암 분류를 위한 음의 상관관계 특징을 이용한 앙상블 분류기

(Ensemble Classifier with Negatively Correlated Features for Cancer Classification)

원 홍 희 † 조 성 배 ††
(Hong-Hee Won) (Sung-Bae Cho)

요 약 최근의 DNA 마이크로어레이 기술로 많은 양의 유전자 데이타를 얻을 수 있는데, 특히 암의 진단과 치료에 적용되어 암의 정확한 분류에 많은 도움을 줄 것으로 기대된다. DNA로부터 얻어지는 유전자 데이타의 양은 매우 방대하므로 이를 효과적으로 분석하는 것은 매우 중요하다. 암의 분류는 진단과 치료에 있어 매우 중요하므로 하나의 분류기에 의존한 분류 결과보다는 다수의 전문화된 분류기 결과를 결합하여 결과를 도출하는 것이 바람직하다. 일반적으로 분류기를 결합함으로써 분류 성능 및 분류 결과에 대한 신뢰도를 높일 수 있다. 앙상블 분류기의 많은 장점에도 불구하고, 오류 의존적인 분류기의 결합은 성능 향상에 한계가 있다. 본 논문에서는 암을 정확하게 분류하기 위해서 음의 상관관계를 갖는 특징으로 학습한 신경망 분류기를 결합하는 방법을 제안하고, 제안한 방법의 유용성을 체계적으로 분석하고자 한다. 세 가지 벤치마크 암 데이타에 대하여 제안한 방법을 적용하여 실험한 결과, 음의 상관관계 특징을 이용한 앙상블 분류기가 다른 분류기보다 높은 성능을 내는 것을 확인할 수 있었다.

키워드 : DNA 마이크로어레이, 유전자 발현 정보, 암 분류, 특징 추출, 분류기, 앙상블 분류기, 음의 상관 관계

Abstract The development of microarray technology has supplied a large volume of data to many fields. In particular, it has been applied to prediction and diagnosis of cancer, so that it expectedly helps us to exactly predict and diagnose cancer. It is essential to efficiently analyze DNA microarray data because the amount of DNA microarray data is usually very large. Since accurate classification of cancer is very important issue for treatment of cancer, it is desirable to make a decision by combining the results of various expert classifiers rather than by depending on the result of only one classifier. Generally combining classifiers gives high performance and high confidence. In spite of many advantages of ensemble classifiers, ensemble with mutually error-correlated classifiers has a limit in the performance. In this paper, we propose the ensemble of neural network classifiers learned from negatively correlated features using three benchmark datasets to precisely classify cancer, and systematically evaluate the performances of the proposed method. Experimental results show that the ensemble classifier with negatively correlated features produces the best recognition rate on the three benchmark datasets.

Key words : DNA microarray, gene expression data, cancer classification, feature selection, classifier, ensemble classifier, negative correlation

1. 서 론

복잡한 조절 기능을 갖는 생명 현상에 대한 분자 수준의 단편적인 이해는 한계가 있기 때문에 인간 게놈 프로젝트(Human Genomic Project: HGP)와 같이 전체적인 이해를 위한 연구의 필요성이 대두되었다. 이를 위해서는 염기서열의 기능을 이해하는 것이 필수적이기 때문에 DNA 칩이 개발되었다. 최근 cDNA 마이크로어

· 본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임(과제고유번호:02-PJ1-PG1-CH04-0001)
† 학생회원 : 연세대학교 컴퓨터과학과
cool@candy.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 부교수
sbcho@cs.yonsei.ac.kr
논문접수 : 2003년 1월 20일
심사완료 : 2003년 8월 13일

레이와 oligonucleotide 마이크로어레이 기술의 발전으로 엄청난 양의 유전자 정보를 얻을 수 있다. DNA 칩 기술의 발전은 유전자 정보의 대량 생산을 가능하게 하였고, 특정한 실험 환경[1]과 조건에 따른 수천 개의 유전자 발현 정도를 동시에 파악할 수 있게 하였으며, 이를 대량으로 처리함으로써 수천 개의 유전자 정보를 굉장히 빠르고 정확하게 분석할 수 있게 되었다[2].

DNA 마이크로어레이 기술은 암의 예측과 진단 분야에도 적용되어 암의 정확한 분류에 많은 도움을 줄 것으로 예상된다. 특히 암의 정확한 분류는 그 치료에 있어서 매우 중요한 문제이기 때문에 유전자 정보를 이용하여 암을 분류하는 문제에 관한 많은 연구가 진행되고 있다. 일반적으로 DNA 마이크로어레이 데이타는 방대한 양의 유전자 정보를 갖고 있으므로, 이를 효율적으로 분석하기 위하여 다양한 데이타 마이닝 기법, 기계학습 알고리즘, 통계적인 방법 등을 적용하고 그 성능을 평가하려는 연구가 진행되고 있다[3,4]. 그러나 대부분의 연구들은 각 특징 추출 방법과 개별 분류 방법에 따른 성능만을 주로 평가하였다.

개별 분류기 만의 성능을 평가하는 데 그치지 않고, 특징의 상관관계를 기준으로 각 분류기의 결과를 결합함으로써 분류 성능을 향상시키려는 연구도 있었다[5]. 특징 추출 방법인 Pearson's correlation coefficient와 Euclidean distance가 음의 상관관계가 있음을 이용하여 특징의 음의 상관관계를 정의하고, 두 가지 특징 추출 방법에 의해 선택된 특징들로 분류기를 학습시킨 후, 이를 결합함으로써 기존의 분류기보다 나은 성능을 보였다[5]. 여기서 말하는 두 특징 추출 방법 간의 음의 상관관계는 Pearson's correlation coefficient는 그 값이 커질수록, Euclidean distance는 그 값이 작을수록 높은 유사도를 갖는다는 성질에 의한 것이다. 하지만 하나의 벤치마크 데이타에 대하여 실험하였기 때문에 실험 결과에 대한 충분한 검증이 되지 않았다. 따라서 다양한 벤치마크 데이타를 이용하여 분류기의 성능을 체계적으로 분석해 볼 필요가 있다. 또한 Cho[5]는 다양한 특징 추출 후에 추출된 특징들간의 상관관계를 분석함으로써 적절한 조합을 찾은 반면에, 본 논문에서 제안하는 방법은 특징 추출 단계 이전에서 데이타에 관한 지식을 이용하여 음의 상관관계를 갖는 특징들을 결합한다는 점에서 앞선 연구와 차이가 있다. 본 논문의 방법은 특징 추출 이전 단계에서 음의 상관관계를 정의하기 때문에 특징 추출 방법에 상관없이 음의 상관관계를 갖는 특징을 선택할 수 있는 장점이 있다.

본 논문에서는 음의 상관관계를 갖는 특징으로 학습한 신경망 분류기를 결합하는 방법을 제안한다. 세 가지 벤치마크 암 데이타에 대하여 제안한 방법을 적용하여 실험하고, 그 유용성을 체계적으로 분석하고자 한다. 특징 추출의 기준이 되는 이상적인 특징 벡터를 두 가지로 정의하고, 각각의 이상적인 특징 벡터와 유사한 유전자들을 선택하여 분류에 이용한다. 두 가지 이상적인 특징 벡터는 클래스 A에서는 높은 값을 갖고 클래스 B에서는 낮은 값을 갖는 벡터와 클래스 A에서는 낮은 값을 갖고 클래스 B에서는 높은 값을 갖는 벡터로 정의하였다. 두 벡터는 음의 상관관계를 가지므로 각 벡터와 유사한 유전자 벡터들의 집합은 서로 음의 상관관계를 갖는다. 음의 상관관계를 갖는 특징들은 학습 데이타의 두 가지 다른 측면을 대표하기 때문에 이러한 특징들을 결합함으로써 보다 넓은 해공간을 탐색할 수 있다[5].

2. 배 경

2.1 cDNA 마이크로어레이

DNA 마이크로어레이는 유전자 발현 정보를 얻기 위해 고형 지지체(substrate) 위에 실험하고자 하는 대량의 유전자를 고정해 놓은 것이다. 한번에 수백 개 이상의 유전자가 DNA 마이크로어레이 위에 고정되어 분석되기 때문에 대량의 유전자 데이타를 신속하게 제공한다. DNA 마이크로어레이 기술은 그림 1과 같이 로봇을 이용하여 고밀도의 어레이에 대량의 유전자 DNA 서열을 고정시키기 때문에 정확하고 효율적으로 유전자 발현정보를 제공한다.

두 개의 샘플로부터 추출한 실험 시료와 참조 시료의 DNA나 RNA 서열을 RT-PCR(reverse transcription polymerase chain reaction)방법을 통하여 역전사(reverse-transcription)시키는 과정에서 각각 다른 형광물질(빨간색의 형광물질 Cy5와 녹색의 형광물질 Cy3)로 염색하여 cDNA를 생성하고, 이를 보합하여 하나의 마이크로어레이 위에 고정시킨다. 두 개의 시료로부터 생성한 cDNA를 보합하여 마이크로어레이에 첨가하면

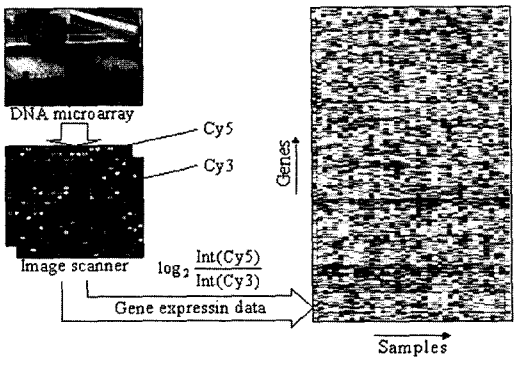


그림 1 cDNA 마이크로어레이로부터 유전자 발현정보를 얻는 과정

유전자의 상보적(complementary) 결합에 의해 시료에 발현된 유전자는 각 지점에 고정되고, 각 지점은 고정된 시료에 따라 빨간색 혹은 녹색을 띄게 된다. 레이져 형광 스캐너를 이용하여 마이크로어레이의 각 지점의 형광정도를 읽어 낸다. 각 유전자의 형광정도는 그 유전자의 발현정도를 나타내며 유전자 발현정보 데이타로 사용하기 위하여 Cy3와 Cy5 형광정도의 상대적 강도를 다음과 같이 구한다[6-8].

$$gene\_expression = \log_2 \frac{\mathrm{Int(Cy5)}}{\mathrm{Int(Cy3)}} \quad (1)$$

수식 (1)에서 Int(Cy5)와 Int(Cy3)는 빨간색 형광물질과 녹색 형광물질의 강도를 의미한다.

2.2 Oligonucleotide 마이크로어레이

고밀도의 oligonucleotide 칩 기술은 부분적으로 빛을 투과하여 화학적으로 합성시키는 기법을 이용하여 빛이 투과하는 유리 기판 위에 수 천 개의 oligonucleotide를 직접 합성하는 방식이다. 그림 2의 (a)와 같이 칩 표면에 빛에 불안정한 보호기(基)를 보유하는 링커 분자를 입히고, 유전자 조각을 심을 부분에 마스크를 이용하여 부분적으로 빛을 투과함으로써 보호기를 제거한다. 보호기가 제거된 부분에 광활성 부위에서만 융합하는 광 보호 nucleotide에 빛을 투과 시켜 줌으로써 nucleotide를 심는다. 이를 반복함으로써 그림 2의 (b)와 같이 유전자 조각의 길이가 대략 20-25 mers가 되도록 칩을 제작한다. Oligonucleotide 마이크로어레이 기술은 고밀도 집적이 가능하지만, 주로 20-25의 염기를 갖는 oligonucleotide만을 이용하게 된다. 이 oligonucleotide는 target DNA의 염기서열에 대한 지식을 기반으로 특정한 target 유전자에만 반응할 수 있도록 제작된다. 고밀도의 oligonucleotide 칩은 동족의 유전자를 고유하게 식별할 수 있도록 제작되므로 비슷한 염기서열을 갖는 유전자와 cross-hybridization 되는 것을 최소화하게 된다[1,9,10].

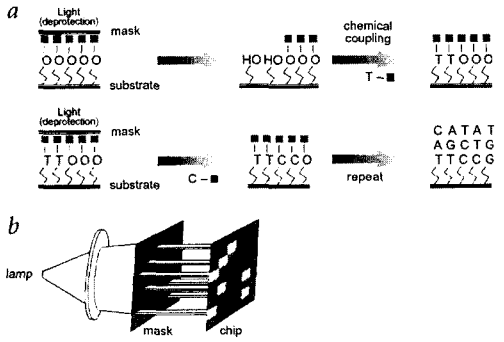


그림 2 photolithography를 이용한 oligonucleotide의 합성[10]

Affymetrix사에서 GeneChip® oligonucleotide 칩을 개발하였으며, 고밀도 DNA probe 칩을 만들기 위하여 photolithography와 solid-phase DNA 합성 기술을 이용하였다.

2.3 관련 연구

DNA 마이크로어레이 데이타의 양은 보통 매우 크기 때문에 이를 효율적으로 분석하는 것이 중요하다. DNA 마이크로어레이의 분석은 클러스터링(clustering), 분류(classification), 유전자 식별(gene identification), 유전자 네트워크 모델링(gene regulatory network modeling)과 같이 크게 네 가지 분야로 나눌 수 있다. 이러한 문제를 풀기 위하여 많은 기계학습과 데이타 마이닝 방법이 적용되고 있다.

정보이론[11]은 유전자 식별 문제에 적용되었으며, 부울 네트워크[12]와 베이시안 네트워크[13], reverse engineering[14] 방법 등이 유전자 네트워크 모델링 문제를 풀기 위하여 사용되고 있다.

Fisher의 선형 식별 분석[3], k 최근접 이웃[15], 결정 트리, 다층 퍼셉트론[16,17], support vector machine [18,19], 자기 구성 지도[20] 등의 다양한 기계학습 방법이 유전자 발현 데이타를 분류하는 문제에 적용되고 있다. 또한, 많은 기계학습 방법이 유전자 발현 데이타를 클러스터링하는 문제에 적용되고 있다[9]. 클러스터링 문제에 적용된 알고리즘으로 계층적 클러스터링[8], 자기 구성 지도[21] 및 그래프이론 기반 방법 등이 있다.

첫 번째 방법인 분류는 일반적으로 교사학습(supervised learning)이라 하고, 두 번째 방법인 클러스터링은 비교사학습(unsupervised learning)이라 한다. 클러스터링 방법은 분할 단계에서 샘플에 대한 정보(예, 암이나 정상)를 사용하지 않는다. 분류 방법은 외부의 "supervision"에 의해 이미 분류된 학습 데이타로 모델을 학습한 후에 이를 이용하여 새로운 샘플에 대하여 그 클래스를 예측한다[4].

분류 문제를 해결하는 데 있어 다양한 분류기를 결합함으로써 성능을 향상시키기 위한 많은 연구들이 진행되고 있다. 분류기의 결합은 분류 성능이 향상될 뿐만 아니라 분류 결과에 대한 신뢰도 또한 증가하는 장점이 있다. 이론적으로, 결합하고자 하는 분류기들이 서로 이질적(heterogeneous)이거나 오류 독립적이면 그 앙상블은 성능이 향상된다. 평균 결합, 투표 결합, 가중치 투표 결합, Bayesian 결합, 신경망 결합은 대표적인 앙상블 방법으로서 데이타 마이닝과 기계학습 등의 다양한 응용분야에 적용되고 있다. 하지만 이 방법들은 각 분류기가 오류 독립이라는 가정을 보장하지 못한다. 반면 boosting(bootstrap resampling)이나 bagging(bootstrap aggregating), arcing(adaptively resampling and

combining)과 같은 앙상블 방법은 다양한 샘플을 생성하고 각각을 학습시킴으로써 이질적인 분류기를 생성하고 결합한다[4,22].

3. 음의 상관관계 특징을 이용한 앙상블 분류기

본 논문에서 제안하는 음의 상관관계 특징을 이용한 앙상블 분류기의 구성은 그림 3과 같다. 음의 상관관계를 갖도록 정의한 두 개의 이상적인 특징 벡터인 *Ideal feature A*와 *Ideal feature B*를 이용하여 유전자 발현 데이타로부터 특징 집단을 추출하고, 분류기를 학습시킨다. 각각의 이상적인 벡터는 음의 상관관계를 갖기 때문에 이를 기준으로 추출한 두 개의 특징 집단은 음의 상관관계를 갖는다. 이 특징 집단은 특징 공간의 서로 다른 양상을 대표하기 때문에 이를 이용하여 학습시킨 분류기는 서로 이질적이다. 따라서 두 개의 분류기를 결합함으로써 보다 높은 분류 성능을 기대할 수 있다.

3.1 음의 상관관계

사용 가능한 다양한 특징이 존재할 때, 더 많은 특징을 사용할수록 많은 정보를 얻을 수 있기 때문에 다양한 특징을 사용하는 것은 분류에 더 효과적일 수 있다[5]. 그러나 특징 공간이 서로 겹쳐지는 특징들은 불필요한 정보의 잉여 문제를 야기하거나 과적합(overfitting)과 같은 문제를 초래하게 된다. 특징 선택 방법이 N개 존재한다고 할 때, 관찰 공간에서 특징 공간으로의 비선형 변환 함수들의 집합을 $\{\phi = \varphi_1, \varphi_2, \varphi_3, \ldots, \varphi_N\}$라 하고, $\phi_k \in 2^{\phi}$일 때, 특징 선택 방법들의 집합 ϕ_k가 분류기에 제공하는 분류 정보량 $I(\phi_k)$는 다음 수식과 같이 표현할 수 있다.

$$I(\phi_k) = \frac{a\sum A_i}{\frac{N}{2}\sum_{j=1, j\neq i}^{N} d_{ij}} + b \quad (2)$$

d_{ij}는 ϕ_k의 i 번째와 j 번째 원소간의 의존관계, A_i는 ϕ_k의 i 번째 원소가 특징 공간에서 차지하는 영역의 넓이, a와 b는 상수이다. 특징 쌍의 의존관계가 높을수록 분류 정보량 $I(\phi_k)$는 작아지고, 특징 선택방법들의 차지 영역의 넓이가 클수록 $I(\phi_k)$는 커진다. 만약 사용할 특징 선택 방법들의 개수를 크게 하면 차지하게 되는 영역이 더욱 넓어지게 되어 분자가 커지게 되지만 d_{ij}를 낮은 상태로 유지하도록 보장하지 못한다면 $I(\phi_k)$ 값이 전체적으로 감소하게 된다. 따라서 특징 선택을 통하여 분류기에 제공되는 정보량 $I(\phi_k)$를 크게 하기 위해서는 특징의 개수를 무조건 늘리는 것보다 상호 독립적인 소수의 특징을 사용하는 것이 더욱 효과적이다. 특징들 간의 상관관계는 특징값들의 분포나 통계학적 분석 기법을 사용하여 추론할 수 있다. 그러한 상호 독립적인 소수의 특징을 추출하기 위하여 특징간의 음의 상관관계를 이용하여 가능한 음의 상관관계 정도가 큰 특징들을 선택하여 이를 분류에 사용하였다.

M개의 샘플과 N개의 유전자를 갖는 $M \times N$ 행렬의 유전자 발현 데이타가 있고, M 개의 샘플은 클래스 A와 클래스 B로 나뉜다고 하면 각 유전자 데이타 g_i는 수식 (3)과 같은 벡터로 표현할 수 있다[5].

$$g_i = (e_1, e_2, e_3, \cdots, e_M) \ (i = 1 \sim N) \quad (3)$$

클래스 간의 뚜렷한 패턴의 차이가 존재하는 이상적인 유전자 벡터를 g_{ideal}이라 하고 수식 (4)와 같은 벡터로 표현했을 때, 분류에 사용하고자 하는 의미있는 유전자들은 이상적인 유전자 벡터 g_{ideal}과의 유사도가 큰 벡터라 정의할 수 있다.

$$g_{ideal} = (e_1', e_2', e_3', \cdots, e_M') \quad (4)$$

본 논문은 수식 (4)에 정의한 이상적인 유전자 벡터 g_{ideal}을 그림 4와 같이 클래스 A에서 높은 값을 갖고, 클래스 B에서는 낮은 값을 갖는 벡터와 클래스 A에서는 낮은 값을 갖고, 클래스 B에서는 높은 값을 갖는 두 개의 벡터로 정의하고, 각각의 이상적인 유전자 벡터와 유사한 유전자 벡터를 추출하고자 한다. 두 개의 이상적인 벡터 간의 Pearson's correlation coefficient는 -1이므로 완전한 음의 상관관계를 갖는다. 따라서 이를 바탕으로 추출한 두 유전자 벡터 집단(feature subset)도 음의 상관 정도가 큰 특징 관계가 된다. 음의 상관관계를 갖는 특징들은 학습 데이타의 두 가지 다른 측면을 대표하기 때문에 이러한 특징들을 결합함으로써 보다 넓은 해공간을 탐색할 수 있다[5].

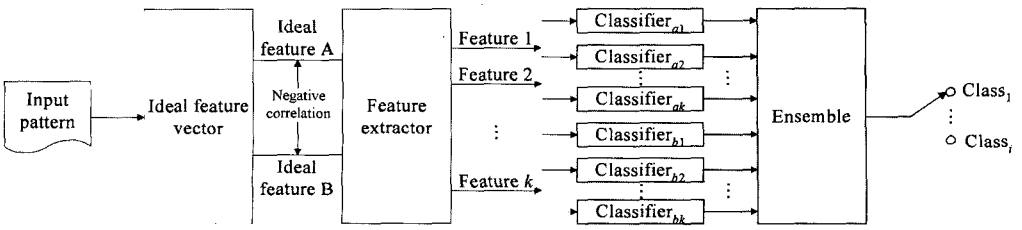


그림 3 음의 상관관계 특징을 이용한 앙상블 분류기의 개요

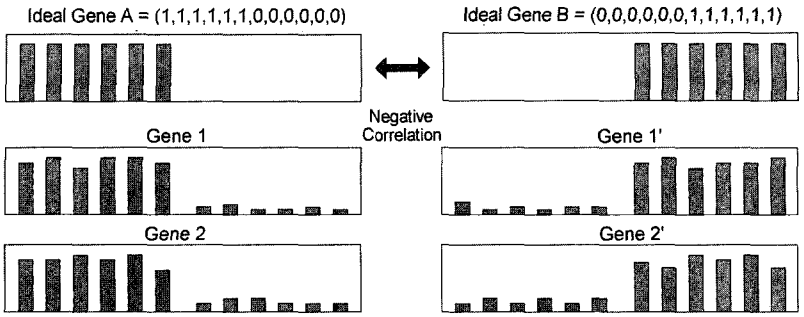


그림 4 음의 상관관계 특징에 의해 선택된 의미있는 유전자

3.2 특징 추출 방법

마이크로어레이로부터 얻어지는 유전자의 수는 대략 수천 개에서 수만 개이다. 하지만 얻어진 데이타에서 각 샘플의 특정 클래스와 연관이 있는 유전자의 수는 그보다 훨씬 작다. 따라서 유전자 데이타를 이용하여 클래스를 분류하기 위해서는 클래스와의 연관성이 높은 유전자를 추출하는 과정이 필요하다. 이러한 과정을 일반적으로 특징 추출 과정이라고 하며, 유전자 선택 과정이라고도 한다[15]. 주로 통계적인 상관관계 분석 방법과 클러스터링 기법 등을 이용하여 클래스와의 상관관계가 높은 유전자를 선택한다. 주성분 분석(principal component analysis)을 이용하여 특징의 차원을 줄이거나 유전자 알고리즘(genetic algorithm)을 이용하여 특징을 추출하는 등의 다양한 연구가 진행되고 있다[16,23].

본 논문에서는 유전자 벡터 g_i와 이상적인 유전자 벡터 A, B와의 유사도에 의하여 의미 있는 유전자를 추출하였다. 이상적인 유전자 벡터 A, B와 가장 유사한 25개의 유전자를 각각 선택하여 이를 분류를 위한 특징으로 사용하였다. 특징으로 사용한 유전자의 수를 변경해 가면서 실험한 결과 25~30개의 유전자를 사용하였을 때의 성능이 가장 우수하고 안정적이었기 때문에 25를 선택하였다. 유전자 벡터 g_i와 이상적인 유전자 벡터 g_{ideal}의 유사도를 측정하기 위하여 통계적 상관관계 분석과 거리 척도 방법을 사용하였다.

통계적 상관관계를 이용하여 두 벡터의 선형 관계 및 관계의 방향성을 분석할 수 있다. 상관관계 지수 r은 -1에서 +1의 값을 갖으며, +1에 가까울수록 두 벡터는 높은 양의 상관관계를, -1에 가까울수록 높은 음의 상관관계를, 0에 가까우면 상관관계가 없음을 나타낸다. 대표적인 통계적 상관관계 방법에는 Pearson's correlation coefficient와 Spearman's correlation coefficient 등이 있다. 두 벡터간의 유사성은 벡터 공간에서의 거리로 정의할 수 있으며, 벡터간의 거리가 작을수록 두 벡터는 유사하다. 대표적인 거리 척도에는 Euclidean distance와 cosine coefficient 등이 있다. 표 1은 음의 상관관계를 이용하여 의미 있는 유전자를 선택하기 위한 4가지 특징 추출법이며, 각 수식은 순서대로 Pearson correlation coefficient(PR), Spearman correlation coefficient (SP), Euclidean distance(ED), cosine coefficient(CC)를 의미한다.

표 1 음의 상관관계를 이용하여 의미 있는 유전자를 선택하기 위한 4가지 특징 추출법

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

$$SP(g_i, g_{ideal}) = 1 - \frac{6\sum (D_g - D_{ideal})^2}{N(N^2-1)}$$

(D_g and D_{ideal} are the rank matrices of g_i and g_{ideal})

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

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

3.3 기본 분류기

본 논문에서 앙상블을 위한 분류기로 다층신경망(multi-layer perceptron, MLP)을 사용하였다. MLP는 인공신경망의 대표적인 기계 학습 알고리즘으로서, 일반적인 패턴 인식 문제에서 강하고 안정적인 성능을 보인다[24]. MLP는 오류 역전파(error back-propagation) 알고리즘을 사용하여 신경망의 결과 값이 분류 목표치에 가까워지도록 연결 강도를 조절해나감으로써 주어진 패턴을 학습한다. 오류 역전파 학습 알고리즘은 크게 두 가지 장점이 있다. 첫째, 연결강도와 바이어스(bias)의 업데이트가 지역적으로 일어난다. 각 은닉층에서의 델타 값을 기준으로 오류가 발생하는 노드의 활성화를 선택적으로 억압하여 학습의 효율을 높여준다. 둘째, 네트워크의 모든 자유 매개변수에 대한 비용함수(cost func-

tion)의 부분 도함수 계산이 효율적이다. 오류 역전파 알고리즘에서의 연결 강도 업데이트 규칙은 수식 (5)와 같다.

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1) \quad (5)$$

$\Delta w_{ji}(n)$은 학습 알고리즘의 n번째 반복에서 업데이트된 연결 강도를 나타내며, η와 δ_j는 각각 학습률과 오류항(error term)을 의미한다. x_{ji}는 i 번째 뉴런의 값이고, $0 \leq \alpha < 1$는 학습의 관성을 주기 위한 상수인 모멘텀(momentum)을 나타낸다.

3.4 앙상블 분류기

분류기를 결합하기 위한 대표적인 앙상블 방법으로 평균 결합, 투표 결합, 가중치 투표 결합, Bayesian 결합, 신경망 결합 등이 있다. 짝수 개의 분류기 결합에서 각 클래스가 동일한 선택을 받는 경우 결과의 선택에 모호함이 발생하는 것을 막기 위해서, 본 논문에서는 각 분류기의 사전 정보를 이용하는 Bayesian 결합 방법을 사용하였다. 투표 방법은 각 분류기의 결과만으로 결합하는 반면, Bayesian 결합 방법은 각 분류기의 오류 가능성이 최종 결과에 영향을 미치도록 한다. 결과적으로 결합하는 분류기에 대한 사전 지식을 이용함으로써 분류기의 가중치를 달리하는 결합 방식이다. k개의 개별 분류기 결합에서 c_i (i=1,...,m)는 샘플의 실제 클래스이고 $c(classifier_j)$는 j 번째 분류기가 낸 클래스이며 η는 각 클래스 c_i의 사전 확률일 때, Bayesian 결합 방법은 수식 (6)과 같이 구한다.

$$c_{ensemble} = \arg\max_{1 \leq i \leq m} \{\eta \prod_{j=1}^{k} P(c_i \mid c(classifier_j))\} \quad (6)$$

4. 실험 결과

4.1 데이타 집합

유전자 발현정보를 이용한 암의 연구에서 사용된 많은 공개된 마이크로어레이 데이타가 있는데, 본 논문에서는 백혈병 데이타, 결장암 데이타, 림프종 데이타를 사용하였다. 백혈병 데이터와 림프종 데이타는 같은 질병의 두 가지 다른 타입으로부터 추출한 데이타이며, 결장암 데이타는 같은 조직의 암 세포와 정상 세포로부터 추출한 데이타이다. 이 데이타는 많은 논문에서 대상으로 하고 있기 때문에 본 논문의 결과를 객관적으로 평가할 수 있다.

• 백혈병 데이타(http://www.genome.wi.mit.edu/MPR): 백혈병 데이타는 72개의 샘플 데이타로 구성되어 있으며, 백혈병의 두 가지 종류인 급성 골수성 백혈병(acute myeloid leukemia, AML) 환자 25명과 급성 림프성 백혈병(acute lymphoblastic leukemia, ALL) 환자 47명으로부터 얻어진 데이타이다. 72개의 샘플 데이타 중에서 63개는 골수로부터 채취하였고, 나머지 9개는 말초 혈액으로부터 채취하여 고밀도 oligonucleotide 마이크로어레이를 사용하여 만들어졌다[4]. 72개의 샘플 중에서 38개를 학습 데이타로 사용하였고, 나머지 34개를 실험 데이타로 사용하였는데, 각 샘플은 7129개의 유전자 발현 정보를 갖고 있다.

• 결장암 데이타(http://www.sph.uth.tmc.edu:8052/hgc/default.asp): 결장암 데이타는 결장암 환자의 결장 상피 세포로부터 추출한 62개의 샘플 데이타이며, 각 샘플은 2000개의 유전자 발현 정보를 갖고 있다. 원래의 데이타는 6000개의 유전자 정보를 갖고 있었지만, 정확하지 않은 정보를 갖고 있는 4000개를 제거한 것이다. 62개의 샘플 데이타 중에서 40개는 암 세포의 샘플이며, 다른 22개는 정상 세포의 샘플이다. 각 샘플은 같은 환자의 암 부위와 정상 부위의 세포에서 채취되었으며, 고밀도 oligonucleotide 마이크로어레이를 사용하여 만들어졌다[4]. 62개의 샘플 중에서 31개를 학습 데이타로 사용하였고, 나머지 31개를 실험 데이타로 사용하였다.

• 림프종 데이타(http://genome-www.stanford.edu/lymphoma): B cell diffuse large cell lymphoma (B-DLCL)은 형태학이나 임상적인 상태와 약물 반응에 있어 이질적인 두 가지 종류가 있다. 한 종류가 germinal center B cell-like DLCL이고 나머지가 activated B cell-like DLCL이다[25]. 림프종 데이타는 GC B-like 샘플 24개와 activated B-like 샘플 23개로 구성되어 있다. 47개의 샘플 중에서 22개를 학습 데이타로 사용하였고, 나머지 25개를 실험 데이타로 사용하였으며, 각 샘플은 4026개의 유전자 발현 정보를 갖고 있다.

4.2 실험 환경

특징 추출 단계에서 표 1의 특징 추출 방법에 의해 각 유전자의 점수를 계산하고, 상위 25개의 유전자를 입력 패턴의 특징으로 사용하였다. 두 개의 이상적인 벡터 중 하나인 Ideal Gene A = (1,1,⋯,1,0,0,⋯,0)에 의해 선택된 특징들로 학습한 MLP를 MLP I으로, 또 다른 이상적인 벡터인 Ideal Gene B = (0,0,⋯,0,1,1,⋯,1)에 의해 선택된 특징들로 학습한 MLP를 MLP II로 정의하였다. 학습 단계에서 MLP의 모멘텀은 0.9로 정하였고 총 레이어의 수는 3으로 고정한 후에 학습률을 0.01에서 0.50으로 변화시키며 실험하였다. 또한 학습 데이타에 과적합되는 것을 막기 위하여 학습과정의 최대 반복은 100으로 고정하였다.

음의 상관관계 집합은 MLP I과 MLP II로 구성된 집합이며, 음의 상관관계 특징의 성능을 평가를 위하여 MLP I 집단과 MLP II 집단으로 각각 앙상블한 결과와 비교하였다. 음의 상관관계 집단으로 학습한 분류기를

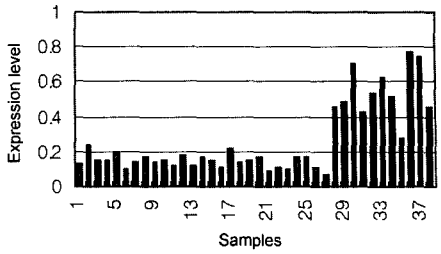
Bayesian 결합 방법을 사용하여 결합하고 그 결과를 분석하고 평가하였다. 음의 상관관계 집합의 개별 분류기는 8개이므로 ${}_8C_k$ (k=1, 2, 3, 4)개의 결합에 대해 앙상블하여 각 방법의 최대 인식률과 평균인식률을 구하였고, MLP I 집단과 MLP II 집단의 개별 분류기는 4개이므로 ${}_4C_k$ (k=1, 2, 3, 4)개의 결합에 대해 앙상블하여 각 방법의 최대 인식률과 평균인식률을 구하였다.

4.3 결과 분석

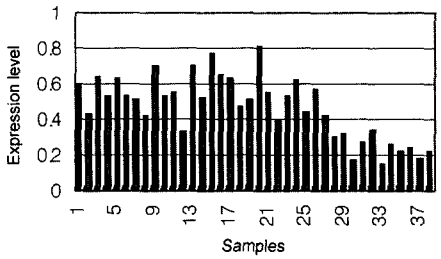
그림 5는 Pearson's correlation coefficient으로 선택된 25개 유전자의 평균 발현 정도를 나타낸다. (a)와 (b)는 각각 Ideal Gene A, Ideal Gene B를 이용하여 Pearson's correlation coefficient로 선택한 유전자의 발현 정도이다. 1~27샘플은 AML 클래스이며 28~38 샘플은 ALL클래스이다. 음의 상관관계에 의해 선택된 25개 유전자의 클래스간 발현 정도가 뚜렷하게 구분됨을 볼 수 있다.

표 2는 각 데이타 집합에 대한 분류기의 인식률을 나타낸다. 백혈병 데이타의 경우 Pearson's correlation coefficient와 MLP I의 조합이 97.1%의 인식률로 가장 우수하였다. 결장암 데이타의 경우 cosine coefficient와 MLP I의 조합이 83.9%의 인식률로 가장 우수하였다. 림프종 데이타의 경우 Spearman's correlation coefficient와 MLP II의 조합이 88.0%의 인식률로 가장 우수하였다.

Ideal Gene A로부터 선택된 특징을 사용한 MLP I과 Ideal Gene B로부터 선택된 특징을 사용한 MLP II는 데이타 집합에 따라 다른 성능을 보였다. 백혈병 데이타의 경우 MLP I이 MLP II 보다 우수하였고, 결장암 데이타의 경우 거의 유사하였다. 또한 림프종 데이타의 경우 MLP II가 MLP I 보다 우수하였다. 이는 각 데이타의 특성에 기인한다. 그림 5의 (a)와 (b)와 같이 각 클래스 마다 유전자의 발현 정도가 Ideal Gene A의 패턴이 Ideal Gene B의 패턴 보다 뚜렷한 경우에는 Ideal Gene A로부터 선택된 특징을 이용하는 것이 보다 나은 성능을 줄 수 있다. 표 2에서 Pearson's correlation coefficient와 MLP I의 조합이 Pearson's correlation coefficient와 MLP II의 조합보다 더 나은 성능을 보이는 것을 알 수 있다.



(a) Pearson's correlation coefficient for MLP I



(b) Pearson's correlation coefficient for MLP II

그림 5 Leukemia에서 Pearson's correlation coefficient로 선택된 25개 유전자의 발현 정도

그림 6, 7, 8은 각 데이타 셋에 대한 앙상블 분류기의 평균, 최대 인식률을 나타낸다. 그림의 x 축은 결합한 분류기의 수이며, y 축은 각 방법의 인식률이다. 음의 상관관계 집합(MLP I + MLP II)의 경우 2개의 이상적인 유전자 벡터(Ideal gene A, Ideal gene B)와 4개의 특징 추출 방법(Pearson's correlation coefficient, Spearman's correlation coefficient, Euclidean distance, cosine coefficient)의 조합으로 총 8개의 서로 다른 특징 집단을 생성하여 분류기를 학습시킨 후 이를 결합하였다. 평균 인식률은 ${}_8C_k$ (k=1, 2, 3, 4)개의 모든 조합에 대한 앙상블 분류기의 평균 인식률을 의미한다. 결합한 분류기의 수를 증가 시킴에 따라 앙상블 분류기의 평균 인식률은 증가하였다. 또한 모든 벤치마크 데이타에 대하여 앙상블 분류기가 개별 분류기보다 우수한 성능을 보였다. 백혈병 데이타의 경우 앙상블 분류기의 최대 인식률은 97.1%이며, 개별 분류기의 최대 인식률

표 2 각 데이타 셋에 대한 분류기의 인식률(%)

	백혈병 데이타		결장암 데이타		림프종 데이타	
	MLP I	MLP II	MLP I	MLP II	MLP I	MLP II
Pearson's correlation coefficient	97.1	79.4	74.2	77.4	64.0	72.0
Spearman's correlation coefficient	82.4	79.4	58.1	64.5	60.0	88.0
Euclidean distance	91.2	61.8	67.8	77.4	56.0	72.0
Cosine coefficient	94.1	76.5	83.9	77.4	68.0	76.0
평균	91.2	74.3	71.0	74.2	62.0	77.0

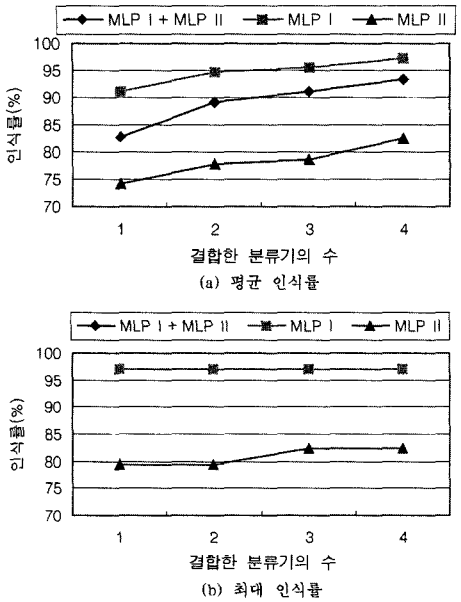


그림 6 백혈병 데이타에 대한 앙상블 분류기의 인식률(%)

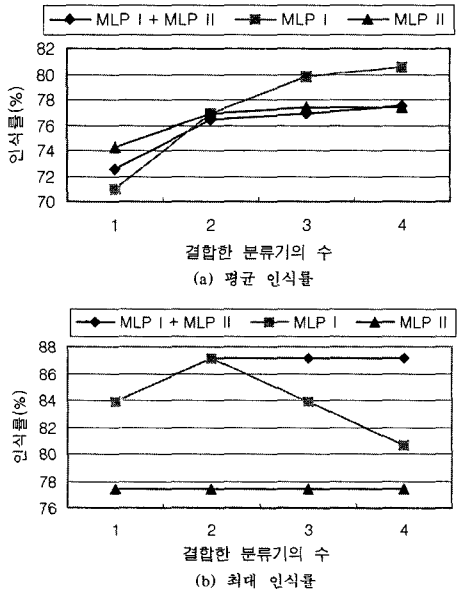


그림 7 결장암 데이타에 대한 앙상블 분류기의 인식률(%)

과 같았고, 다른 데이타의 경우 앙상블 분류기의 최대 인식률은 개별 분류기의 최대 인식률보다 높았다. 그림 6의 (b)에서 MLP I + MLP II의 성능과 MLP I의 성능이 동일하였다. 결장암 데이타의 경우 앙상블 분류기의 최대 인식률은 87.1%이고, 개별 분류기의 최대 인식

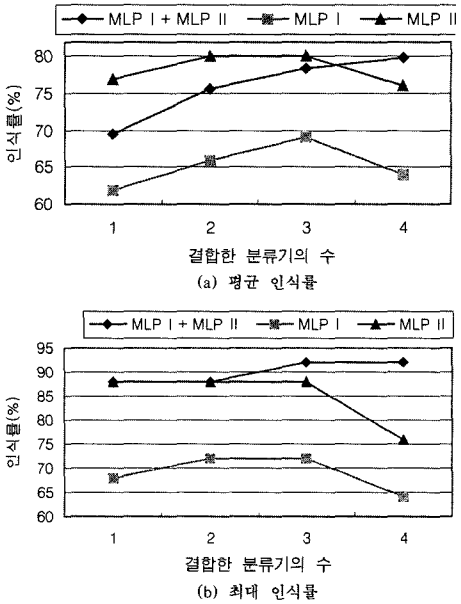


그림 8 림프종 데이타에 대한 앙상블 분류기의 인식률(%)

률은 83.9%이다. 림프종 데이타의 경우 앙상블 분류기의 최대 인식률은 92.0%이고, 개별 분류기의 최대 인식률은 88.0%이다. 표 3의 관련 연구 결과와 비교하였을 때, 제안한 방법은 관련 연구 결과에 비하여 향상된 성능을 보이거나 대등한 성능을 보였다.

MLP I, MLP II와 음의 상관관계 집합(MLP I + MLP II)의 앙상블 결과를 비교하였을 때, 평균 인식률에서는 음의 상관관계 집합의 성능이 크게 뛰어나진 않았지만, 최대 인식률에서는 음의 상관관계 집합의 성능이 가장 우수하였다. MLP I 앙상블과 MLP II 앙상블의 최대 인식률은 결합한 분류기의 수를 늘려감에 따라 떨어지는 경향을 보였지만, 음의 상관관계 집합 앙상블의 최대 인식률은 상승하였다.

본 논문은 음의 상관관계를 이용한 앙상블 분류기가 개별 분류기에 비하여 뛰어난 분류 성능을 보이고, 더욱이 음의 상관관계를 이용함으로써 기존의 앙상블 분류기보다 높은 분류 성능을 얻을 수 있음을 보여주었다.

5. 결 론

본 논문은 클래스를 구분하는 데 있어 유의한 정보를 제공하는 유전자 집단을 선택하기 위하여 음의 상관 관계를 갖는 두 개의 이상적인 벡터를 정의하고, 이와 유사한 패턴의 유전자들을 선택하여 독립적인 다수의 분류기를 생성하고 결합하는 방법을 제안하였다. 제안한 방법의 유용성을 평가하기 위하여 세 개의 공개된 벤치

표 3 암 분류에 대한 관련 연구

저자	데이타	방법		인식률(%)
		특징 추출	분류기	
Furey *et al.* [18]	백혈병	Signal to noise ratio	SVM	94.1
	결장암			90.3
Li *et al.* [26]	백혈병	Model selection with Akaike information criterion and Bayesian information criterion with logistic regression		94.1
Li *et al.* [15]	림프종	Genetic Algorithm	KNN	84.6~
	결장암			94.1~
Ben-Dor *et al.* [4]	백혈병	All genes, TNoM score	Nearest neighbor	91.6
	결장암			80.6
	백혈병		SVM with quadratic kernel	94.4
	결장암			74.2
	백혈병		AdaBoost	95.8
	결장암			72.6
Dudoit *et al.* [3]	백혈병	The ratio of between-groups to within-groups sum of squares	Nearest neighbor	95.0~
	림프종			95.0~
	백혈병		Diagonal linear discriminant analysis	95.0~
	림프종			95.0~
	백혈병		BoostCART	95.0~
	림프종			90.0~
Nguyen *et al.* [27]	백혈병	Principal component analysis	Logistic discriminant	94.2
	림프종			98.1
	결장암			87.1
	백혈병		Quadratic discriminant analysis	95.4
	림프종			97.6
	결장암			87.1
	백혈병	Partial least square	Logistic discriminant	95.9
	림프종			96.9
	결장암			93.5
	백혈병		Quadratic discriminant analysis	96.4
	림프종			97.4
	결장암			91.9

마크 암 데이타에 대해 음의 상관관계에 의한 특징 추출 방법을 적용하고, 생성된 다수의 전문화된 분류기를 Bayesian 결합 방법을 이용하여 결합하였다.

음의 상관관계 방법에 의해 선택된 유전자들은 클래스간 유전자 발현 정도의 차가 뚜렷하게 나타났으며, 이는 선택된 유전자를 기반으로 클래스를 구분하기 위한 정보를 제공한다. 이상적인 유전자 벡터 Ideal Gene A와 Ideal Gene B에 의해 선택된 특징들로 각각 학습한 MLP I과 MLP II는 데이타 셋에 따라 다른 성능을 보였다. 이러한 성능의 차이는 데이타의 특성에 기인한다. 음의 상관관계 집합에 Bayesian 결합 방법을 적용하여 앙상블한 결과 음의 상관관계 특징을 이용한 앙상블 분류기가 모든 데이타에서 가장 우수한 성능을 보였다.

세 가지 벤치마크 암 데이타 집합에 적용하여 실험한 결과, 본 논문에서 제안한 음의 상관관계 방법이 단일 분류기에 비해 우수한 성능을 보임을 알 수 있었다. 또한 음의 상관관계를 이용하지 않은 집합의 앙상블 결과와 비교할 때, 음의 상관관계 집합의 앙상블 결과가 가장 우수하였다. 본 논문에서 음의 상관관계 특징들은 서로 다른 신경망에 상호 독립적인 정보를 제공함으로써, 앙상블 분류기의 성능을 향상시킴을 확인할 수 있었다.

본 논문에서 제안한 방법은 암의 분류라는 특정한 문제에 적용되어 그 성능이 우수함을 보였지만, 일반적인 패턴인식 문제에도 적용가능하며 문제에 맞는 이상적인 특징 벡터를 정의하고 이를 이용하여 특징을 추출하고 결합함으로써 기존의 분류기 보다 개선된 성능을 보일 수 있겠다.

참 고 문 헌

[1] Harrington, C. A., Rosenow, C., and Retief, J., "Monitoring gene expression using DNA microarrays," *Curr. Opin. Microbiol.*, vol. 3, pp. 285-291, 2000.

[2] Eisen, M. B. and Brown, P. O., "DNA arrays for analysis of gene expression," *Methods Enzymbol*, vol. 303, pp. 179-205, 1999.

[3] Dudoit, S., Fridlyand, J. and Speed, T. P., "Comparison of discrimination methods for the classification of tumors using gene expression data," *Technical Report 576*, Department of Statistics, University of California, Berkeley, 2000.

[4] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, N., "Tissue classification with gene expression profiles," *Journal of Computational Biology*, vol. 7, pp. 559-584, 2000.

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

[6] Lashkari, D., Derisi, J., McCusker, J., Namath, A., Gentile, C., Hwang, S., Brown, P., and Davis, R., "Yeast microarrays for genome wide parallel genetic and gene expression analysis," *Proc. of the Natl. Acad. of Sci. USA*, vol. 94, pp. 13057-13062, 1997.

[7] Derisi, J., Iyer, V. and Brosn, P., "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, vol. 278, pp. 680-686, 1997.

[8] Eisen, M. B., Spellman, P. T., Brown, P. O. and Bostein, D., "Cluster analysis and display of genome-wide expression patterns," *Proc. of the Natl. Acad. of Sci. USA*, vol. 95, pp. 14863-14868, 1998.

[9] Shamir, R. and Sharan, R., "Algorithmic approaches to clustering gene expression data," *Current Topics in Computational Biology*. In Jiang, T., Smith, T., Xu, Y. and Zhang, M. Q. (eds), MIT press, 2001.

[10] Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J., "High density synthetic oligonucleotide arrays," *Nature Genetics*, vol. 21, pp. 20-24, 1999.

[11] Fuhrman, S., Cunningham, M. J., Wen, X., Zweiger, G., Seilhamer, J. and Somogyi, R., "The application of Shannon entropy in the identification of putative drug targets," *Biosystems*, vol. 55, pp. 5-14, 2000.

[12] Thieffry, D. and Thomas, R., "Qualitative analysis of gene networks," *Pacific Symposium on Biocomputing*, vol. 3, pp. 66-76. 1998.

[13] Friedman, N., Linial, M., Nachman, I. and Pe'er, D., "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, pp. 601-620, 2000.

[14] Arkin, A., Shen, P. and Ross, J., "A test case of correlation metric construction of a reaction pathway from measurements," *Science*, vol. 277, pp. 1275-1279, 1997.

[15] Li, L., Weinberg, C. R., Darden, T. A. and Pedersen, L. G., "Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131-1142, 2001.

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

[17] Xu, Y., Selaru, M., Yin, J., Zou, T. T., Shustova, V., Mori, Y., Sato, F., Liu, T. C., Olaru, A., Wang, S., Kimos, M. C., Perry, K., Desai, K., Greenwood, B. D., Krasna, M. J., Shibata, D., Abraham, J. M. and Meltzer, S. J., "Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer," *Cancer Research*, vol. 62, pp. 3493-3497, 2002.

[18] 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*, vol. 16, no. 10, pp. 906-914, 2000.

[19] Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M. Jr. and Haussler, D., "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. of the Natl. Acad. of Sci. USA*, vol. 97, pp. 262-267, 2000.

[20] 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., Blomfield, C. D., and Lander, E. S., "Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[21] Tamayo, P., "Interpreting patterns of gene expression with self-organizing map: Methods and application to hematopoietic differentiation," *Proc. of the National Academy of Sciences of the United States of America*, vol. 96, pp. 2907-2912, 1999.

[22] Dettling, M. and Bühlmann, P., "How to use boosting for tumor classification with gene expression data," *Technical Report*, Department of Statistics, ETH Zürich, 2002.

[23] Liu, J. and Iba, H., "Selecting informative genes with parallel genetic algorithms in tissue classification," *Genome Informatics*, vol. 12, pp. 14-23, 2001.

[24] Lippman, R. P., "An introduction to computing with neural nets," *IEEE ASSP Magazine*, 4-22, 1987.

[25] Lossos, I. S., Alizadeh, A. A., Eisen, M. B., Chan, W. C., Brown, P. O., Bostein, D., Staudt, L. M., and Levy, R., "Ongoing immunoglobulin somatic mutation in germinal center B cell-like but not in activated B cell-like diffuse large cell lymphomas," *Proc. of the Natl. Acad. of Sci. USA*, vol. 97, no. 18, pp. 10209-10213, 2000.

[26] Li, W. and Yang, Y., "How many genes are needed for a discriminant microarray data analysis," *Critical Assessment of Techniques for Microarray Data Mining Workshop*, 2000.

[27] Nguyen, D. V. and Rocke, D. M., "Tumor classification by partial least squares using microarray gene expression data," *Bioinformatics*, vol. 18, no. 1, pp. 39-50, 2002.

원 홍 희

2002년 2월 연세대학교 컴퓨터과학과 학사. 2002년 3월~현재 연세대학교 컴퓨터과학과 석사과정. 관심분야는 인공지능, 데이타마이닝, 생물정보학

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 부교수. 관심분야는 신경망, 패턴인식, 지능정보처리