유전자 알고리즘과 Feature Wrapping을 통한 마이크로어레이 데이타 중복 특징 소거법

(Removing Non-informative Features by Robust Feature Wrapping Method for Microarray Gene Expression Data)

이 재 성 † 김 대 원 ††
(Jae-sung Lee) (Dae-won Kim)

요 약 본 논문에서는 유전자 사이의 상관계수가 높은 마이크로어레이 데이타에 대하여 제안하는 알고리즘을 통해 상관계수가 낮은 유전자들의 부집합을 만들고, 이에 대해 적합 함수를 통한 평가로 기존 방법론이 가지는 한계를 극복할 수 있도록 하였다. 기존 방법론은 개별 특징의 평가를 통해 중복 특징을 제거하며, 상관계수에 대한 고려가 없어 선택된 유전자 부집합들의 상관계수가 높은 문제가 있었다. 이에 따라 제안하는 알고리즘은 특징간의 관계를 평가하는 Feature Wrapping 기법을 활용하여, 추출된 유전자 부집합에 포함된 유전자 사이의 상관관계가 낮고, 클래스 구분력이 높은 특징을 갖도록 하였다.

키워드 : 생물정보학, 계층적 클러스터링 알고리즘, 유전자 알고리즘, 특징 선별, 상관 계수

Abstract Due to the high dimensional problem, typically machine learning algorithms have relied on feature selection techniques in order to perform effective classification in microarray gene expression datasets. However, the large number of features compared to the number of samples makes the task of feature selection computationally inprohibitive and prone to errors. One of traditional feature selection approach was feature filtering; measuring one gene per one step. Then feature filtering was an univariate approach that cannot validate multivariate correlations. In this paper, we proposed a function for measuring both class separability and correlations. With this approach, we solved the problem related to feature filtering approach.

Key words : Bioinformatics, HCA, Genetic algorithm, Feature selection, Correlation coefficient

1. 서 론

클래스 예측에 있어서 전통적인 접근방법들은 수학적인 방법을 통해 클래스의 형태를 예측하고, 또한 방법론의 성능을 클래스에 대한 예측률을 통하여 평가하여 좋은 결과를 얻어내었다. 이와 같은 방법들은 통계학적인 방법을 통하여 특징 공간에서 나타나는 클래스의 형태를 확률로 표현함으로써 이후 분석에서 활용할 수 있도록 하였다[1,2].

한편 생물 유전자에 대한 관심이 최근 높아지면서 이를 데이타화하여 표현하는 방법을 연구하게 되었고, 이에 따른 결과로써 마이크로어레이(Microarray)가 개발되었다. 마이크로어레이는 수많은 유전자를 일종의 배열판에 놓고, 유전자 염기쌍 중 일부와의 반응을 측정하여 데이타로 저장하는 기술이다. 마이크로어레이 데이타에 대한 분석은 암과 같은 특정 질병과 같은 문제에 대해 직접적으로 영향을 끼치는 유전자를 찾아내는 것과 찾아낸 유전자에 대한 분석이 목표이며, 이러한 특징 유전자들을 표지 유전자라고 한다. 표지 유전자를 분석하기 위해서는 마이크로어레이 데이타를 분석하는 작업이 선행되어야 하는데, 기존의 방법론들을 사용할 경우 다음과 같은 문제가 발생하게 된다[3-6].

- 마이크로어레이 데이타에 포함된 여러 클래스를 나타내는 샘플(환자 및 정상인)의 개수가 서로 다르다. 특

† 학생회원 : 중앙대학교 컴퓨터공학과
curseor@hotmail.com
†† 정 회 원 : 중앙대학교 컴퓨터공학과 교수
dwkim@cau.ac.kr
논문접수 : 2007년 12월 18일
심사완료 : 2008년 8월 5일

정보과학회논문지: 소프트웨어 및 응용 제35권 제8호(2008.8)

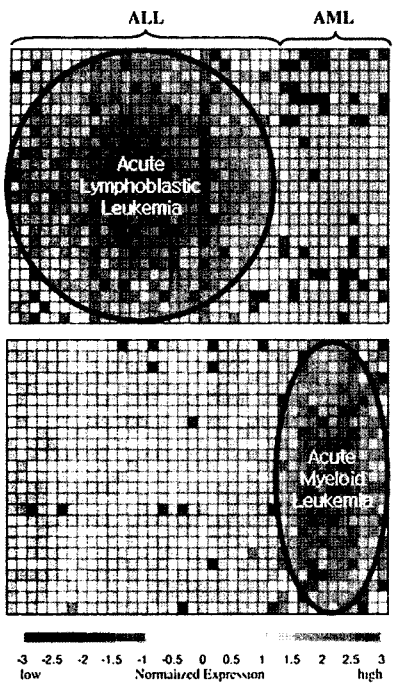


그림 1 Golub의 논문에서 사용된 마이크로어레이 데이타[7]. 세로에 해당하는 부분이 샘플(환자 및 정상인)을 나타내며, 가로는 유전자(특징)를 나타냄. 붉은 색으로 표현된 부분이 특정 질병(ALL, AML)을 나타냄

정 클래스의 경우 샘플의 개수가 매우 적어서 샘플의 개수가 많은 클래스에 대한 정확도가 전체 데이타에 대한 정확도에 영향을 미치는 등, 알고리즘의 성능을 파악하기 위한 정확도 검증 작업에 문제가 발생하게 된다.

- 마이크로어레이 데이타에서 유전자들은 샘플을 표현하는 특징(Feature)으로써 나타나게 되는데, 유전자의 개수가 많은 상태에서 실제 얻을 수 있는 샘플의 개수가 적어서 기존 방법론들을 그대로 적용하기 어렵다. 이는 기존의 방법론들이 데이타를 하나의 매트릭스로써 다루게 되는데, LDA와 같은 특징 축의 추출과 같은 방법론에서 기본적으로 사용되는 변환(Transform)을 적용하기 어렵게 된다. 또한 추출을 하더라도, 추출된 결과가 유전자에 대한 가중치로 나타나게 되어 표지 유전자에 대한 분석을 난해하게 한다.
- 마이크로어레이 데이타에서 클래스를 나타내는 샘플들이 불균형한 상태에서, 특정 클래스의 샘플이 너무 적어 기존의 방법론에서 클래스 분석에 사용하는 평균과 분산을 적용하기 어렵게 된다.
- 인간의 유전자를 예로 들었을 때, 각각의 유전자는 서로 영향을 미치고 있다. 그러나 기존의 방법론들은 데이타를 나타내는 특징들이 서로 상관관계가 없음을 가정하고 있어 마이크로어레이 데이타를 다루는데 있어 적합하지 않다[8,9]. 이와 같은 문제는 클래스를 표현하는 특징들을 선별(Feature selection)하는 과정에서 많이 나타나게 되는데, 선별된 특징들이 클래스 예측력

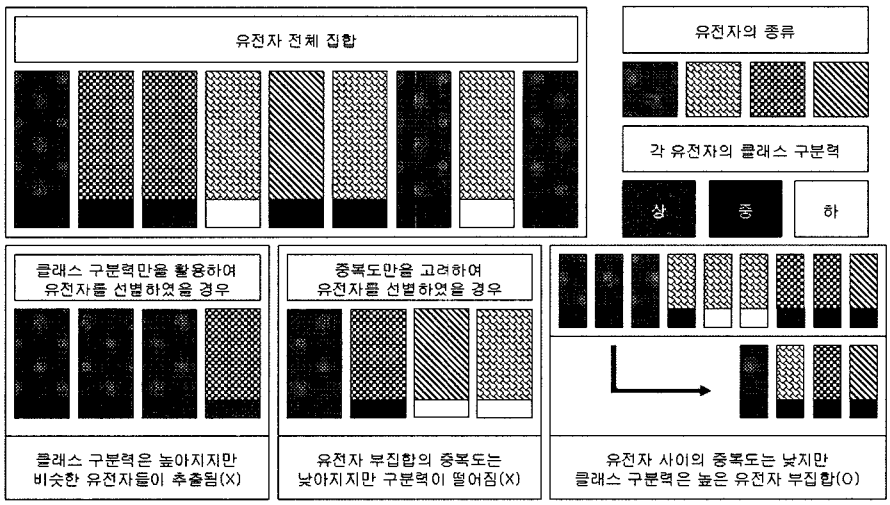


그림 2 유전자 중복도가 높을 경우 발생하는 문제. 클래스 구분력을 통해 유전자를 추출할 경우 비슷한 기능을 담당하고 있는 유전자들이 추출될 가능성이 높음. 이는 유전자들이 기능에 따라 군집화가 가능하며 동일한 군에 포함된 유전자들은 특정 질병에 대해 비슷한 실험값을 보이기 때문임. 또한 순수하게 클래스 구분력에 의해 추출된 유전자들은 그 기능이 이미 밝혀진 경우가 많아 새로운 표지 유전자를 알아내기 위한 분석에는 적합하지 않음

은 뛰어나지만 서로 중복도(Correlation)가 높은 문제를 가지고 있다. 이는 그림 2와 같이 비슷한 기능을 가지는 표지 유전자를 추출할 가능성이 높아져 추후 이어질 표지 유전자 분석에 좋지 않은 영향을 주게 된다.

이에 따라 마이크로어레이 데이타를 다루기 위한 여러 방법론들이 개발되었으나 기존의 방법론을 확장하여 사용하고 있어 실제 마이크로어레이 데이타의 특징을 고려하지 못하는 등 많은 문제가 있다.

2. 선행 연구

유전자 데이타를 매트릭스로써 다루는 접근 방법에 있어서, 마이크로어레이 데이타와 같이 다수의 유전자가 특징으로 나타날 경우 변환과 같은 접근 방법을 적용하기 어렵다. 따라서 각각의 특징에 대해 측정을 한 뒤, 높은 점수를 얻은 특징을 선별하는 방법이 사용되어 왔다[7]. 대표적인 개별 특징에 대한 평가 방법(Feature filtering)으로써, 클래스 구분력을 평가하는 방법으로 사용할 수 있는 t-test와 개별 특징들의 중복도를 평가하는 방법으로 사용되는 상관 계수(Correlation Coefficient)가 있다.

2.1 t-test

t-test는 단일 특징 f_k에서 클래스의 평균과 분산을 통해 클래스의 구분력을 평가하는 방법이며, 그 형태는 식 (1)과 같다[10].

$$t(f_k) = \frac{\bar{c_1} - \bar{c_2}}{s_{\bar{c_1}-\bar{c_2}}}$$
$$where\ s_{\bar{c_1}-\bar{c_2}} = \sqrt{\frac{s_{c_1}^2 + s_{c_2}^2}{n}} \quad (1)$$

이 때, $\bar{c_i}$는 특징 f_k에 포함된 i번째 클래스 c_i의 평균을 의미하고, s_i는 클래스 c_i의 편차를 의미하며, $s_{\bar{c_1}-\bar{c_2}}$는 클래스를 표현하고 있는 특징 f_k 전체의 편차를 의미한다. 식 (1)에서 두 클래스의 평균 사이의 거리가 멀고, 분산이 작을수록 $t(f_k)$값이 커질 수 있는데 값이 클수록 클래스 구분력이 높은 특징이며 이는 그림 3과 같이 표현할 수 있다.

t-test는 평균과 분산이 편중되지 않을 것을 가정하고 있는데, 이러한 특성은 마이크로어레이 데이타에서 기대하기 어렵다. 마이크로어레이 데이타를 이루고 있는 샘플의 개수가 부족하기 때문인데, 이에 따라 데이타에 포함되어 있는 일부의 노이즈에 의해 그림 4와 같이 평균이 크게 영향을 받게 된다. 마이크로어레이 데이타의 경우, 일부의 유전자가 측정 장비의 오류와 마이크로어레이의 표면에 생긴 상처 및 오염 등에 의한 원인으로, 측

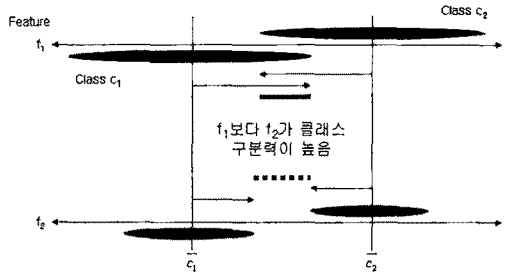


그림 3 t-test 비교. x축은 각각 f_1과 f_2를 의미하며, 아래쪽 타원은 클래스 c_1을, 위쪽 타원은 c_2를 의미함. 각 타원은 특징 f_1과 f_2에서 나타나는 클래스의 분포를 나타냄. f_1과 f_2에서 나타나는 클래스의 중심은 같으나 분산이 서로 달라서 f_2에서 클래스 사이의 공백이 나타나므로, f_1보다 f_2가 클래스 구분력이 더 높음

Sample	f_k	Class
:	:	:
Normal	333	1
Normal	170	1
Normal	100	1
Normal	1545	1
Normal	45	1
Normal	112	1
Normal	63	1
Patient	176	-1
Patient	74	-1
Patient	237	-1
Patient	-68	-1
:	:	:

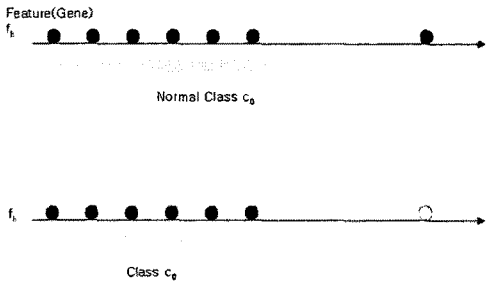


그림 4 노이즈에 의한 클래스 형태 변화

정된 값이 일반적인 값들의 범주를 넘게 된다. 이에 따라 t-test를 통한 클래스 구분력(Separability)의 판단에 영향을 주게 되어, 일부의 노이즈에 의해 클래스 구분력이 좋은 특징임에도 불구하고 t값이 낮게 평가된다.

또한, 개별 특징에 대한 평가를 하여 선별을 하기 때문에 실제 t-test를 통해 선별된 특징들의 집합은 서로에 대해 높은 상관관계를 보이게 된다. 이는 t-test가 실제 특징을 평가할 때, 특징들은 서로 상관관계가 없음을 가정하고 있기 때문이며, 이러한 특성 때문에 유전자 사이의 상관계수가 높은 마이크로어레이 데이타에서 적용하기 어렵게 된다.

2.2 상관 계수

개별 특징에 대한 중복도 평가는 상관계수(Correlation Coefficient)를 통해 측정할 수 있다[11]. 상관계수는 대상이 되는 두 특징에 대한 공분산을 활용하여 서로의 값에 대한 변화를 측정하는 방법이다. 식 (2)의 f_m과 f_k에서 각각 유전자 m과 유전자 k를 나타내는 값의 집합을 의미한다. 또한, $cov(f_m,f_k)$는 f_m과 f_k의 공분산을 의미하며, $var(f_m)$는 f_m에서 나타나는 값들의 분산을 의미한다.

$$cc(f_m,f_k)=\frac{|cov(f_m,f_k)|}{\sqrt{var(f_m)\cdot var(f_k)}} \quad (2)$$

측정된 값은 0에서 1사이의 값을 가지게 되는데, 측정된 값이 0에 수렴할수록 두 특징은 서로 중복되지 않았음을 의미하고, 1에 수렴할수록 중복이며, 특징에 포함된 값의 단위에 무관하다. 상관계수는 주어진 특징들에 대해서 중복도를 측정하는 강력한 방법을 제공하지만, 클래스 예측에 큰 영향을 미치는 클래스 구분력을 고려하지 않은 방법이므로, 실제 클래스 예측에 사용하기에 적합하지 않아, 상관 계수만으로는 클래스 예측 문제에 활용하기 어렵다는 단점이 있다[12].

2.3 Feature Wrapper

개별 특징에 대한 평가에 의한 한계를 극복하고, 전체 집합을 평가하기 위해 필요한 시간을 줄이기 위하여 Feature Wrapper와 같은 방법들이 제시되었다[13]. Feature Wrapper는 주어진 훈련 데이타와 클래스 예측기를 활용하여 생성하고자 하는 특징 부집합의 클래스 예측률을 높이기 위한 방법이다.

Feature Wrapper는 크게 특징 공간을 탐색하는 알고리즘과 각 단계마다 생성된 특징 부집합을 평가하기 위한 클래스 예측기로 이루어지는데, 특징 공간을 탐색하는 알고리즘으로는 크게 SFS(Sequential Forward Search)와 SFFS(Sequential Forward Floating Search) 등이 있다. 이 두 알고리즘의 큰 차이점은 이미 생성된 부집합에 포함된 특징이 제거되는 등의 변경 과정 유무에 있다. SFS는 다음과 같은 과정을 통해 특징 부집합을 선택한다.

- 최초 단일 특징에 대해 클래스 예측률을 평가한다. 클래스 예측률이 가장 높은 특징을 선택하는데, 선택된 특징을 f_1이라고 한다.
- 두 번째 특징을 선택하는데, 이 때 f_1을 포함하면서 새로운 특징을 추가하였을 때 클래스 예측률이 최적인 특징 부집합을 탐색하여, 새로운 특징을 추가한다.
- 위의 과정을 반복한다.

SFS에 의한 특징 부집합 선택 방법은 전체 공간을 탐색하는 알고리즘이나 GA에 비해 빠른 탐색 시간과 높은 클래스 예측률이 장점이지만, 실제 생성되는 특징 부집합이 SFS의 각 단계에서 선택되는 특징들의 순서에 크게 영향을 받는 단점이 있다.

3. 제안하는 알고리즘

개별 특징에 대한 평가를 기반하는 방법(Feature Filtering)은 근본적으로 각 특징들이 서로에 대해 상관관계가 없음을 가정하고 있어 마이크로어레이 데이타에 적합하지 않다. 그러나 개별 특징에 대한 가정을 배제할 경우, 특징들로 이루어진 모든 집합에 대해 고려를 해야 하기 때문에 실세계에서의 문제를 해결하기에는 적합하지 않으며, 데이타에 포함된 클래스의 개수가 많아질수록 해결하기 어려워진다. 따라서 본 논문에서는 전체 조합에 대해 평가하는 것보다는 유전자 알고리즘(Genetic algorithm, 이하 GA)을 통한 Feature Wrapping적인 접근을 사용하였다.

3.1 GA의 인자 설정

GA는 유전자 풀(Pool)을 구성하고, 유전자 풀에서 임의의 유전자를 선택한 뒤 다음 세대의 유전자를 생성하여, 기존 유전자 풀을 평가값에 의하여 갱신하는 방법을 주로 사용한다. 본 논문에서 활용한 GA 알고리즘은 유전자 풀을 최초에 무작위로 선택한 다음, 이의 절반을 그림 5와 같이 제안한 알고리즘을 통해 평가하고, 나머지 절반을 K-nn(K-nearest neighbor)의 정확도에 의해 평가를 하도록 하였으며, LOOCV(Leave-one-out cross validation) 방법으로 훈련 데이타와 시험 데이타를 생성하도록 하였다.

또한, 다음 세대에 만들어지는 유전자는 다음과 같은 방법으로 생성하도록 하였다.

- 유전자 풀에서 다음 세대의 생성에 참여할 2개의 유전자를 선택한 뒤, 각 유전자의 우선순위를 결정하고, 이에 따라 어떤 부모의 유전자를 복사할 것인지에 대한 복사 확률을 결정한다. 만약 부모 유전자를 각각 G_1, G_2라고 하고, G_1은 제안한 알고리즘에 의해 평가된 유전자 이며, G_2는 K-nn에 의해 평가된 유전자라

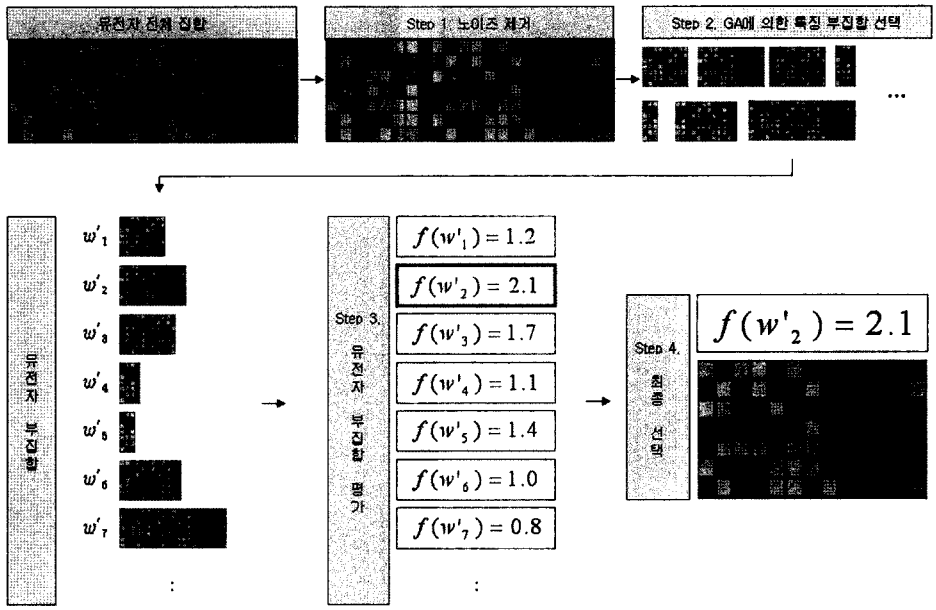


그림 5 제안된 알고리즘의 구조. 전체 유전자로 이루어진 마이크로어레이 데이타에 대해 노이즈를 제거하고, GA에 의해 유전자 부집합을 선택한 뒤, 평가함수를 통해 평가값을 얻어 유전자 부집합에 대해 평가함

하면 G_1과 G_2의 평가값에 의한 우선순위를 결정할 수 있다. 이 때, G_1의 순위가 3번째이며, G_2의 순위가 2번째라면 다음 세대의 유전자는 40%확률로 G_1의 유전자를 복사하며, G_2의 유전자는 60%의 확률로 복사한다.

- GA에서 활용되는 모든 유전자들은 1과 0으로 이루어진 문자열이므로, 부모 유전자의 우선순위에 따라 새로운 문자열을 순차적으로 복사 확률에 따라 생성한다. 생성된 문자열에 대해 80%의 확률로 돌연변이(Mutation)를 생성하는데 해당 문자열의 무작위 위치의 문자에 대해 1을 0으로, 0을 1로 치환한다. 이 과정은 80%의 확률을 통과할 때까지 반복된다.
- 생성된 유전자는 제안된 알고리즘과 K-nn에 의해 동시에 평가된다. 이 때, 유전자 풀에서 제안된 알고리즘에 의해 평가된 유전자들의 집합을 P_1이라고 하고, K-nn에 의해 평가된 유전자들의 집합을 P_2라고 할 때, 해당 유전자에 대해 제안된 알고리즘의 평가값이 P_1의 최하위 평가값을 가진 유전자 보다 좋다면 P_1에 반영된다. 만약, P_1의 최하위 평가값을 가진 유전자보다 평가값이 좋지 않다면 P_2의 평가값을 비교하여, P_2의 최하위 평가값을 가지는 유전자 보다 평가값이 좋을 경우 P_2에 갱신된다.

본 논문에서 활용한 GA에서는 유전자 풀에 포함된 유전자의 개수를 100개로, 각 세대당 50개의 새로운 유전자를 생성하였으며, 1000번의 세대를 거치도록 하였다. 최종 세대에서 만들어진 유전자 100개에 대해서 K-nn에 의한 평가를 거쳐 1개의 유전자를 결정하도록 하였다.

3.2 노이즈 제거

데이타에 포함된 샘플은 클래스 구분력을 판단하는데 결정적인 영향을 주게 되는데, 이는 샘플의 개수가 적은 마이크로어레이 데이타에서 그 영향력이 훨씬 커진다. 또한 클래스 구분력은 질병과 연관된 표지 유전자를 추출하는데 중요한 척도가 되기 때문에 데이타에서 유전자를 의미하는 각각의 특징들이 가지고 있는 클래스 구분력은 곧 표지 유전자일 가능성을 판단하는 척도가 된다. 그러나 먼지나 계측장비의 오류 등으로 인하여 실험값이 정상의 범주를 크게 벗어나는 문제가 있기 때문에 클래스 구분력의 추정에 있어서 좋지 않은 영향을 주게 되며, 이는 곧 노이즈 데이타에 의해 표지 유전자일 가능성이 높은 유전자가 누락될 수 있음을 의미한다.

표 1은 Golub의 논문에서 사용된 백혈병(Leukemia) 데이타에서 임의의 특징을 보인 것이다. 앞서 언급하였듯이 t-test와 같은 클래스 구분력을 판단하는 방법들에 있어서, 각 특징에서 나타나는 음영으로 표시된 값이 클래스의 형태를 결정함에 있어서 악영향을 끼치고 있다. 실제 백혈병 데이타에서는 이러한 값들이 전체 특징에 걸쳐 나타나는데, 이러한 값들을 제거함으로써 클래스 구분력을 판단하는데 미치는 영향을 최소화할 수 있으며, 이렇게 클래스의 형태를 판단하는데 큰 영향을 미치는 소

표 1 Leukemia data[3]

Sample	f_1	···	f_k	···	f_n	Class Label
:	:	:	:	:	:	:
Normal	-161		-192		16	1
Normal	-48		187		-73	1
Normal	-176	···	13868	···	-60	1
Patient	-214		-126		-37	-1
Patient	-139		2267		-14	-1
:	:	:	:	:	:	:

수의 노이즈들은 HCA(Hierarchical Clustering Algorithm, 이하 HCA)를 통해 제거하는 것이 가능하다. 이러한 노이즈 값들은 특징상에서 정상 반응한 결과를 측정한 값들에 비해 크게 벗어난 값으로 측정되며, 따라서 각 값들의 유사도를 관찰함으로써 노이즈를 결정 할 수 있다. 특징 k 상에서 나타난 n개의 샘플을 $v_i(i=1,\ldots,n)$라 할 때, 특징 k상에서 나타나는 값 v_j의 비유사도(Dissimilarity)는 식 (3)으로 평가할 수 있다[14].

$$c(v_j)=\sum_{i=1}^{n}(v_i-v_j)^2 \quad (3)$$

그림 6과 같이 식 (3)을 통해 얻은 각 값에 대한 비유사도를 이용하여, HCA를 통해 클러스터 트리를 생성하는 것이 가능하다. 이 때, 각각의 클러스터들은 유사도에 의해 생성되었으므로, 유사도가 낮은 값들로 이루어진 클러스터가 생성될 수 있다. 클러스터 트리에서 높이는 각 유사도간의 거리를 의미하므로, 유사도간의 거리를 통해 최적 클러스터 개수를 추론할 수 있으며, 이를 통해 그림 7과 같이 노이즈 클러스터를 구분하는 것이 가능하다. 즉, 최적 클러스터 개수 c^*을 추론한 뒤, 클러스터 중 클러스터를 이루는 데이타의 개수를 통해 노이즈 클러스터를 판단하는 것이 가능하며, 노이즈 클러스터를 제거함으로써 목적을 달성할 수 있다[15].

차후 평가 함수에서 클래스들의 중첩 공간을 비율로

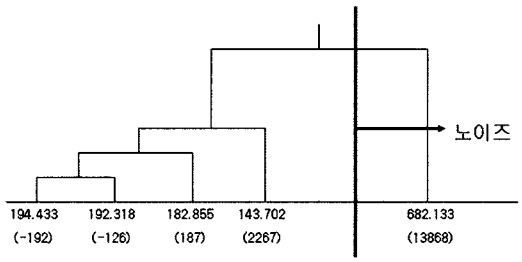


그림 7 Hierarchical tree에 의한 노이즈 결정, 작성된 Tree에서 y축은 각 원소의 비유사도에 대한 거리를 의미하며, 거리에 따라 노이즈를 판별하는 것이 가능함

측정하게 되는데, 각 클래스의 평균과 분산을 측정할 때, 노이즈로 결정된 값들을 제외하고 측정함으로써 앞서 언급한 노이즈에 의한 클래스 위치 추정의 편중 현상을 회피할 수 있도록 한다. 본 논문에서는 7129개의 유전자로 이루어진 백혈병 데이타의 전체 72개 샘플 중에서 정상 샘플 47개, 비정상 샘플 25개에 대해 노이즈 클러스터와 노이즈가 아닌 클러스터가 생성될 것을 가정하여 c^*을 2로 가정하고 실험하였으며, 각 클러스터에 포함된 샘플의 개수가 전체 샘플의 5% 미만일 경우 노이즈로 정의하였다. 그림 8은 Golub의 백혈병 데이타[3]에 대해서 c^*에 의해 노이즈로 결정된 값과 전체 샘플

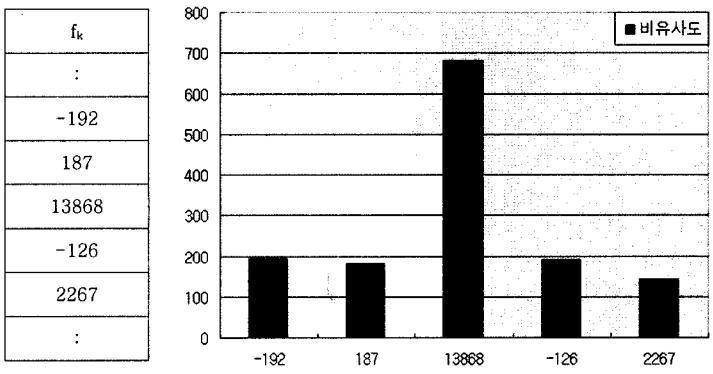


그림 6 식 (3)을 통해 얻은 비유사도 히스토그램

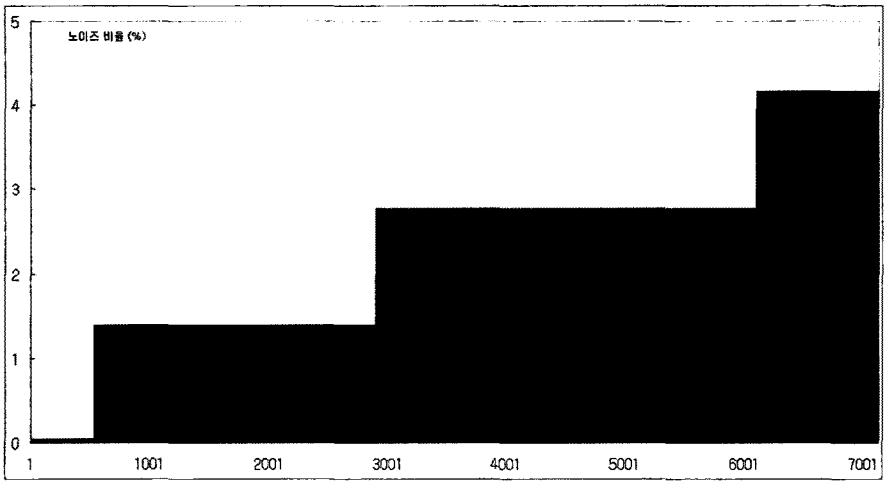


그림 8 백혈병 데이타[3]에서 c^*에 의한 노이즈 판단과 각 유전자에서 결정된 노이즈의 비율, x축은 유전자에 포함된 노이즈의 개수에 따라 정렬하여 표시한 것임. 예를 들어, x축의 5001번째 값은 노이즈의 개수에 따라 정렬된 결과 5001번째의 유전자가 가지고 있는 노이즈의 비율을 의미함. 즉, Golub의 데이타에는 2500개 정도의 유전자가 1.5% 정도의 노이즈를 가지고 있고, 1000개 미만의 유전자가 4.5% 정도의 노이즈를 가지고 있음을 의미함. 또한, 노이즈가 포함된 샘플은 이후 각 특징을 평가하는 과정에서 사용되지 않음

의 개수 비율을 나타낸 것이다.

3.3 GA를 활용한 특징 부집합 선택

본 논문에서는 그림 9와 같이 GA를 통해 유전자 부집합을 선별하고, 선별된 유전자 부집합에 포함된 유전자들의 상관관계가 낮아졌는지를 상관계수를 통해 평가하도록 하였다. GA를 통해 선별된 유전자 부집합은 부집합에 포함된 각 유전자들에 대해 다른 유전자들과의 상관계수 및 각 유전자에서의 클래스 구분력을 평가함으로써 부집합에 포함된 유전자들의 상관계수는 낮고, 클래스 구분력은 높은 마이크로어레이 데이타 부집합을 얻을 수 있다.

3.4 평가 함수

GA를 통해 선별된 유전자 부집합 w에 대한 평가 함수 $e(w)$는 식 (4)와 같다. 이 때, w는 선별된 유전자 부집합이며, w'는 노이즈 제거를 거친 유전자 부집합이다. 분자에 해당하는 $D(w')$ 함수는 유전자 부집합에서

GA에 의한 유전자 부집합 선택

f_1	f_2	f_3	f_4
1	0	1	1

유전자 전체 집합(w')

Sample	f_1	f_2	f_3	f_4	Class Label
Patient	23	NaN	9	51	-1
Patient	4	45	53	42	-1
Normal	56	95	159	14	1
Patient	123	53	46	NaN	-1
Normal	18	87	87	47	1
Normal	NaN	73	19	86	1
Patient	552	159	35	53	-1

→

유전자 부집합(w'_1)

Sample	f_1	f_3	f_4	Class Label
Patient	23	9	51	-1
Patient	4	53	42	-1
Normal	56	159	14	1
Patient	123	46	NaN	-1
Normal	18	87	47	1
Normal	NaN	19	86	1
Patient	552	35	53	-1

그림 9 GA에 의한 유전자 부집합 선택, GA에 의해서 f_1, f_3, f_4가 선택되었을 때, GA에 의해 생성된 유전자 부집합은 Cancer Data의 우측에 위치한 표와 같음. 이 때, NaN은 노이즈 제거에 의해서 제거된 값을 의미하며, 이후 유전자 부집합의 평가에서 제외됨

개별 특징들과 다른 특징들과의 비교를 통해 현재 부집합을 이루고 있는 유전자들의 상관관계를 측정하는 함수이며, $K(w')$ 함수는 유전자 부집합의 특징(유전자) 공간에 나타나는 클래스들의 중첩 비율을 측정한다.

$$e(w) = \frac{D(w')}{K(w')} \quad (4)$$

t-test를 통해서 특징을 추출할 경우 클래스 구분력이 높은 유전자 부집합을 찾는 것이 가능하지만, 유전자 사이의 중복도가 높아, 앞서 언급한 표지 유전자에 대한 분석에 좋지 않다. 그러나 중복도를 고려하여 상관계수를 통해 특징을 추출하게 되면, 클래스 구분력이 낮아지는 것을 회피할 수 없다.

이 때, 그림 10과 같이 평가 함수를 통해 클래스 구분력이 높은 유전자 부집합이라도 중복도가 높으면 평가값을 낮게 할당하여 최종 선택에서 제외하도록 하였다. 다시 말해, 클래스 구분력이 높더라도 유전자 중복도가 높은 유전자 부집합에 대한 평가가 낮아지도록 평가 함수를 분수 형태로 작성하였다.

3.4.1 특징 비중복도 계산 : $D(w')$

유전자 부집합을 이루는 특징 사이의 비중복도는 식 (5)로 평가할 수 있다. 이 때, f_i는 유전자 부집합을 이루는 특징을 의미한다.

$D(w')$ 함수는 유전자 부집합을 이루는 특징 중 임의의 특징 f'_i의 비중복도에 대한 평균을 의미하며, 유전자 부집합 w에 i개의 특징이 있을 때, 특징 각각의 비중복도를 $d(f'_i)$ 함수를 통해 구하고 이에 대한 평균을 구한다. 상관계수에 의한 비중복도는 GA에 의해 선택된 유전자 부집합에 의해 상대적으로 결정되는데 현재 유전자들의 집합 안에서 각각의 유전자들의 중복도가 어떠한지 측정하여 계수로 표현하는 것이 목적이다.

$$D(w') = \frac{1}{n}\sum_{i=1}^{n} d(f'_i) \quad (5)$$

식 (5)에서 사용되는 각 유전자의 비중복도는 식 (6)을 통해 구할 수 있다. 식 (6)은 임의의 특징 f'_i가 나머지 특징들과의 비중복도가 얼마나 되는지 평가하기 위하여 작성되었다. 상관계수를 통해 얻은 값은 0에서 1 사이의 값을 나타내는데, 이 때 1에 가까워질수록 중복

유전자 전체 집합(w)

	f_1	f_2	f_3	f_4	Class Label
Patient	23	[illegible]	9	51	-1
Patient	4	45	53	42	-1
Normal	56	95	159	14	1
Patient	123	53	46	[illegible]	-1
Normal	18	87	87	47	1
Normal	[illegible]	73	19	86	1
Patient	552	159	35	53	-1

↓ ↓

GA에 의해 선택된 유전자 부집합

유전자 부집합(w'_1)

	f_1	f_2	f_3	Class Label
Patient	23	[illegible]	9	-1
Patient	4	45	53	-1
Normal	56	95	159	1
Patient	123	53	46	-1
Normal	18	87	87	1
Normal	[illegible]	73	19	1
Patient	552	159	35	-1

유전자 부집합(w'_2)

	f_1	f_2	f_4	Class Label
Patient	23	[illegible]	51	-1
Patient	4	45	42	-1
Normal	56	95	14	1
Patient	123	53	[illegible]	-1
Normal	18	87	47	1
Normal	[illegible]	73	86	1
Patient	552	159	53	-1

부집합 평가	f_1, f_2, f_4	f_1, f_2, f_3
	$D(w')$ 0.8, $K(w')$ 0.8	$D(w')$ 0.2, $K(w')$ 0.9
최종 선택	O	X

그림 10 유전자 부집합의 최종 선택. 클래스 구분력을 평가하는 $K(w')$ 값이 높게 평가되더라도, 유전자 비중복도를 평가하는 $D(w')$ 값이 낮게 평가되면 최종 선택에서 제외됨

도가 높으며, 0에 가까워질수록 중복도가 낮으므로, 1과의 차이를 측정하는 것으로 비중복도를 측정하는 것이 가능하다. 식 (6)에서 $\alpha(f'_i,f'_j)$는 f'_i와 f'_j의 상관계수를 구하는 수식이며, 노이즈가 포함된 샘플은 사용하지 않는다.

$$d(f_i)=\frac{1}{n-1}\sum_{j=1}^{n}(1-|\alpha(f'_i,f'_j)|) \qquad (6)$$

그림 11은 상관계수를 통하여 유전자 부집합 w에 대한 비중복도를 측정하는 방법에 대해 보이고 있다.

3.4.2 클래스 중첩도 계산 : $K(w')$

클래스의 중첩도는 식 (7)을 통해 평가할 수 있다. 이 때, 유전자 부집합에 포함되어 있는 중첩도를 측정할 경우 측정 함수를 설계하는데 있어서 몇 가지 중요한 요소가 있다.

- 일반적으로 클래스 중첩도는 클래스를 표현하는 특징의 개수가 증가할수록 높아진다. 이는 특징의 개수가 증가할수록 클래스 구분력이 낮아지도록 하는 특징이 점점 증가하기 때문이며, 이에 따라 특징의 추가에 민감하지 않아야 한다.
- 마이크로어레이 데이타를 이루고 있는 유전자들은 서로 연관성이 있지만, 그 단위(Scale)가 서로 일치하지 않는 경우가 있다. 이는 중첩 공간이 실제로는 특징상에서 나타나는 클래스의 분산과 비교하였을 때 크지 않음에도 불구하고, 측정되는 값이 단위에 의해 크게 측정될 수 있음을 의미한다. 따라서 평가 함수는 유전자를 표현하는 값의 단위에 민감하지 않아야 한다.
- 클래스 구분력에 있어서 실제 클래스 자체의 구분력은 개별 특징 중에 단 하나에서 클래스가 서로 떨어져 있을 경우 다른 특징들에 관계없이 클래스 구분력이 좋아지게 된다. 그러나 단일 특징을 찾는 방법은 이후의 표지 유전자를 찾기 위한 분석을 위해서 좋은 접근 방법이 아니며, 다양한 특징들을 선별하였더라도 개별 특징 중에서 클래스 구분력이 좋은 경우 이를 잘 평가할 수 있어야 한다.

식 (7)은 노이즈가 제거된 유전자 부집합 w'에서 클래스 중첩 공간을 측정하는 수식이다. 유전자 부집합 w'는 GA가 선택한 유전자에 의해서 다양한 조합으로 나타날 수 있는데, 앞서 언급하였듯이 클래스 중첩도는 클래스를 표현하는 특징의 개수가 증가할수록 클래스 구분력이 낮아질 가능성이 높아진다. 이를 고려하여 유전자 부집합을 이루는 특징이 n개라고 하였을 때, 그림 12와 같이 n제곱근을 통해 극복할 수 있도록 하였다.

$$K(w')=\sqrt[n]{\prod_{i=1}^{n}k(f'_i)} \qquad (7)$$

그림 13은 제곱근에 의한 중첩 비율 평가와 평균에 의한 중첩 비율 평가를 보이고 있다. 클래스 구분력에 있어서 다양한 특징들이 데이타에 포함되어 있더라도 단일 특징에서 클래스 구분이 이루어진다면 곧 다양한 특징에서도 클래스가 명확하게 구분이 된다. 즉, 클래스 구분력만을 위해서 특징 선별을 한다면 단일 특징에 대한 평가를 통해 선별하는 것이 적합할 수 있으나, 마이크로어레이 데이타에서는 다양한 유전자 중에서 표지 유전자를 찾아내는 것이 목적이므로, 클래스 구분이 잘 되는 특징에 대해 가중치를 두어 평가를 하고, 표지 유전자가 존재할 수 있는 다른 특징들은 남겨두는 것이 적합하다. 그림 13에서 보는 바와 같이 특징의 개수에

유전자 부집합(w')					
	f_1	f_2	f_3	f_4	Class Label
Patient					-1
Patient	4	45	53	42	-1
Normal	56	95	159	14	1
Patient					-1
Normal	18	87	87	47	1
Normal					1
Patient	552	159	35	53	-1

$d(f'_1)$	$d(f'_2)$	$d(f'_3)$	$d(f'_4)$
0.37	0.53	0.45	0.44

1-|상관계수 매트릭스|					$d(f_i)$
	f_1	f_2	f_3	f_4	
f_1	0	0.09	0.48	0.53	0.37
f_2	0.09	0	0.79	0.72	0.53
f_3	0.48	0.79	0	0.07	0.45
f_4	0.53	0.72	0.07	0	0.44

유전자 부집합 w'의 비중복도 $D(w')$

0.4475

그림 11 유전자 전체 집합에 대해 비중복도를 구하는 과정의 예제. $d(f'_i)$를 통해 개별 특징의 비중복도를 구하고, 개별 특징의 비중복도에 해단 평균으로 w'의 비중복도를 구함. 이 때, 노이즈가 포함된 샘플은 상관계수를 구하는 과정에서 사용되지 않음

전체 유전자 집합		f_1	f_2	f_3	f_4	f_5	f_6
	중첩 공간 비율	0.9	0.1	0.8	0.7	0.1	0.6

GA에 의해 선택된 유전자 부집합(f_1, f_3, f_4)			
유전자 부집합(w_1')	f_1	f_3	f_4
중첩 공간 비율	0.9	0.8	0.7
$K(w_1') = \sqrt[3]{0.9 \cdot 0.8 \cdot 0.7} \approx 0.795$			

GA에 의해 선택된 유전자 부집합(f_2, f_4, f_5, f_6)				
유전자 부집합(w_2')	f_2	f_4	f_5	f_6
중첩 공간 비율	0.1	0.7	0.1	0.6
$K(w_2') = \sqrt[4]{0.1 \cdot 0.7 \cdot 0.1 \cdot 0.6} \approx 0.254$				

그림 12 $K(w')$에 의한 유전자 부집합의 중첩 비율 측정. 유전자의 개수(특징 공간의 차수)에 관계없이 유전자 부집합의 중첩 비율을 측정하는 것이 가능하며, 유전자 부집합 w'_2가 유전자 부집합 w'_1보다 중첩 비율이 낮은 것을 알 수 있음. 또한, 중첩 공간 추정시 노이즈가 포함된 샘플은 사용하지 않음

전체 유전자 집합		f_1	f_2	f_3	f_4	f_5	f_6
	중첩 공간 비율	0.9	0.1	0.8	0.7	0.1	0.6

GA에 의해 선택된 유전자 부집합(f_1, f_3, f_4)	
제곱근	$K(w'_1) = \sqrt[3]{0.9 \cdot 0.8 \cdot 0.7} \approx 0.795$
평균	$\frac{1}{3}(0.9+0.8+0.7) = 0.8$

GA에 의해 선택된 유전자 부집합(f_2, f_4, f_5, f_6)	
제곱근	$K(w'_2) = \sqrt[4]{0.1 \cdot 0.7 \cdot 0.1 \cdot 0.6} \approx 0.254$
평균	$\frac{1}{4}(0.1+0.7+0.1+0.6) = 0.375$

그림 13 제곱근과 평균에 의한 클래스 구분력 평가 비교. 유전자 부집합을 이루는 중첩 비율이 상대적으로 클 경우 평균을 통한 평가와 큰 차이가 없지만, 중첩 비율이 낮은 특징이 포함된 유전자 부집합에 대해서 가중치를 두어 평가하는 것을 볼 수 있음

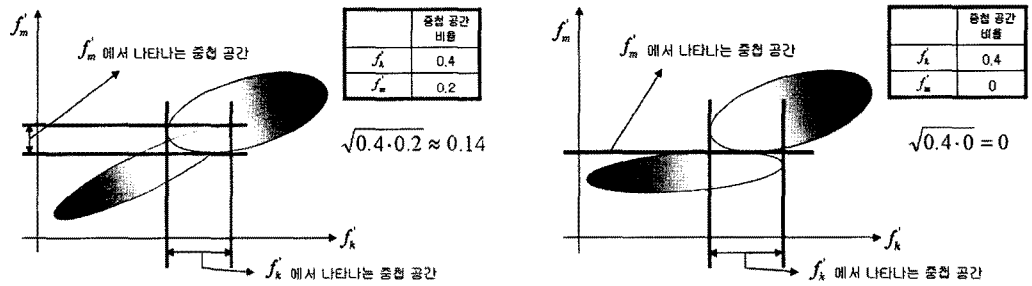


그림 14 특이점 문제(Singularity problem). 식 (9)에 의해 중첩공간을 측정하면 중첩공간의 비율이 0이 되어 식 (7)이 성립하지 못하므로, 식 (8)을 통해 계수 $k(f'_i)$를 구하여 보정함. (a) 특징 f_m과 f_k 모두 중첩 공간이 있을 경우 식 (9)를 통해 중첩 공간을 계측할 수 있음. (b) 특징 f_m에서 중첩 공간이 0이 될 경우, 식 (9)가 0이 되어 f_k에 관계없이 유전자 부집합의 중첩공간이 0으로 측정됨

상관없이 선택된 유전자 부집합의 중첩 비율을 측정할 수 있으며, 중첩 비율이 낮은 유전자 부집합에 대해 더 낮은 값으로 측정되는 것을 알 수 있다.

식 (8)은 식 (9)에서 구한 중첩 공간의 비율에서 특이점 문제를 회피하기 위해 작성되었으며, f'_i는 노이즈가 제거된 유전자 부집합의 i번째 특징을 의미한다. 유전자 부집합의 중첩 공간을 측정할 때, 실제 비율을 통해 부피를 측정하는 경우 실제 중첩 공간이 없거나, 정확하게 0일 경우 그림 14와 같이 특이점 문제(Singularity Problem)가 발생할 수 있다. 이러한 문제를 회피하기 위해 식 (9)에서 $O(f'_i)$의 값이 0으로 측정되더라도 식 (8)을 통해 보정하였으며, 클래스 중심의 거리가 멀어질

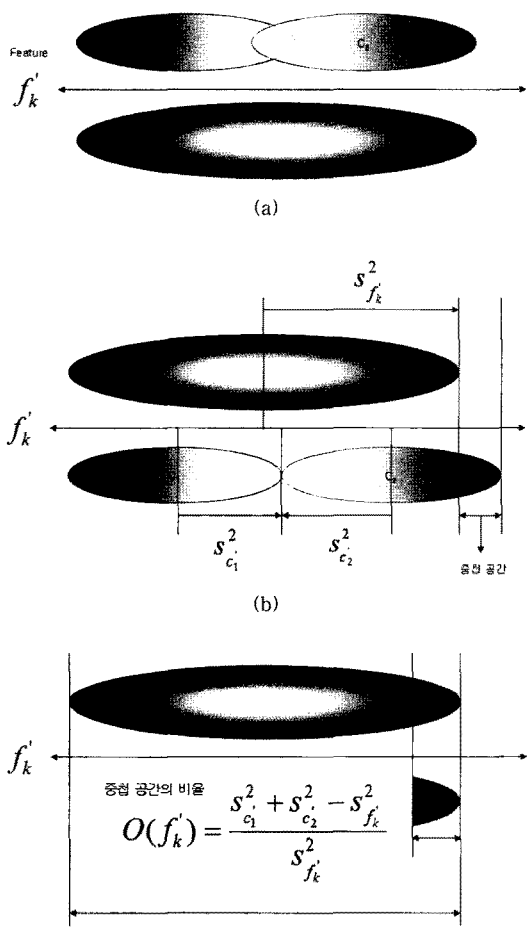


그림 15 식 (9)의 기하학적인 형태. 평가 중인 특징 f'_k와 f'_k에서 나타나는 클래스 c'_1과 c'_2의 분산을 통해 중첩 공간의 비율 $O(f'_k)$를 구함. (a) 특징 f_k에 클래스가 c_1과 c_2가 있으며, 아래쪽 타원은 특징 f_k의 분산, 위쪽 타원은 각 클래스의 분산을 나타냄. (b) 각 클래스의 분산을 합할 경우, f_k의 분산을 통해 실제 중첩되는 공간을 알아낼 수 있음. (c) 각 특징의 단위에 클래스 구분력이 영향을 받지 않도록 f_k의 분산과 중첩 공간 분산을 통해 비율을 구하여 $O(f'_k)$를 구함

수록(클래스 구분력이 좋아질수록) 0으로 수렴하도록 하였다.

$$k(f_i') = \frac{1}{-O(f_i')+2} \tag{8}$$

식 (9)는 특징 f'_i에서 나타나는 n개의 클래스들의 분산을 통해 실제 중첩되는 공간의 비율을 측정하도록 하였다. 이는 마이크로어레이를 통해 얻은 측정값들의 단위(Scale)에 영향을 받지 않도록 하기 위함이다. 이 때, $s_{f'_i}$는 특징 f'_i에서 측정된 값들의 분산이며, $s_{c'_j}$는 특징 f'_i의 클래스가 j개 일 때, j번째 클래스의 분산을 의미한다.

$$O(f_i') = \frac{1}{s^2_{f_i'}} \sum_{j=1}^{n} s^2_{c_j'} - s^2_{f_i'} \tag{9}$$

그림 15는 식 (9)에서 측정하는 개별 특징에서의 중첩 공간을 측정하는 방법에 대해 설명하고 있다. 각 클래스에 대한 분산과 그 합을 통하여 중첩 공간을 측정하고, 앞서 언급한 단위(Scale)에 영향을 받지 않기 위해 f'_i의 분산에 대한 비율을 구한다.

4. 실 험

4.1 실험 디자인

실험 마이크로어레이 데이타는 백혈병(Leukemia)과 결장(Colon) 데이타 및 NCI60 데이타를 사용하였으며, 백혈병 데이타에는 7129개의 유전자가 포함되어 있고 [3], 결장 데이타에는 2000개의 유전자가 포함되어 있으며[8], NCI60 데이타에는 백혈병 데이타와 마찬가지로 7129개의 유전자가 포함되어 있다.

실험은 먼저 GA를 통해 좋은 부집합을 선택하고, GA가 선택한 유전자 부집합의 유전자 개수와 동일한 숫자의 유전자를 t-test와 mRMR(minimum Redundancy, Maximize Relevancy), SFS를 통해 유전자를 선별하고, 각 부집합에 대해 LOOCV K-nn을 통하여 평가하였으며, 참조회수는 3으로 설정하였다[2,14]. 본 실험에서 보이고자 하는 목표는 제안한 알고리즘을 통해 선별된 유전자 부집합이 원본 마이크로어레이 데이타에 노이즈가 포함되어 있음에도 불구하고, 클래스 예측률이 높으므로 표지 유전자를 잘 추출함을 보이는 것이다.

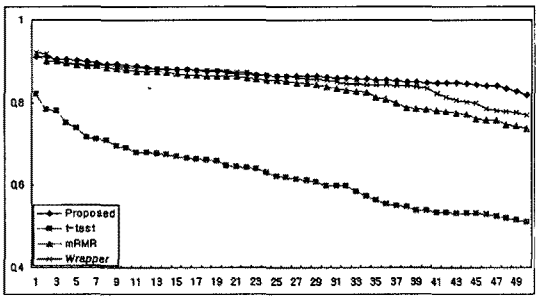
4.2 실험 결과 분석

4.2.1 클래스 예측률 및 유전자 중복도

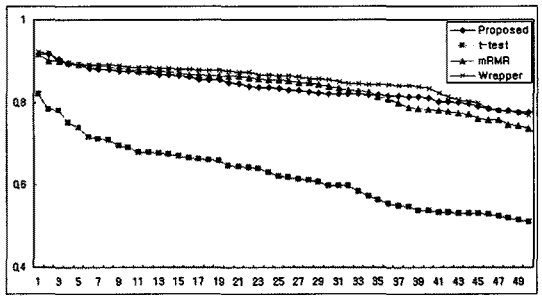
실험은 동일한 데이타에 대하여 3가지 특징 추출 방법을 적용하고, 추출된 특징 부집합의 정확도를 평가하였다. 비교 실험에 활용된 t-test와 mRMR 및 SFS의 경우 추출할 특징의 개수를 결정할 수 있으므로, 각각 50개의 특징을 추출한 뒤 클래스 예측률을 측정하였다. 백혈병 데이타에서 mRMR의 경우 클래스 예측률이 매우 낮은 것을 확인할 수 있는데, 이는 mRMR 데이타가 연속적인 값(Continuous value)으로 이루어진 데이타에 대해 이산화(Discretize) 과정에 크게 영향을 받으며, 클래스의 개수가 늘어나면 늘어날수록 취약하기 때문이다. 또한 SFS의 경우 실제 클래스 예측률을 높일 수 있는 특징이 부집합에 포함되어 있지 않을 경우 정확도를 높

표 2 각각의 특징 선별 방법을 통해 유전자를 선별한 후, K-nn을 통해 클래스를 예측한 결과 정확도. 백혈병 데이타에는 2개의 클래스로 이루어진 7129개의 유전자가 포함되어 있고, t-test를 통한 평가 결과 상위 50개의 유전자로 이루어진 특징 부잡합의 클래스 예측률이 79.1%임

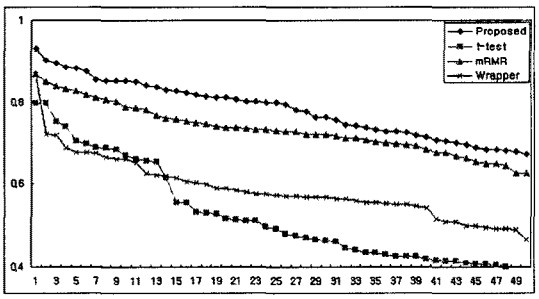
Dataset	Class	Features	No Feature Selection		t-test		mRMR		SFS		Proposed			
											Noise Filtering		No noise Filtering	
			Feature	Acc.	Feature	Acc.	Feature	Acc.	Feature	Acc.	Feature	Acc.	Feature	Acc.
Leukemia	2	7129	7129	76.4%	50	79.1%	50	80.0%	50	70.8%	50	97.2%	48	97.2%
Colon	2	2000	2000	85.5%	50	74.0%	50	78.5%	50	80.6%	26	93.5%	26	95.2%
NCI60	9	7129	7129	45.0%	50	53.3%	50	28.7%	50	13.3%	35	80.0%	35	80.0%



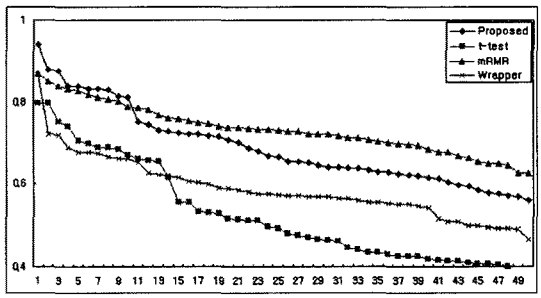
(a) Leukemia Data



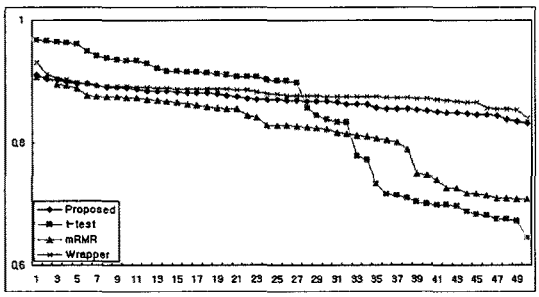
(a) Leukemia Data



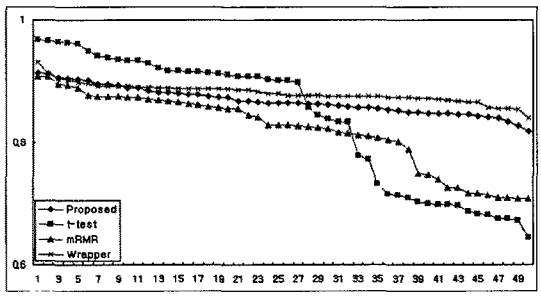
(b) Colon Data



(b) Colon Data



(c) NCI60 Data



(c) NCI60 Data

그림 16 백혈병과 결장암 및 NCI60 데이타에 대한 비중복도 비교 결과 도표. 제안한 알고리즘이 추출한 유전자들의 비중복도가 비슷하거나 높아졌음을 볼 수 있으며, 추출된 유전자 부집합에 포함된 유전자들은 기능이 서로 다른 표지 유전자일 가능성이 높음. 중복도를 전혀 고려하지 않는 t-test와는 현저한 차이가 나며, 중복도를 고려한 mRMR의 경우 클래스 예측률이 낮아 표지 유전자일 가능성이 낮음

일 수 없음을 확인할 수 있다. 이에 반하여 제안된 알고리즘은 클래스의 개수나 값의 형태에 관계없이 비교적 높은 클래스 예측률을 보이며, 기존 특징 추출 기법보다 상대적으로 적은 유전자들을 추출하였음에도 불구하고 정확도가 높으므로, 추출된 특징들이 표지 유전자일 가능성이 매우 높다. 노이즈가 제거된 데이타로 유전자 부집합을 추출한 뒤, 실제 평가 때는 노이즈가 포함되어 평가를 하게 되므로 정확도가 다소 감소하지만, 유전자 부집합의 비중복도 검사에서 노이즈를 제거하여 유전자 부집합을 추출한 것에 비해 비중복도가 낮았다. 이는 비슷한 기능을 하는 유전자를 이용하여 만들어진 유전자 부집합이 정확도는 높아지지만 비중복도에서 낮아지는 반비례 관계가 있다고 생각할 수 있다.

유전자 중복도의 경우 그림 16을 통해 비교하였다. GA는 추출될 특징의 개수를 결정할 수 없으므로, GA가 추출해낸 부집합들 중에서 추출된 빈도가 높은 유전자를 50개 선정하여 중복도를 비교하였다. t-test와의 비교 결과 특징 비중복도가 명확하게 차이가 남을 알 수 있으며, 비슷한 비중복도를 보이는 mRMR의 경우 클래스 예측률이 현저하게 떨어져 추출된 유전자들이 표지 유전자가 아닐 가능성이 높아진다. 또한, SFS의 경우 특징을 50개 이후로 추가할수록 정확도가 더욱 상승할 것을 예상할 수 있지만, 유전자 부집합에 포함된 유전자의 숫자가 증가하므로 표지 유전자 추출에는 부적합함을 알 수 있다.

4.2.2 생물학적 유효성 분석

그림 17은 유전자 알고리즘이 선별한 유전자 부집합에 포함된 유전자들의 빈도를 나타낸 것이다. 또한 표 3의 대상 데이타는 백혈병 데이타이며, 선택 유전자의 빈도수를 얻기 위해 1000개의 유전자 부집합을 구하여 얻은 유전자 부집합에 대해 각 유전자 부집합에서 공통적으로 선택된 유전자의 빈도순으로 상위 25개의 유전자에 대한 목록을 나타내고 있다. 표 3은 제안한 알고리즘과 t-test를 통해 선별된 유전자들이 실제 생물학 분야에서 불리는 명칭과 관련 질병 및 참고 문헌을 나타내고 있으며, 표 4는 각 유전자들의 세부 내용을 나타내고 있다. 이 때, 제안된 알고리즘을 통해 찾아낸 유전자 중 일부는 이미 많은 연구가 이루어져 실제 백혈병과의 연관성이 있음을 알 수 있는 실험결과들이 풍부하게 제공되고 있으나, 일부의 유전자들은 유전자 자체에 대한 연구가 활발히 이루어지지 않아 앞으로 백혈병과의 연관성을 밝힐 수 있는 연구가 필요할 것으로 보인다.

5. 결 론

본 논문에서는 GA를 통해 특징을 선별하고, 선별된 특징에 대해서 평가 함수를 적용하여 클래스 구분력은 높고, 중복도는 낮은 특징 부집합을 선별하는 알고리즘을 제안하였다. 클래스 구분력을 측정하기 위해 마이크로어레이 데이타에 포함된 노이즈를 판별하고, 노이즈를 회피하여 특징 공간에서 노이즈에 편중되지 않은 클래

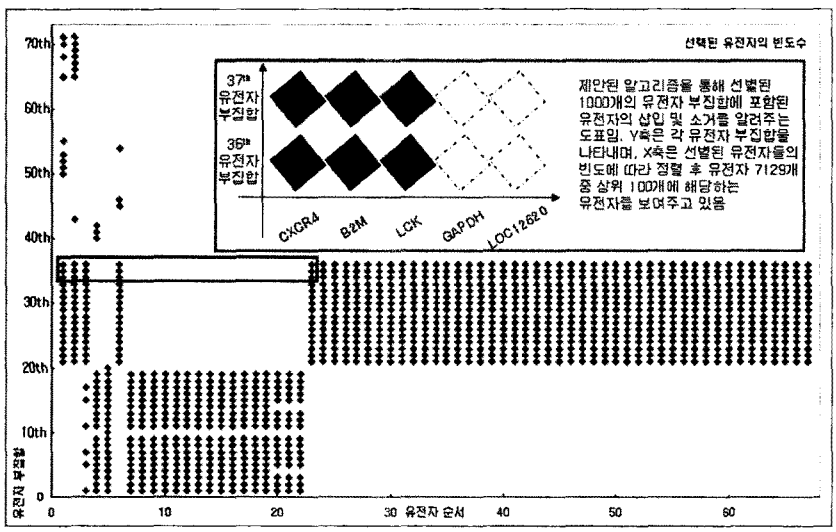


그림 17 선택 유전자 빈도 도표의 예제. 제안된 알고리즘을 통해서 얻은 유전자 부집합에 포함된 횟수를 나타내는 도표임. 제안한 논문에서는 백혈병 데이타(유전자 7129개)에 대해 총 1000개의 유전자 부집합을 얻었으며, 빈도순으로 100개의 유전자만을 나타내고 있음. X축은 선택된 유전자를 나타내며, Y축은 유전자 부집합을 의미함

표 3 제안된 알고리즘을 통해 얻은 표지 유전자 비교표. 빈칸은 관계가 밝혀지지 않았거나 해당 유전자에 대한 연구가 필요한 상태임. 유전자의 연관성은 PubMed(www.ncbi.nlm.nih.gov)에서 논문 검색을 통해 구성하였으며, 그 의미는 해당 유전자가 영향을 주는 질병이나 기관 또는 세포임

Seq.	Official Name	Relation						
		Leukemia	Tumor	Cancer	Leukocyte	B cell	T cell	Interleukin
163	TYMS	O	O	O				
781	ATF4	O	O	O			O	
1086	MME	O	O	O	O			O
1109	IFITM1		O	O	O	O		
1262	TNFSF8	O	O	O	O	O	O	O
1433	EIF1		O	O				O
1962	CD81		O	O	O	O	O	O
2186	NFKBIA	O	O	O			O	O
2259	RPS19	O	O		O		O	
2354	CCND3	O	O	O		O	O	O
2402	AZU1		O		O			O
4017	HLA-DRA		O	O	O	O	O	
4050	CD3D	O	O				O	O
4643	TRAF4		O	O		O		O
5088	HSPB1	O	O	O			O	
5552	CXCR4	O	O	O	O	O	O	O
5688	CD24	O	O	O	O	O	O	O
5926	B2M	O	O	O	O	O	O	O
6218	ELA2	O	O	O	O			O
6226	GNAS		O	O	O	O		
6510	LCK	O	O	O		O	O	O
6613	PABPC1		O	O			O	
6659	SOD2		O	O				O
6696	MAL		O	O			O	
7070	HLA-DRB1	O	O		O			O

스의 구분력을 측정할 수 있도록 하였다. 또한 특징 선별에 있어서 특징 사이의 상관관계가 없음을 가정하고 있는 기존 알고리즘들의 한계를 극복하기 위하여, 소모적인 방법을 택하지 않고 GA를 통해 특징 부집합을 선별하도록 하여 마이크로어레이 데이타에서도 효과적으로 특징 부집합을 선별할 수 있도록 하였다. 그 결과 클래스 예측률은 기존 방법론과 비교하여 상대적으로 높았고, 특징 부집합의 비중복도가 높아 추출된 표지 유전자들이 서로 유사하지 않으므로 표지 유전자들의 실제 기능이 유사하지 않음을 예측할 수 있다. 또한 클래스 예측률을 높이기 위한 유전자 부집합을 선별하는 것과 중복도가 낮은 유전자 부집합은 서로 상관관계에 있음을 알 수 있었다.

본 논문의 한계는 다음과 같다. 첫째, 유전자 부집합을 찾는 방법으로써 GA를 사용하고 있는데, GA 알고리즘의 단점은 알고리즘 종료시점이 명확하지 않고, 유전자 부집합을 찾아가는 과정이 비교적 무작위에 의존하는 편이기 때문에 GA가 생성하는 최종 결과가 비록 좋은 결과를 얻더라도 항상 동일한 결과를 얻기는 어렵다. 본 논문에서 다양한 유전자 부집합을 보인 것도 동일한 맥락에서 설명할 수 있다. 따라서 향후 연구에는 이러한 문제를 극복하기 위해 효과적인 선형 검색(Linear Search) 알고리즘으로의 재설계가 있어야 할 것이다. 둘째, 본 논문에서 제안한 적합 함수는 그 설계 과정에서 많은 부분이 고려되었지만, 실제 클래스의 중첩 영역 결정에 있어서 다양한 형태를 고려하진 못하였다. 또한 유전자들의 비중복도를 취합하는 방법 역시 기존에 제공된 상관 계수를 반복 사용하고 있어 유전자 원본 집합의 중복도가 너무 높거나 낮을 경우에 대해 오히려 둔감한 문제가 있었으므로 중복도 측정 함수의 재정의와 정리가 필요할 것으로 보인다. 마지막으로 중복도를 고려하여 유전자를 추출하였지만, 특징을 추출하는 다른 알고리즘들과의 비교에서 현격한 차이가 나지는 않았다. 이를 통해 마이크로어레이 데이타에 포함된 유전자들의 실험값들을 상관계수만을 통해 각 유전자들 고유 기능의 중복도를 추정하는데 한계가 있음을 알 수 있었다.

표 4 표 3에서 언급된 유전자들의 상세 설명 도표. 표 4에서 사용된 Sequence에 해당하는 유전자의 상세 설명을 표현하고 있음. Gene Accession Number는 PubMed에서 제공하는 GeneBank에서 해당 유전자의 정보를 얻을 수 있는 ID를 의미함

Seq.	Gene Accession Number	Gene Description
163	D00596	TYMS Thymidylate synthase
781	D90209	ATF4 CAMP-dependent transcription factor ATF 4 (CREB2)
1086	J03779	MME Membrane metallo-endopeptidase (neutral endopeptidase, enkephalinase, CALLA, CD10)
1109	J04164	RPS3 Ribosomal protein S3
1262	L09753	CD30LG CD30 antigen ligand
1433	L26247	RPL3 Ribosomal protein L3
1962	M33680	26-kDa cell surface protein TAPA 1 mRNA
2186	M69043	MAJOR HISTOCOMPATIBILITY COMPLEX ENHANCER BINDING PROTEIN MAD3
2259	M81757	40S RIBOSOMAL PROTEIN S19
2354	M92287	CCND3 Cyclin D3
2402	M96326_rna1	Azurocidin gene
4017	X00274	HLA CLASS II HISTOCOMPATIBILITY ANTIGEN, DR ALPHA CHAIN PRECURSOR
4050	X03934	GB DEF = T-cell antigen receptor gene T3-delta
4643	X80200	MLN62 mRNA
5088	Z23090	HSPB1 Heat shock 27kD protein 1
5552	L06797_s	PROBABLE G PROTEIN-COUPLED RECEPTOR LCR1 HOMOLOG
5688	L33930_s	CD24 signal transducer mRNA and 3' region
5926	J00105_s	BETA-2-MICROGLOBULIN PRECURSOR
6218	M27783_s	ELA2 Elastatse 2, neutrophil
6226	M21142_cds2_s	Guanine nucleotide-binding protein G s-alpha-3 gene extracted from Human guanine nucleotide-binding protein alpha-subunit gene (G s alpha)
6510	U23852_s	GB DEF = T-lymphocyte specific protein tyrosine kinase p56lck (lck) abberant mRNA
6613	Z48501_s	GB DEF = Polyadenylate binding protein II
6659	X65965_s	GB DEF = SOD 2 gene for manganese superoxide dismutase
6696	X76223_s	GB DEF = MAL gene exon 4
7070	M33600_f	HLA-DRB1 Major histocompatibility complex, class II, DR beta 5

참 고 문 헌

[1] Stephen Erickson, Hierarchical empirical Bayes analysis of genomic microarrays, University of California, Los Angeles, AAT 3247476, 2006.

[2] Peng H.C., Long, F., Ding, C., "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.27, pp. 1226-1238, 2005.

[3] Ian A. Wood, Peter M. Visscher, Kerrie L. Mengersen, "Classification based upon gene expression data: bias and precision of error rates," Bioinformatics, Vol.23, pp. 1363-1370, 2007.

[4] Yudi Pawitan, Karuturi R. Krishna Murthy, Stefan Michiels, Alexander Ploner, "Bias in the estimation of false discovery rate in microarray studies," Bioinformatics, Vol.21, p. 3865, 2005.

[5] Dan Nettleton, "A Discussion of Statistical Methods for Design and Analysis of Microarray Experiments for Plant Scientists," Plant Cell, Vol.18, pp. 2112-2121, 2006.

[6] Kevin Dobbin, Richard Simon, "Sample size determination in microarray experiments for class comparison and prognostic classification," Biostatistics, Vol.6, p. 27, 2005.

[7] T. R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, Vol.286, pp. 531-537, 1999.

[8] Danh V. et al., "Tumor classification by partial least squares using microarray gene expression data," Bioinformatics, Vol.18, No. 1, pp. 39-50, 2001.

[9] David P. Kreil, Roslin R Russell, "There is no silver bullet - a guide to low-level data transforms and normalisation methods for microarray

data," Briefings in Bioinformatics, Vol.6, pp. 86-97, 2005.
[10] Seo Young Kim, Jae Won Lee, In Suk Sohn, "Comparison of various statistical methods for identifying differential gene expression in replicated microarray data," Statistical Methods in Medical Research, Vol.15, p. 3, 2006.
[11] Carla S. Möller-Levet, Catharine M. West, Crispin J. Miller, "Exploiting sample variability to enhance multivariate analysis of microarray data," Bioinformatics, Vol.23, pp. 2733-2740, 2007.
[12] Guo Yu, Statistical issues in microarry data analysis: Array-to-array normalization, Empirical Bayes batch effect adjustment, and Pearson's correlation coefficient in the context of replicated experiments, Harvard University, AAT 3217745, 2006.
[13] Cianluca B., "A Blocking Startegy to Improve Gene Selection for Classification of Gene Expression Data," IEEE/ACM Trans. Computational Biology and Bioinformatics, pp. 293-300, 2007.
[14] Miin-Shen, Kuo-Lung Wu, "A Similarity-Based Robust Clustering Method," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.26, pp. 434-448, 2004.
[15] Yvan Saeys, Iñaki Inza, Pedro Larrañaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, Vol.23, pp. 2507-2517, 2007.

이 재 성
2007년 중앙대학교 컴퓨터공학과 학사 2008년~현재 중앙대학교 컴퓨터공학과 석사과정. 관심분야는 바이오정보학, 데이타마이닝

김 대 원
1997년 경북대학교 컴퓨터공학과 학사 1999년 KAIST 전산학과 석사. 2004년 KAIST 전자전산학과 박사. 2005년~현재 중앙대학교 컴퓨터공학부 조교수. 관심분야는 바이오정보학, 의료정보학, 데이타마이닝