in Human Cancers Challenge.", icmla, 2009 International Conference on Machine Learning and Applications, p.p.439-442, 2009.

[22] Aik Choon Tan, Daniel Q. Naiman, LeiXu, Raimond L. Winslow and Donald Geman, "Simple decision rules for classifying human cancers from gene expression profiles.", BIOINFORMATICS, Vol.21, No.20, p.p. 3896-3904, 2005

[23] Singh D., Febbo P. G., Ross K., Jackson D. G., Manola J., Ladd C., "Gene expression correlates of clinical prostate cancer behavior", cancer cell, vol.1, pp.203-209, 2002

[24] Welsh J. B., Sapinoso L, M., Su A. I., Kern S. G., Wang-Rodriquez J., Moskaluk C. A., "Analysis of gene expression identifies candidate markers and pharmachological targets in prostate cancer", Cancer Research, Vol. 61, pp.5974-5978, 2001

[25] LaTulippe E., Satagopan J., Smith A., Scher H., Scardino P., Reuter V., "Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease", Cancer Research, Vol.62, pp.4499-4506, 2002

[26] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Written ; The WEKA Data Mining Software, An Update, SIGKDD Explorations, Vol.11, Issue 1, 2009.

[27] *http://www.hsl.creighton.edu/hsl/searching/Recall-Precision.html*

[28] 박윤정, 박승수, "암 분류를 위한 분류기법의 성능비교", 한국 컴퓨터 종합 학술대회 논문집, Vol.33, No.1, 220-222쪽, 2006년.

저 자 소 개

정 현 이

2007년 3월 : 가천의과학대학교 IT학부 유비쿼터스컴퓨팅학과 입학.

관심분야 : 바이오인포매틱스, 데이터마이닝.

윤 영 미

1981년 2월 : 서울대학교 자연과학대학 졸업(학사).

1981년 6월~1983년 6월 : 오하이오 주립대학 수학과(학사수료)

1987년 3월 : 스탠포드대학교 컴퓨터과학과 졸업(이학석사)

2008년 8월 : 연세대학교 컴퓨터과학과 졸업(공학박사)

1987년 5월~1993년 5월 : IntelliGenetics Inc., California, USA, Software Engineer

1995년 2월~현재 : 가천의과학대학교 정보공학부 교수

관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스