2 단계 접근법을 통한 통합 마이크로어레이 데이타의 분류기 생성

(Building a Classifier for Integrated Microarray Datasets through Two-Stage Approach)

윤 영 미 † 이 종 찬 †† 박 상 현 †††
(Youngmi Yoon) (Jongchan Lee) (Sanghyun Park)

요 약 마이크로어레이 데이타는 동시에 수 만개 유전자의 발현 값을 포함하고 있기 때문에 질병의 발현 형질 분류에 매우 유용하게 쓰인다. 그러나 동일한 생물학적 주제라 할지라도 여러 독립된 연구 집단에서 생성된 마이크로어레이의 분석결과는 서로 다르게 나타날 수 있다. 이에 대한 주된 이유는 하나의 마이크로어레이 실험에 참여한 샘플의 수가 제한적이기 때문이다. 따라서 개별적으로 수행된 마이크로어레이 데이타를 통합하여 샘플의 수를 늘리는 것은, 보다 정확한 분석을 하는데 있어 매우 중요하다. 본 연구에서는 이에 대한 해결 방안으로 두 단계 접근방법을 제안한다. 제 1 단계에서는 개별적으로 생성된 동일 주제의 마이크로어레이 데이타를 통합한 후 인포머티브(Informative) 유전자를 추출하고 제 2 단계에서는 인포머티브 유전자만을 이용하여 클래스 분류(Classification) 과정 후 분류자를 추출한다. 이 분류자를 다른 테스트 샘플 데이타에 적용한 실험결과를 보면 마이크로어레이 데이타를 통합하여 샘플의 수를 증가시킬수록, 비교 방법에 비해 정확도가 최대 24.19% 높은 분류자를 만들어 내는 것을 알 수 있다.

키워드 : 바이오 인포마틱스, 마이크로어레이 데이타 분석, 마이크로어레이 데이타 통합, 클래스 분류자 찾기, 인포머티브 유전자 추출

Abstract Since microarray data acquire tens of thousands of gene expression values simultaneously, they could be very useful in identifying the phenotypes of diseases. However, the results of analyzing several microarray datasets which were independently carried out with the same biological objectives, could turn out to be different. One of the main reasons is attributable to the limited number of samples involved in one microarry experiment. In order to increase the classification accuracy, it is desirable to augment the sample size by integrating and maximizing the use of independently-conducted microarray datasets. In this paper, we propose a novel two-stage approach which firstly integrates individual microarray datasets to overcome the problem caused by limited number of samples, and identifies informative genes, secondly builds a classifier using only the informative genes. The classifier from large samples by integrating independent microarray datasets achieves high accuracy up to 24.19% increase as against other comparison methods, sensitivity, and specificity on independent test sample dataset.

Key words : Bioinformatics, Microarray data analysis, Microarray data Integration, Microarray classification, Informative gene selection

1. 서 론

암세포가 가지고 있는 분자적 특성을 암의 진단 및 분류에 적용하기 위하여 암세포 특이적 유전자 발현 패턴을 연구하고자 하는 흐름이 있어왔고, 마이크로어레이는 소수의 실험만으로 많은 수의 유전자 발현 정도를 밝힐 수 있다는 점에서 암 진단 분야에 획기적 진전을 가져올 것으로 기대되고 있다. 하나의 실험에서 얻어진 마이크로어레이 데이타는 그림 1과 같이 각 열은 하나

· 본 연구는 과학기술부 과학재단 목적기초연구(R01-2006-000-11106-0)로 수행되었음
† 정 회 원 : 연세대학교 컴퓨터과학과
amyyoon@cs.yonsei.ac.kr
†† 학생회원 : 연세대학교 컴퓨터과학과
jclee@cs.yonsei.ac.kr
††† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr
논문접수 : 2006년 6월 15일
심사완료 : 2006년 11월 12일

의 샘플을, 각 행은 하나의 유전자를 의미하며 각 셀의 값은 특정 유전자가 특정 샘플에서 발현된 정도를 나타내는 수치이다. 또한 각 샘플은 암, 정상 등과 같은 클래스 레이블(label)을 갖는다. 마이크로어레이 실험 자료는 소수의 유전자만을 대상으로 하는 분자생물학 실험과 달리, 수천 내지 수만 개의 프로브(probes)를 대상으로 발현 수치를 얻게 되므로 이의 분석 및 해석 과정에 있어서 필연적으로 통계 기법의 적용이 필요하다.

	C_1			C_2		
	S_1	S_2	S_3	S_4	S_5	S_6
G_1	3	5	7	9	11	13
G_2	15	32	23	12	2	3
G_3	....	....	....	....	....	....
G_4						
G_5						
G_6						

그림 1 하나의 마이크로어레이 데이타 실험결과
(Si는 샘플, G_i는 유전자, C_1과 C_2는 클래스 레이블)

이러한 분석에 통계 기법이 적용되기 때문에 실험에 사용된 샘플의 수가 많으면 많을수록 결과에 대한 신빙성이 높아질 수 있다. 특히 암과 관련된 연구의 경우 충분한 수의 샘플을 대상으로 분석하는 것이 신뢰할 수 있는 결과를 도출하는 데 필수적이다.

최근 들어 각 연구자들에 의해 독립적으로 얻어진 마이크로어레이 데이타 중 동일한 종류의 세포나 조직이 사용된 데이타 세트를 통합하여 분석함으로써 각각의 독립적 분석보다 더 유의한 결과를 얻을 수 있을 것이라는 가능성이 제기되고 있다[1]. 그러나 현재까지는 동일 주제에 대해 실험한 마이크로어레이 데이타를 대상으로 한다 하더라도 사용된 플랫폼(platform), 유전자 세트, 각 실험실마다 적용된 프로토콜(protocol)이 다를 경우 이를 직접 통합하는 솔루션이 완전하지 않은 상태이기 때문에, 이를 통합하여 분석하는데 많은 한계점을 가지고 있다. 또한 독립적으로 수행된 마이크로어레이의 유전자 발현 값은 직접적인 비교가 불가능하기 때문에, 서로 상이한 환경에서 실험된 다른 마이크로어레이 데이타를 통합 한다는 것은 매우 어려운 문제이다. 따라서 본 논문에서는 두 단계 접근을 통하여 독립적인 마이크로어레이 데이타를 통합하여 신뢰도가 높은 분류자를 찾는 방법을 제안한다.

제 1 단계에서는 독립적으로 생성된 동일 주제의 마이크로어레이 데이타를 복잡한 정규화 과정 없이 효율적으로 통합하는 새로운 방법을 제안한다. 또한, 통합 마이크로어레이 데이타로 부터 발현형질(phenotype)과 관련이 있는, 인포머티브 유전자만을 효과적으로 추출하는 방법을 제안한다. 본 논문에서 제안하는 인포머티브 유전자 추출 방법은 샘플 내 순위 기반 접근(Rank Based Approach)을 통해 간단한 정규화 과정을 거친 통합 마이크로어레이 데이타에 효과적으로 적용할 수 있다.

제 2 단계인 클래스 분류 단계에서는 제 1 단계에서 생성된 인포머티브 유전자만을 사용하여 생물학적으로 해석이 용이한 간단한 클래스 분류 규칙을 생성한다. 기존의 TSP(Top Scoring Pair) 방법[2]은 분류 규칙에 참여하는 유전자의 개수를 중복을 허용하지 않는 2개로 한정하였으며 규칙의 개수도 1개로 제한하였다. 그러나 이러한 제약 조건은 생물학적인 근거가 미흡할 뿐만 아니라 규칙에 참여하는 유전자가 테스트 샘플에 존재하지 않을 가능성도 배제할 수 없다. 따라서 본 논문에서는 규칙에 참여하는 유전자의 개수를 일반화하고 규칙의 개수도 확장하여 좀 더 신뢰성 있고 정확한 규칙을 찾는 방법을 제안한다. 본 논문의 분류자는 $K(\geq 5)$개의 규칙으로 이루어져 있으며, 한 개의 규칙은 세 개유전자 간의 관계식과 클래스 레이블로 이루어져 있다. 이 방법은 제 1 단계에서 클래스 분류와 연관이 있는 유전자만을 추출한 후에 적용하기 때문에 클래스 분류의 정확도를 높일 수 있을 뿐만 아니라 제 2 단계의 계산 량을 줄일 수 있다는 장점이 있다. 제 1단계에서 통합에 사용된 학습데이타(training dataset)의 샘플수가 커질수록 테스트 데이타의 분류 정확도가 더 높아지는 것을 기존의 알고리즘과의 비교실험을 통해 알 수 있었다. 본 논문의 순차적 두 단계 접근 방법은 기존의 실험[19-21]으로 얻어진 마이크로어레이데이타의 활용성을 획기적으로 향상하고, 마이크로어레이 통합분야에 새로운 가치를 창출하는 패러다임으로 볼 수 있다.

2. 관련 연구

2.1 마이크로어레이 데이타 통합

최근까지 마이크로어레이 데이타 통합을 위하여 사용된 방법으로는 마이닝 기법의 하나인 메타 마이닝이 있다. 이 방법은 개별적으로 얻어진 마이크로어레이 실험의 결과를 통합하여 분석하는 방식이다[3]. 그러나 각 개별 연구의 샘플의 수가 일반적으로 적기 때문에 개별 연구 결과 자체가 좋지 않은 경우가 많고 이런 결과의 통합은 더 좋지 못한 분석을 낳을 수도 있다. 또 다른 통합의 방법으로는 개별 연구로 얻어진 데이타 값을 공통의 스케일을 갖는 값으로 정규화 하여 결합시키는 방법이 있다[4]. 가장 대표적인 예는 Z-Score로 변형시켜 결합하는 경우이다. 그러나 복잡한 정규화 과정을 거쳐

야 하는 이러한 방법은 전처리(preprocessing) 단계에 많은 비용을 지불하게 된다. 이 밖에 데이타 통합의 모델을 제시한 연구로는 이질 마이크로어레이 데이타 통합에 상관서명(Correlation Signature)을 사용한 방법이 있다[5].

2.2 인포머티브 유전자 식별

마이크로어레이 데이타를 분석 하는데 있어 가장 큰 제약 조건은 실험에 참여하는 샘플의 수에 비해 유전자의 수가 너무 많다는 점이다[6]. 그러나 실질적으로 클래스 결정에 영향을 미치는 관련(relevant) 유전자의 수는 매우 한정적이고 대부분의 유전자들은 클래스 판별에 영향을 미치지 않는 잡음(noise) 유전자이다. 인포머티브 유전자는 아래 그림 2의 두 번째와 같이 클래스 A에서는 모두 높은 발현 값을 나타내고 클래스 B에서는 모두 낮은 발현 값을 나타내는 유전자로 정의할 수 있다. 반면에 그림 2의 세 번째와 같이 특정 클래스에 대해 일관성 있는 발현 값을 제공하지 못하는 유전자는 연관성이 없는 잡음 유전자로 판단할 수 있다[7]. 따라서 특정 질병에 관여하는 의미 있는 유전자만을 추출한 후 이를 대상으로 분류 방법을 적용하는 것이 합리적이다.

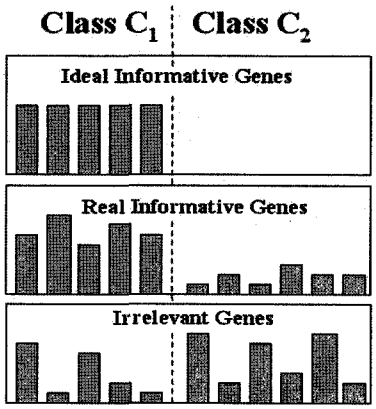


그림 2 발현 값의 유형으로 본 인포머티브 유전자

발현형질에 관여하지 않는 유전자를 제거하고 인포머티브 유전자만 식별해 내는 작업을 특징추출(Feature Selection)이라 하는데 이는 마이크로어레이 데이타 분석에 매우 중요하다[8]. 현재 이러한 인포머티브 유전자를 정확하고 효과적으로 추출하기 위한 여러 방법들이 제시되고 있다. 특징추출의 대표적 방법으로 PCA(Principal Component Analysis)[9]와 같은 방법이 있다. 그러나 PCA방법은 아이겐 벡터로 마이크로어레이 데이타의 차원을 줄이기는 하지만 클래스분류와 연관성이 있는 유전자를 개별적으로 찾아 주지는 못한다. 순위 방법 중 모수적(parametric) 방법은 t-statistics나 Golub [10]의 방법 같이 데이타를 대표하는 통계적인 모델을 가정하여 그 모델을 대표할 수 있는 모수(예: 평균과 분산) 를 저장하게 된다. 이 방법은 수 만개의 유전자의 발현 값을 매우 적은 수의 모수로 치환하기 때문에 정보의 손실을 발생시킬 수 있다는 문제점이 있다. 반면 비모수적(non-parametric) 방법은 하나의 유전자에 대한 모든 샘플 값을 정렬하여 그 유전자가 두 개의 클래스 그룹에서 다르게 발현된 정도에 대한 점수(완벽한 분리를 방해하는 정도)를 계산한다[8,11]. 유전자를 특징으로 보았을 때 특징추출 방법 중에서 가장 일반적으로 많이 쓰이는 방법은 순위 기반 방법이다. 순위 기반 특징 추출 방법은 각 특징이 다른 특징보다 더 유의한지의 정도를 통계적 수치로 측정한 후 정렬하여 상위의 특징을 선택하는 것이다.

대표적인 인포머티브 유전자 추출 방법으로는 인포메이션 게인(Information Gain)[12], 릴리프-에프(Relief-F)[13], 켄달의 상관계수(Kendall's Correlation Coefficient)[14]를 응용한 Park방법[11]이 있다.

인포메이션 게인 방법은 무질서한 정도의 척도인 엔트로피를 활용하는 알고리즘이다. 우선 X를 유전자로, Y를 클래스 레이블(정상(Normal), 암(Tumor))로 정하여 각각의 엔트로피를 계산한다. 엔트로피의 계산식은 아래와 같이 정의된다.

$$H(Y) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log P(y|x)$$

그리고 아래 정의한 인포메이션 게인(IG) 수식에 의해 IG값을 구하여 이 값이 큰 유전자를 추출하는 방법이다.

$$IG(Y|X) = H(Y) - H(Y|X)$$

릴리프-에프는 각 샘플에 대해 가장 가까운 f개의 동일 클래스의 샘플(hit sample)과 다른 클래스의 샘플(miss sample)을 찾아 모든 유전자에 대해 해당 샘플과의 발현 값 차이를 구한다. 그리고 해당 샘플과 동일 클래스의 샘플 사이의 차이는 가중치를 감소시키고 해당 샘플과 다른 클래스의 샘플 사이의 차이는 가중치를 증가시킨다. 이러한 계산을 모든 샘플에 대해 모두 누적하여 각 유전자의 가중치를 계산한 후 이 값이 큰 순서대로 유전자를 추출하는 방법이다.

Park[11]은 켄달의 상관계수를 이용하여 두 클래스 그룹에서 그 유전자가 다르게 발현된 정도를 측정하는 점수 함수를 정의하였다. 그러나 이 방법은 각 유전자의 발현 값을 그대로 사용하기 때문에 서로 상이한 환경에서 실험된 다른 마이크로어레이 데이타에는 적용할 수

없다.

위에 언급한 세 개의 방법들은 모두 유전자 발현 값을 그대로 사용하였으며 마이크로어레이 데이타의 통합 또는 정규화에 대한 고려가 전혀 없다.

2.3 클래스 분류

많은 클래스 분류 방법 중에 대표적인 것으로는 SVM[15,16], k-Nearest Neighbor[17]의 방법이 있다. SVM은 기계학습 알고리즘에 바탕을 둔 것으로 하이퍼플레인(Hyper plane)으로 대표되는 선형 결정 규칙(Linear Decision Rules)을 학습해 나아가는 것이다. SVM은 마이크로어레이의 클래스 분류뿐만 아니라 회귀분석, 밀도 예측 같은 다양한 분야에도 사용된다. 그러나 마이크어레이 데이타에 적용하기 위해서는 실험적으로 여러 가지 종류의 파라미터 조정을 필요로 하기 때문에 다소 복잡하다는 단점이 있다. k-Nearest Neighbor(k-NN)는 새로운 샘플에 대하여 학습 데이타 개체 중에서 유사한 것들을 선택하여 샘플의 클래스를 분류하는 알고리즘이다. 그러나 k-NN 알고리즘은 모든 유전자에 동일한 가중치를 부여하였을 경우 좋은 성능을 제공하지 못한다는 단점을 가지고 있다.

또 다른 클래스 분류 방법 중 파라미터를 사용하지 않고 데이타에 따라 처리되는 기계학습 방법으로 Xu가 제안한 TSP[2]와 Tan이 제안한 k-TSP[18] 방법이 있다. TSP는 가장 높은 점수를 갖는 유전자의 쌍을 찾는 알고리즘이다. 모든 유전자 쌍 (X_i, X_j)에 대하여, "$X_i < X_j$" 관계가 두 클래스에서 나타나는 상대 빈도를 각각 구하여 그 차이를 계산하여 점수 함수로 사용한다. 이 점수가 높을수록 그 유전자 쌍은 두 클래스를 잘 구별한다고 말할 수 있으며 점수가 가장 큰 한 쌍의 유전자가 TSP 분류자로 사용된다. k-TSP는 TSP 방법을 확장한 것으로 k개의 상위 점수를 얻은 유전자 쌍을 분류자로 사용하는 방법이다. TSP의 경우 단지 두 개의 유전자가 분류자가 되기 때문에 생물학적 해석의 용이함은 있으나 학습데이타를 약간만 변동시켜도 TSP 분류자 자체가 변할 수 있다. 또한 분류 규칙에 참여하는 유전자가 테스트 샘플에 존재하지 않을 가능성도 배제할 수 없다.

TSP와 k-TSP 연구에서는 인포머티브 유전자를 먼저 추출하는 단계 없이 바로 클래스 분류자를 찾는 작업을 수행하며 전체 유전자가 분류규칙 생성에 관여하게 되므로, 데이타를 통합함에 따라 계산 량의 증가를 초래한다.

3. 연구 방법

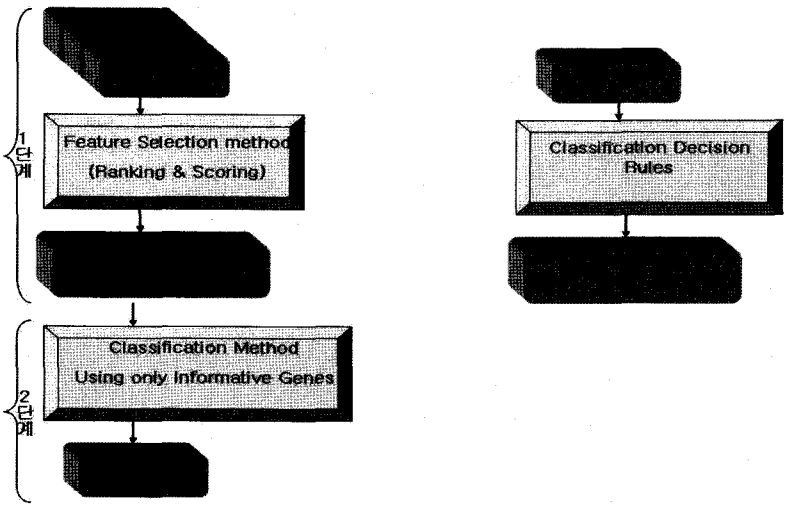
본 장에서는 시스템의 전체적인 개요를 먼저 기술하고 두 단계 알고리즘을 기술한다. 3.2절에서는 마이크로어레이 데이타 통합의 과정과 인포머티브 유전자를 효과적으로 추출할 수 있는 방법을 기술하고, 3.3절에서는 추출된 인포머티브 유전자를 이용하여 두 클래스를 잘 구분 짓는 유전자 집합 및 대소 관계를 알아낼 수 있는 k-TST(k-Top Scoring Triple) 분류 방법에 대해 설명한다.

3.1 시스템 개요

본 논문에서 제안하는 시스템의 전체적인 개요는 그림 3과 같다. 본 논문에서 제안하는 시스템은 두 단계로 구성되는데 제 1단계에서는 개별적으로 생성된 마이크로어레이데이타의 통합과 인포머티브 유전자의 추출이 이루어진다. 개별적으로 시행된 마이크로어레이 실험의 경우, 각 실험의 발현 값의 척도가 다르며 유전자 세트의 종류도 다르다. 따라서 이를 통합하기 위해서 우선 모든 마이크로어레이 데이타에서 공통적으로 사용된 유전자만을 추출한다. 그리고 각 실험에서의 샘플의 발현 값을 각 샘플 내에서의 순위 값으로 치환하여 통합한다. 일단 발현 값을 순위 값으로 치환하면 각기 다른 실험에서 유래한 샘플들도 유전자 순서가 같다면 통합이 가능하다. 이 후 각각의 유전자에 대하여 두 클래스 그룹에서 그 유전자가 다르게 발현된 정도를 측정하는 점수를 켄달의 상관계수를 이용하여 계산한다. 이 때 점수 값이 아주 작거나 혹은 아주 큰 값을 가진 유전자가 인포머티브 유전자가 될 수 있다.

제 2단계에서는 1단계에서 추출한 인포머티브 유전자만을 이용하여 클래스 분류 규칙을 만들어 낸다. 임의의 3개의 유전자를 X_i, X_j, X_k이라 하자. 모든 3개의 유전자 X_i, X_j, X_k에 대하여, 값의 대소 관계를 생성하면 총 6가지의 분류규칙을 만들어 낼 수 있다. 모든 샘플은 두 개의 클래스 C_1, C_2 중의 하나에 속한다. 클래스 C_1을 레이블로 갖는 샘플 중에서 각각의 관계를 만족하는 샘플의 수를 C_1 전체 샘플수로 나누어 C_1 클래스에서 해당 관계가 나타날 확률을 구한다. 그리고 클래스 C_2를 레이블로 갖는 샘플 중에서 각각의 관계를 만족하는 샘플의 수를 C_2 전체 샘플수로 나누어 C_2 클래스에서 해당 관계가 나타날 확률을 구한다. 그리고 각각의 분류규칙에 대하여 이 두 확률 값의 차이를 계산한다. 이 값의 차이가 큰 k개의 분류규칙이 두 클래스를 잘 구분할 수 있는 분류자의 역할을 한다. 분류규칙의 최적 개수 k는 학습데이타에 LOOCV[1](Leave One Out Cross Validation)를 적용하여 얻는다. 각 분류규칙은 (3개의 유전자로 이루어진 집합, 유전자 간의 크기를 비교한 관계식, 이 관계식의 클래스레이블)로 이루어져 있다. 새로

1) LOOCV 방법은 하나의 마이크로어레이 데이타 내에서 하나의 샘플을 제외한 나머지 샘플들을 이용하여 규칙을 생성하고 이를 제외한 하나의 샘플에 적용하여 규칙의 정확도를 측정하는 방법이다.



(a) 클래스 분류 규칙을 얻기 위한 2 단계 접근방법 (b) 테스트 샘플의 클래스 레이블 판정

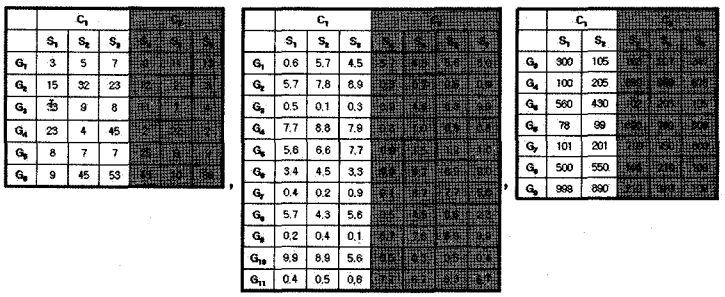
그림 3 시스템의 전체적인 개요

운 테스트 샘플이 주어지면 이 분류자를 적용하여 다수의 판정을 얻은 클래스 레이블을 계산하여 그 샘플의 실질적인 클래스레이블과 비교하여 정확도를 측정한다.

3.2 마이크로어레이 데이타 통합과 인포머티브 유전자 추출

동일한 주제에 대한 마이크로어레이 실험이라 할지라도 실험에 사용된 프로브(Probe, 탐침유전자) 세트 및 유전자의 종류는 다를 수 있다. 그렇기 때문에 개별적으로 수행된 마이크로어레이 실험을 통합하기 위해서는 그림 4와 같이 먼저 독립적으로 수행된 동일 주제의 마이크로어레이 데이타들 중에서 그림 5와 같이 공통적으로 사용된 유전자 집합(G_3, G_4, G_5, G_6)만을 추출해야 한다.

또한 공통적으로 사용되는 유전자집합이 동일 순서를 가졌다 할지라도 서로 다른 실험조건 (프로토콜)으로 인해 마이크로어레이 데이타들의 발현 값의 스케일은 다를 수 있기 때문에 정규화 과정 없이 직접적인 통합은 불가능하다. 따라서 본 논문에서는 그림 6에서와 같이 실제 발현 값이 아닌 각 샘플 내에서의 해당 유전자의 발현 값의 순위를 이용하여 통합을 가능케 하였다.

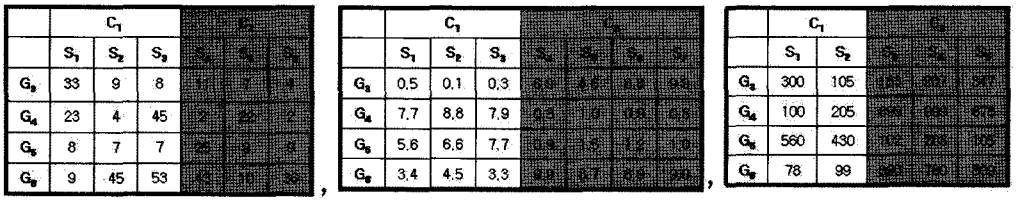


	C_1			C_2		
	S_1	S_2	S_3	S_4	S_5	S_6
G_1	3	5	7			
G_2	15	32	23			
G_3	33	9	8			
G_4	23	4	45			
G_5	8	7	7			
G_6	9	45	53			

	C_1			C_2		
	S_1	S_2	S_3	S_4	S_5	S_6
G_1	0.6	5.7	4.5			
G_2	5.7	7.8	8.9			
G_3	0.5	0.1	0.3			
G_4	7.7	8.8	7.9			
G_5	5.6	6.6	7.7			
G_6	3.4	4.5	3.3			
G_7	0.4	0.2	0.9			
G_8	5.7	4.3	5.6			
G_9	0.2	0.4	0.1			
G_{10}	9.9	8.9	5.6			
G_{11}	0.4	0.5	0.8			

	C_1		C_2		
	S_1	S_2	S_3	S_4	S_5
G_3	300	105			
G_4	100	205			
G_5	560	430			
G_6	78	99			
G_7	101	201			
G_8	500	550			
G_9	998	890			

그림 4 독립적으로 수행된 마이크로어레이 데이타



	C_1			C_2		
	S_1	S_2	S_3	S_4	S_5	S_6
G_3	33	9	8			
G_4	23	4	45			
G_5	8	7	7			
G_6	9	45	53			

	C_1			C_2		
	S_1	S_2	S_3	S_4	S_5	S_6
G_3	0.5	0.1	0.3			
G_4	7.7	8.8	7.9			
G_5	5.6	6.6	7.7			
G_6	3.4	4.5	3.3			

	C_1		C_2		
	S_1	S_2	S_3	S_4	S_5
G_3	300	105			
G_4	100	205			
G_5	560	430			
G_6	78	99			

그림 5 마이크로어레이 데이타 간 공통 유전자만 추출

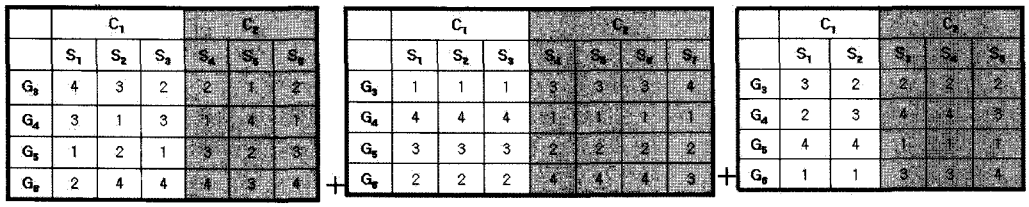


그림 6 각 샘플 내 순위로 표현된 마이크로어레이 데이타

순위로 치환된 값을 각 유전자에 대해 크기순으로 정렬한 후(클래스 레이블도 함께 정렬된다.) 모든 샘플에 대해 클래스 레이블을 0 또는 1의 값으로 치환한다. 예를 들어 정상 클래스 샘플을 0, 암 클래스 샘플을 1로 치환한다. 그리고 이렇게 얻어진 이진 문자열과 초기의 이진 문자열(정렬하기 이전에 클래스 레이블을 이진 문자열로 치환한 것)과의 거리를, 점수 함수를 이용하여 구한다. 점수 함수는 해당 유전자가 인포머티브 유전자가 될 가능성에 대한 척도로서, 정렬 후 이진 문자열이 정렬하기 이전의 이진 문자열로 변환되기 위해 필요한 연속된 0과 1의 교환 횟수로 정의된다. 표 1은 초기의 이진 문자열 "000111"과 정렬 후 얻어진 이진 문자열 "011001"과의 점수를 예로 보인 것으로 여기서 점수는 4로 계산된다. 점수 함수를 통해 얻어진 값이 아주 크거나 아주 작은 유전자가 인포머티브 유전자로 선택될 수 있다.

표 2는 인포머티브 유전자 추출을 위한 알고리즘을 나타낸다. 알고리즘의 이해를 돕기 위해 아래 표 3과 같은 마이크로어레이 데이타가 있다고 가정해 보자. 이를 각 샘플 내에서의 순위를 기반으로 변환한 것(알고리즘 단계 2)이 표 4이다. 그리고 각 유전자 별로 크기순으로 정렬한 것이 표 5이고 클래스 레이블을 이진 문자열로 변환한 것(알고리즘 단계 4)이 표 6이다.

알고리즘 단계 1을 수행하게 되면 초기의 이진 문자열 S를 "000111"로 얻을 수 있다. 이를 연속된 0과 1의 교환 횟수로 정의한 함수를 이용하여 표 6의 각 유전자

표 1 점수 함수의 예

점수 (score)	이진 문자열 (binary sequence)						교환 위치 (positions swapped)
	0	1	1	0	0	1	
+1	0	1	**0**	**1**	0	1	3 과 4
+1	0	**0**	**1**	1	0	1	2 와 3
+1	0	0	1	**0**	**1**	1	4 와 5
+1	0	0	**0**	**1**	1	1	3 과 4

표 3 발현 값으로 표현된 데이타

	정상	정상	정상	암	암	암
G_1	13	32	3	24	13	42
G_2	25	12	26	3	1	2
G_3	23	6	2	102	59	13
G_4	7	20	63	4	7	27

표 4 순위로 표현된 데이타

	정상	정상	정상	암	암	암
G_1	2	4	2	3	3	4
G_2	4	2	3	1	1	1
G_3	3	1	1	4	4	2
G_4	1	3	4	2	2	3

표 5 정렬된 데이타와 클래스 레이블

	정상/암	정상/암	정상/암	정상/암	정상/암	정상/암
G_1	2 (정상)	2 (정상)	3 (암)	3 (암)	4 (정상)	4 (암)
G_2	1 (암)	1 (암)	1 (암)	2 (정상)	3 (정상)	4 (정상)
G_3	1 (정상)	1 (정상)	2 (암)	3 (정상)	4 (암)	4 (암)
G_4	1 (정상)	2 (암)	2 (암)	3 (정상)	3 (암)	4 (정상)

표 2 인포머티브 유전자 추출 알고리즘

Input:	NI (인포머티브 유전자의 개수), V[][] (발현 값)
Output:	IG[][] (인포머티브 유전자 정보)
1:	정상 샘플은 0으로 암 샘플은 1로 치환한 이진 문자열 S를 생성한다.
2:	모든 i,j에 대하여 발현 값으로 표현된 V[G_i][S_j]를 각 샘플 내에서 발현 값으로 정렬하였을 때의 순위인 R[G_i][S_j]로 대체한다.
3:	선택되지 않은 유전자 중 임의의 한 유전자 G_i를 선택한다.
4:	모든 j에 대해 R[G_i][S_j]을 오름차순으로 정렬한 후 정상 샘플은 0으로 암 샘플은 1로 치환한 이진 문자열 T를 생성한다.
5:	연속된 0과 1의 교환 횟수로 정의된 점수 함수를 이용하여 S와 T의 점수를 계산한 후 우선순위 큐에 삽입한다.
6:	모든 유전자가 선택될 때까지 단계 3으로 돌아가 계속 수행한다.
7:	우선순위 큐에서 상위 NI/2개, 하위 NI/2개의 인포머티브 유전자를 추출한다.

표 6 이진 문자열로 표현된 데이타

G_1	0	0	1	1	0	1
G_2	1	1	1	0	0	0
G_3	0	0	1	0	1	1
G_4	0	1	1	0	1	1

의 이진 문자열에 대하여 계산해 보면 0과 1의 교환의 횟수가 가장 적은 유전자는 G_3로 총 1회이고 가장 많은 유전자는 G_2로 9회이다. 이는 G_2와 G_3가 G_1이나 G_4보다 인포머티브 유전자가 될 가능성이 높다는 것을 의미한다. 점수가 낮은 G_3 유전자는 해당 질병에 대해 긍정적인 상관관계(positive correlation), 즉 암 세포에 대해 발현을 많이 하는 유전자로 설명될 수 있으며 점수가 큰 G_2 유전자는 부정적인 상관관계(negative correlation), 즉 정상 세포에 대해 발현을 많이 하는 유전자로 설명될 수 있다.

모든 발현 값을 샘플 내에서의 순위로 치환을 하였기 때문에 이를 정렬하는 데 있어 같은 순위 값이 발생할 수 있다. 만일 동일한 클래스를 가진 샘플 사이에 같은 순위 값이 발생하였을 경우에는 점수에 영향을 미치지 않는다. 그러나 서로 다른 클래스를 가진 샘플 사이에 같은 순위 값이 발생하였을 경우에는 어떤 클래스를 우선순위에 두느냐에 따라 점수가 다를 수 있다. 위의 예제의 경우 초기의 이진 문자열 S가 "000111"이기 때문에 암 샘플을 정상 샘플보다 우선시 한다면 그렇지 않을 경우보다 점수가 더 크게 나오게 된다. 본 논문에서는 정상 샘플에 암 샘플보다 더 높은 우선순위를 주었다. 그러나 우선순위에 따른 이러한 점수의 차이가 인포머티브 유전자를 구별하는 데에는 크게 영향을 미치지 않는다. 왜냐하면 서로 다른 클래스를 가진 샘플들 사이에 동일한 순위 값이 발생하였다면 이는 해당 유전자가 인포머티브 유전자가 될 가능성이 적은 유전자임을 의미하기 때문이다.

3.3 k-TST (k-Top Scoring Triple) 분류 방법

본 연구에서는 암과 정상 샘플을 구분하는 분류자의 신뢰성을 높이기 위하여 규칙에 참여하는 유전자의 수를 일반화하는 것이 목표이다. 이러한 목표의 첫 단계로서 기존의 k-TSP 방식을 확장한 k-TST를 제안한다.

k-TST에서는 분류규칙에 참여하는 유전자의 수를 세 개로 한정한다. 세 개의 유전자의 순위 값의 크기를 비교하여 아래의 표 9와 같이 R_1 ... R_6의 6개의 관계식을 확립하고, 각 관계식에 대하여 발생한 빈도 수를 계산한다. 점수는 하나의 관계식이 클래스 1에서 일어난 확률과 클래스 2에서 일어난 확률의 차로 계산된다. 이 확률의 차가 가장 큰 관계식의 세 개의 유전자조합이 클래스를 가장 잘 분류다고 할 수 있다. 각 관계식의 우세한 클래스 레이블은 두 확률 값을 비교하여 큰 쪽의 클래스로 정의된다. 본 논문은 모든 세 개의 유전자 조합을 대상으로 한다. 각 조합에 대하여 6개의 관계식 모두에 대하여 위의 점수를 계산하여, 우선순위 큐에 정렬한다. 이 중에서 k개의 상위 관계식이 본 논문의 분류자가 된다. 따라서 분류자는 k개의 분류 규칙으로 이루어져 있으며 각각의 분류규칙은 1) 세 개의 유전자 집합, 2) 세 개 유전자간의 크기를 비교한 관계식, 3) 그 관계식의 우세한 클래스레이블 로 이루어져 있다. k-TST에 대한 알고리즘은 표 7과 같다. 그리고 표 7의 단계 3에서 사용한 점수 함수 (scoring function) 의 정의는 표 8과 같다.

예를 들어 표 9와 같은 데이타가 있다고 가정해 보자. C_1과 C_2는 각각 정상 또는 암 클래스를 의미하고 X는 유전자를, R_i는 세 개의 유전자 조합의 대소 관계를 의미한다. 여기서 해당되는 모든 Δ의 값을 구해 보면 R_3

표 8 점수 함수 정의

$P_{ijk}(1)$	Class 1에서 $X_i < X_j < X_k$ 인 관계가 나타날 확률 (X_i, X_j, X_k는 각 샘플 내에서의 순위 값을 나타냄)
Δ_{ijk}	$\lvert P_{ijk}(1) - P_{ijk}(2) \rvert$

표 9 TST 예제

	R_1	R_2	R_3	R_4	R_5	R_6	Total
C_1	2	1	29	4	2	4	42
C_2	4	5	1	14	8	1	33

(R의 정의) R_1: $X_i < X_j < X_k$ $R2$: $X_i < X_k < X_j$

$R3$: $X_j < X_i < X_k$ $R4$: $X_j < X_k < X_i$

$R5$: $X_k < X_i < X_j$ $R6$: $X_k < X_j < X_i$

표 7 k-TST 분류 방법 알고리즘

Input:	K (원하는 규칙의 개수), IS[][] (인포머티브 유전자)
Output:	K개의 분류 규칙 집합
1:	인포머티브 유전자 중 이전에 선택되지 않은 세 개의 유전자 조합을 선택한다.
2:	모든 샘플에 대해 세 유전자의 발현 값 간의 대소 관계를 판별한다.
3:	정의된 점수 함수를 이용하여 각 분류 규칙의 점수를 계산한다.
4:	계산된 점수와 대소 관계, 유전자 조합, 우세한 클래스 레이블로 이루어 진 규칙을 크기 K인 우선순위 큐에 삽입한다.
5:	가능한 세 개의 유전자 조합이 남아 있다면 단계 1로 돌아가 다시 수행한다.
6:	우선순위 큐에서 상위 K개의 규칙을 추출한다.

($X_j < X_i < X_k$) 관계일 경우에,

$$\Delta_{jik} = |P_{jik}(1) - P_{jik}(2)| = |\frac{29}{42} - \frac{1}{33}| \approx 0.66$$

가 가장 큰 점수를 가지는 것을 알 수 있다. 이는 유전자 X_i, X_j, X_k에 대해서 $X_j < X_i < X_k$ 의 규칙이 나올 가능성이 가장 높다는 것을 의미한다. k-TST 알고리즘은 모든 인포머티브 유전자 X_i, X_j, X_k 조합에 대하여 Δ_{ijk}를 구하여 이 값이 큰 K개의 규칙을 반환한다. 위의 예제의 경우 규칙은 (점수: 0.66, 대소 관계: $X_j < X_i < X_k$, 유전자 조합: X_i, X_j, X_k, 우세한 클래스 레이블: C_1) 로 표현될 수 있다. 만일 테스트 샘플의 클래스 레이블을 판별하고자 할 때, 위의 규칙에 기반 한다면 X_i, X_j, X_k 유전자의 값을 확인하여 $X_j < X_i < X_k$ 의 관계를 만족시킨다면 테스트 샘플의 클래스를 C_1으로 판별하게 된다. 실제로는 이러한 규칙의 개수가 k개이기 때문에 과반수 투표(majority voting)에 따라 테스트 샘플의 클래스를 판별한다. 과반수 투표의 과정은 아래 수식과 같다.

- r_i *is the* i^{th} *rule*
- S *is a test sample*
- k *is the number of rules*
- NC *is the number of normal count*

$L(r_i)$ = *Class Label of the* r_i ($L(r_i) = \{normal, t$

$$P_{ri}(S) = \begin{cases} L(r_i) & \text{if } S \text{ satisfies the } r_i \\ \overline{L(r_i)} & \text{Otherwise} \end{cases}$$

$$V(r_i) = \begin{cases} 1 & \text{if } P_{ri}(S) \text{ is a Normal Sample} \\ 0 & \text{Otherwise} \end{cases}$$

$$NC = \sum_{i=1}^{k} V(r_i)$$

위 수식에서 $L(r_i)$는 규칙 r_i의 우세한 클래스 레이블을 의미하고 $P_{ri}(S)$는 테스트 샘플 S가 규칙 r_i를 만족할 경우에는 규칙 r_i의 우세한 클래스 레이블로, 그렇지 않을 경우에는 규칙 r_i의 우세한 클래스 레이블과 반대의 클래스 레이블로 정의된다. $L(r_i)$는 normal과 tumor만을 원소로 가지고 있기 때문에 테스트 샘플의 클래스는 규칙 r_i의 우세한 클래스 레이블이 normal이고 테스트 샘플 S가 규칙을 만족한다면 normal로, 만족하지 않을 경우 tumor로 정의되게 된다. 그리고 $V(r_i)$는 규칙 r_i의 판정 결과가 정상 샘플일 경우에는 1로, 암 샘플일 경우에는 0으로 정의된다. NC는 최종적인 판정을 하기 위해 정상 샘플이라고 판정한 횟수를 모두 누적하게 된다. 만일 NC의 값이 $\frac{k}{2}$보다 크면 테스트 샘플의 클래스를 정상으로 판정하게 되고 그렇지 않을 경우에는 암으로 판정하게 된다. 규칙의 개수 k를 홀수로 고정하였기 때문에 동점이 발생하지 않고 항상 테스트 샘플의 클래스를 판정할 수 있다.

4. 실험 결과

본 장에서는 논문이 제안하는 두 단계 마이크로어레이 분석 기법의 정확도와 효율성을 검증하기 위한 실험을 기술한다. 본 실험에서는 논문을 통하여 공개된 데이타가 비교적 많은 전립선암(Prostate Cancer) 마이크로어레이 데이타를 사용하였다. 실험에 사용된 마이크로어레이 플랫폼은 Affymetrix사의 HG_95AV2이다. 아래 실험에서 사용된 데이타는 각 실험자의 이름인 Singh [19], Welsh[20], Latulippe[21]로 표기한다.

4.1 LOOCV를 통한 최적 규칙의 개수(K) 결정

본 절에서는 LOOCV를 이용해 규칙의 개수인 K를 변경해 가며 정확도가 가장 높은 최적의 K의 값을 얻는 실험에 대해 기술한다. 보통 인포머티브 유전자의 개수는 50에서 200개로 선택하는 것이 일반적이므로[10,22], 본 실험에서는 총 유전자의 개수인 12600개의 1%인 126개를 선택하였다. 아래 그림 7에서 그림 12까지는 K의 최대값을 10으로 설정하였을 때 K를 변경시켜 가며 각 마이크로어레이 데이타에 대한 LOOCV를 실험한 그래프이고 표 11은 실험 결과 얻은 최적의 K의 값들이다. 여기서 비교 척도로서 정확도(Accuracy)와 민감도(Sensitivity) 그리고 특이성(Specificity)을 사용하였다. 이를 수식으로 나타내면 아래와 같다.

$$Accuracy = \frac{The\ Nmber\ of\ Correctly\ Predicted\ Samples}{The\ Nmber\ of\ Samples}$$

$$Sensitivity = \frac{The\ Nmber\ of\ Correctly\ Predicted\ Cancer\ Samples}{The\ Nmber\ of\ Cancer\ Samples}$$

$$Specificity = \frac{The\ Nmber\ of\ Correctly\ Predicted\ Normal\ Samples}{The\ Nmber\ of\ Normal\ Samples}$$

표 10 전립선암(prostate cancer) 마이크로어레이 데이타

데이타	프로브 세트 (유전자 개수)	정상 샘플의 수	암 샘플의 수	총 샘플의 수
Singh	12600	50	52	102
Welsh	12626	9	24	33
LaTulippe	12626	3	23	26

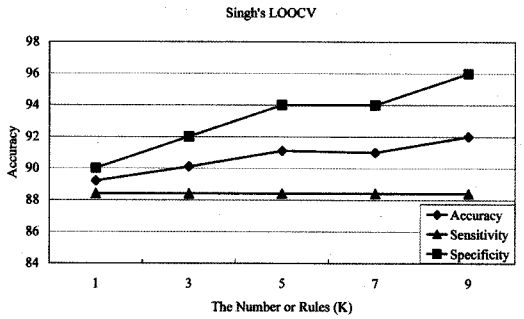


그림 7 Singh 데이타로 LOOCV

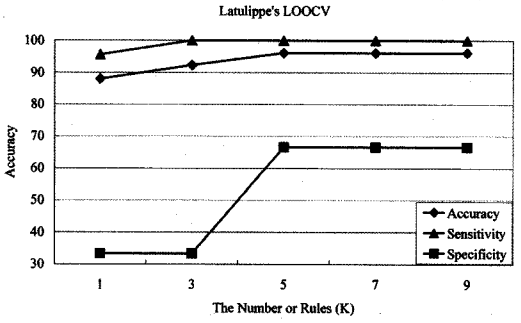


그림 8 Latulippe 데이타로 LOOCV

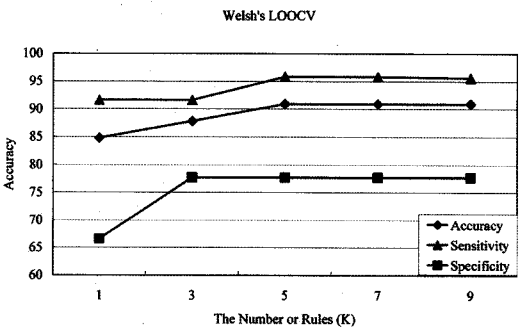


그림 9 Welsh 데이타로 LOOCV

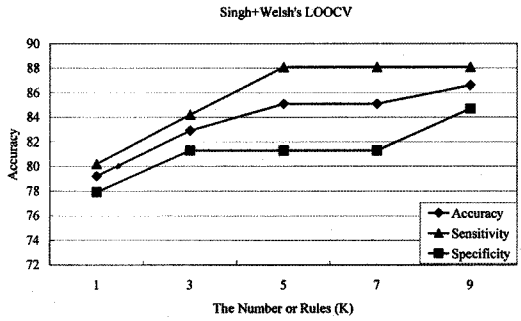


그림 10 Singh과 Welsh의 통합 데이타로 LOOCV

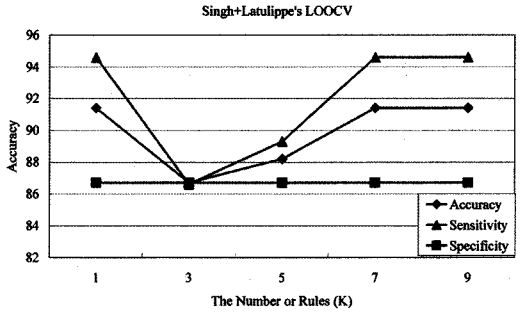


그림 11 Singh과 Latulipple의 통합 데이타로 LOOCV

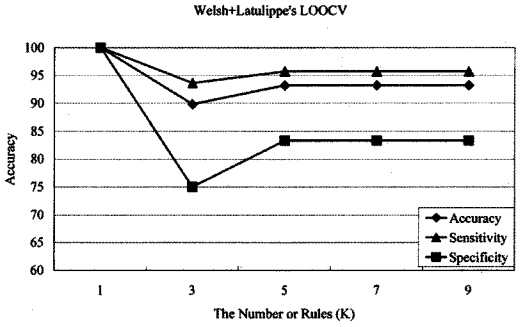


그림 12 Welsh와 Latulippe의 통합 데이타로 LOOCV

표 11 LOOCV를 통해 얻어진 최적 K

Traing Data	최적 K
Singh	9
Latulippe	5
Welsh	5
Singh + Welsh	9
Singh + Latulippe	7
Welsh + Latulippe	5

본 실험에서는 규칙의 개수를 최소 5개 이상으로 한정하였다. 왜냐하면 규칙의 개수가 너무 적을 경우 분류자의 신뢰성을 보장하기 어렵기 때문에 실제 다른 환경에서 얻어진 독립적인 데이타를 테스트 데이타로 하여 실험하였을 경우 정확도가 떨어질 수 있기 때문이다.

4.2 인포머티브 유전자 추출 방법의 정확도 비교

본 절에서는 논문에서 제안한 인포머티브 유전자의 추출 방법에 대한 정확도를 비교 분석한다. 비교 대상으로는 대표적인 인포머티브 유전자 추출 방법인 인포메이션 게인(Information Gain)과 릴리프 에프(Relief-F)를 사용하였다. 이 때 두 방법 모두 통합된 데이타에 그대로 적용하는 것이 불가능하기 때문에 가장 일반적인 정규화 방법인 Z-score를 사용하여 통합하였다. 그리고 두 번째 단계인 클래스 분류 방법의 비교 대상으로는

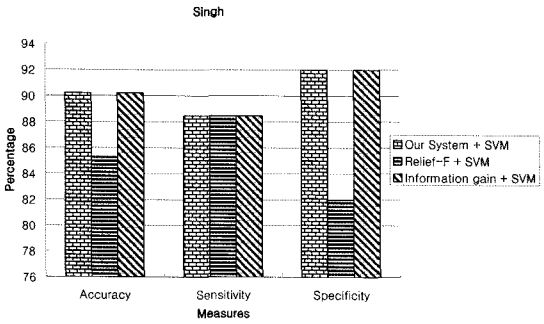


그림 13 Singh 데이타에 대한 LOOCV 결과

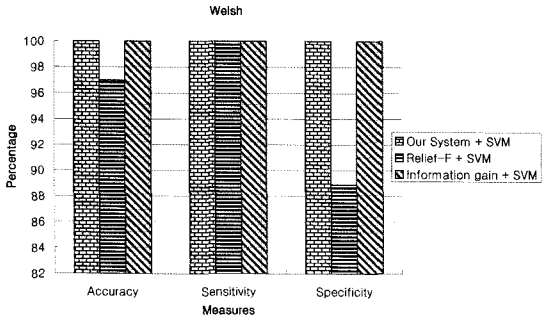


그림 14 Welsh 데이타에 대한 LOOCV 결과

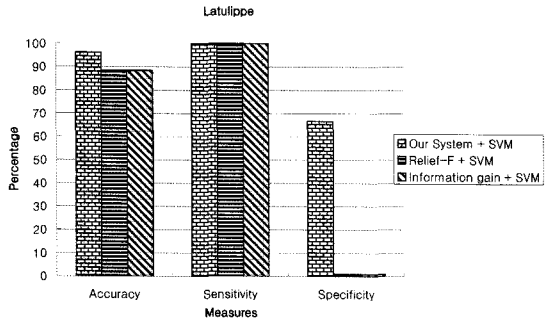


그림 15 Latulippe 데이타에 대한 LOOCV 결과

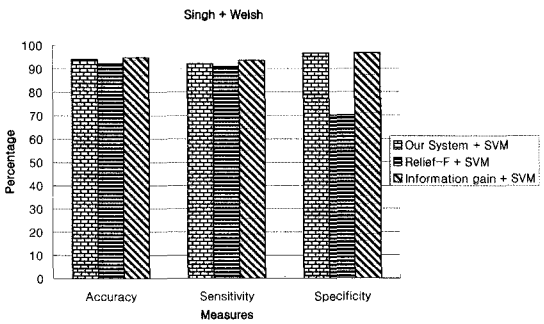


그림 16 Singh과 Welsh 통합 데이타에 대한 LOOCV 결과

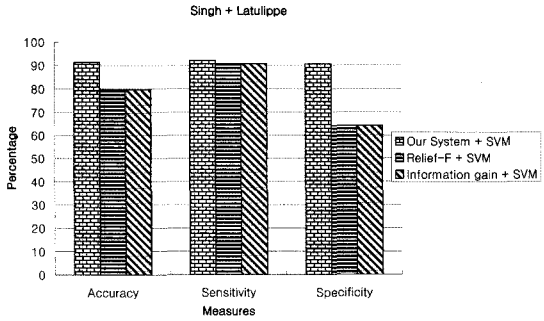


그림 17 Singh과 Latulippe 통합 데이타에 대한 LOOCV 결과

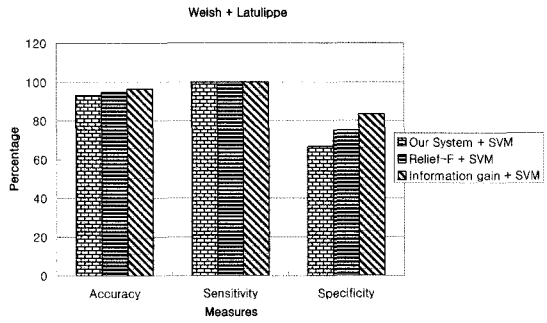


그림 18 Welsh와 Latulippe 통합 데이타에 대한 LOOCV 결과

현재 패턴 인식 분야에서 가장 좋은 정확도를 나타내고 있는 SVM(Support Vector Machine)을 사용하였다. 그림 13에서 그림 18은 각 데이타 세트에 대한 LOOCV의 정확도를 나타내는 그래프이다.

위의 그래프들을 보면 논문에서 제안한 인포머티브 추출 방법이 다른 방법에 비해 동일하거나 더 좋게 나옴을 알 수 있다. Singh과 Welsh의 단일 데이타에 대한 LOOCV의 정확도(그림 13,14)는 인포메이션 게인을 사용한 경우와 동일하게 나왔으며 Latulippe 데이타(그림 15)에 대해서는 더 좋은 성능을 나타내었다. 그리고 릴리프 에프보다는 대부분 좋은 정확도를 보임을 알 수 있다. 특히 Singh과 Latulippe의 통합 데이타(그림 17)에 대한 LOOCV의 정확도는 다른 방법들에 비해 10% 이상의 높은 정확도를 나타내고 있다. 이러한 결과로 비추어 볼 때, 본 논문에서 제안하는 인포머티브 유전자 추출 방법이 다른 추출 방법들에 필적하거나 좀 더 좋은 성능을 보인다는 것을 알 수 있다.

아래 그림 19에서 그림 21은 세 개의 데이타 중 하나를 테스트 데이타로 사용하고 나머지 데이타의 조합을 학습데이타로 사용하여 정확도를 실험한 그래프이다. 예

를 들어 Singh을 테스트데이타로 사용할 경우, 학습데이타 세트는 Welsh, Latulippe, Welsh+Latullipe가 된다. 본 알고리즘은 각 학습데이타로부터 분류자를 얻어서 Singh데이타에 적용하여 정확도를 측정, 비교하였다. 그리고 Singh에 대하여 분류자를 찾는 세 가지 방법 1) 본 논문에서 제안한 인포머티브 유전자 추출 방법 + SVM, 2) 릴리프 에프 + SVM, 3) 인포메이션 게인 + SVM의 정확도를 비교하였다.

독립적인 데이타로 테스트를 수행한 위의 그래프들을 보면, LOOCV 실험에서와 마찬가지로 논문에서 제안한 인포머티브 유전자 추출 방법이 인포메이션 게인과 동일한 성능을 보이거나 조금 더 좋은 성능을 보였다. 인포메이션 게인 방법이 통합 데이타에 대해 항상 더 좋은 정확도를 보이지 않는 반면 (그림 19), 본 논문에서 제안한 방법은 데이타를 통합하여 샘플의 수를 증가 시킬수록 정확도가 좋아지는 것을 알 수 있다.

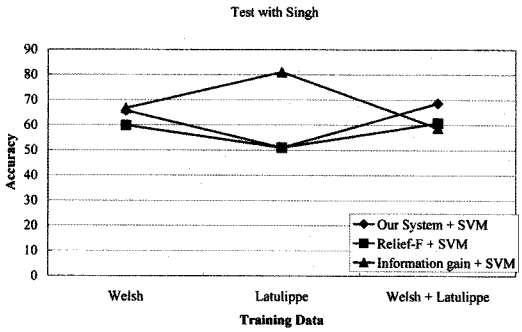


그림 19 Singh을 테스트 데이타로 사용

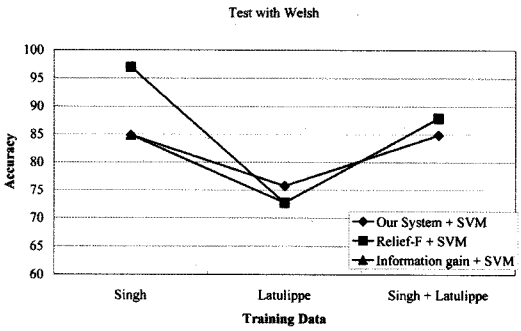


그림 20 Welsh를 테스트 데이타로 사용

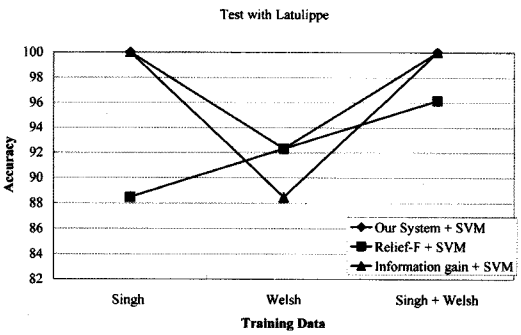


그림 21 Latulippe를 테스트 데이타로 사용

4.3 클래스 분류 방법의 정확도 비교

본 절에서는 4.1절에서 구한 각 학습 마이크로어레이 데이타에 대한 최적의 K개의 규칙을 이용하여 독립적인 데이타에 적용하였을 때 제안하는 두 단계 접근 시스템의 정확도를 기존의 k-TSP 방법과 4.2절에서 비교 실험한 기존의 방법 중 더 좋은 성능을 보인 인포메이션 게인을 적용한 후 SVM을 적용하는 방법과 비교 실험한다. 그림 22에서 그림 24는 Singh, Welsh, Latulippe를 각각 독립적인 테스트 데이타로 이용하고 그 이외의 마이크로어레이 데이타의 조합으로 학습하여 규칙을 만들어 테스트한 정확도 그래프이다. 표 12는 실험에 사용된 K의 값을 나타낸다.

위의 세 그래프를 보면 알 수 있듯이 단일 마이크로어레이 데이타에 대한 테스트 결과는 k-TST가 다른 방법들에 비해 항상 좋은 결과를 내지는 못한다. 그러나 단일 마이크로어레이 데이타는 학습 데이타로 사용하기에 샘플의 수가 충분치 않다. 특히 Welsh, Latulippe에서는 샘플의 수가 약 30개 내외이고 정상 샘플과 암

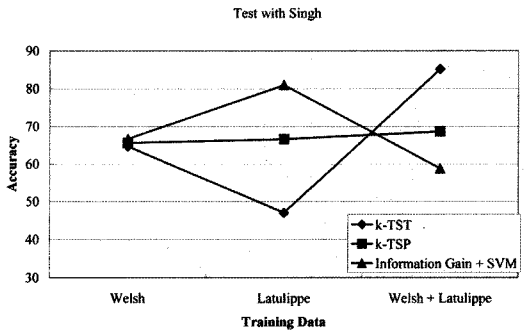


그림 22 Singh을 테스트 데이타로 사용

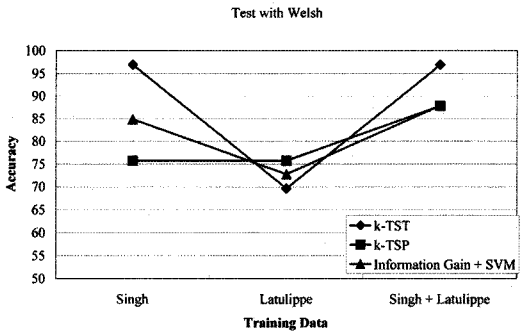


그림 23 Welsh를 테스트 데이타로 사용

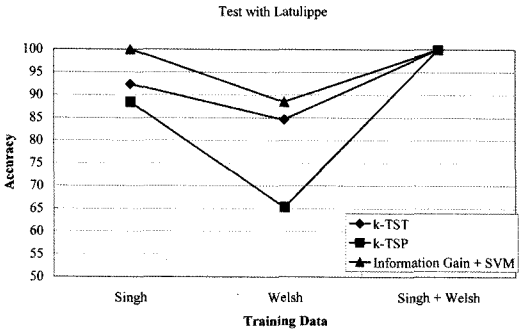


그림 24 Latulippe를 테스트 데이타로 사용

표 12 실험에 사용된 규칙의 개수(K)

Test data	Training data	k-TSP	k-TST
Singh	Welsh	3	5
	Latulippe	3	5
	Welsh + Latulippe	1	5
Welsh	Singh	1	9
	Latulippe	3	5
	Singh + Latulippe	5	7
Latulippe	Singh	1	9
	Welsh	3	5
	Singh + Welsh	9	9

샘플 수의 분포가 비대칭적이기 때문에 이를 학습 데이타로 쓰는 것은 신뢰할 만한 규칙을 만들어 내지 못할 가능성이 있다. 그러나 학습 데이타의 샘플의 수가 늘어날수록, 즉 단일 마이크로어레이 데이타를 통합할수록, k-TST의 정확도가 k-TSP나 SVM에 비해 15%에서 최대 24.19% 좋아짐(그림 22, 23)을 알 수 있다. 위의 실험들을 근거로 하였을 때 본 논문에서 제안한 두 단계 접근법이 단일 마이크로어레이 데이타를 통합하면 할수록, 신뢰도 높은 규칙을 만들어 내는 것을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 독립적으로 생성된 마이크로어레이 데이타를 통합하고 인포머티브 유전자를 찾은 후, 클래스 분류 규칙을 생성하는 두 단계 접근 방법을 소개하였다. 본 논문의 주 공헌은 다음과 같다.

마이크로어레이 데이타의 양이 폭발적으로 늘고 있는 상황에서 동일한 생물학적 주제로 다른 연구기관에서 독립적으로 시행된 데이타를 통합할 수 있는 새로운 방법을 제 1 단계에 적용하였다. 데이타의 통합으로 학습 데이타의 샘플수가 증가되었기 때문에 높은 정확도를 보이는 분류자를 찾을 수 있었다. 또한 두 단계 접근을 취함으로서 제 2 단계에서는 1단계에서 추출된 인포머티브 유전자만을 사용함으로써 계산량을 획기적으로 줄일 수 있었다. 그리고 본 논문에서 제시한 클래스 분류 방법은 관련 유전자의 개수가 비교적 적고 해석이 용이하기 때문에 실질적으로 암 진단키트에 사용될 경우 진단비용을 크게 낮출 수 있다. 왜냐하면 수 만개의 유전자를 이용하는 마이크로어레이 실험을 매 진단에 사용하는 것을 비용 면에서 비효율적이기 때문이다.

본 논문에서 제시한 알고리즘에 대한 시스템 구현이 완성되어 진행된 실험결과를 보면, 본 논문의 인포머티브 유전자 추출방법이 다른 방법에 비해 우수하거나 필적함을 알 수 있고, 독립 마이크로어레이 데이타를 통합하여 학습데이타의 샘플 수가 늘어날수록, 본 논문의 분류자의 정확도가 다른 방법에 비해 월등히 좋아짐을 알 수 있었다. 위의 실험들을 근거로 하였을 때 본 논문에서 제안한 두 단계 접근법이 단일 마이크로어레이 데이타를 통합하면 할수록, 신뢰도 높은 규칙을 만들어 내는 것을 알 수 있었다.

현재 향후 연구로 1) Affymetrix사의 플랫폼 뿐만 아니라 다른 기준으로 발현 값을 표현하는 cDNA 플랫폼과 연동하는 방법 2) 상관관계 분석 또는 클러스터링 기법을 활용하여, 관련 유전자들 간의 존재하는 중복(redundancy)을 없애는 방법 3) 클래스 분류 규칙의 개수, 각 규칙에 참여하는 유전자의 개수를 좀 더 일반화하는 방법 등이 있다.

참 고 문 헌

[1] S. Mukherjee, P. Tamayo, S. Rogers, R. Rifkin, A. Engle, C. Campbell, T. Golub and J. Mesirov, "Estimating dataset size requirements for classifying DNA microarray data", Journal of Computational Biology, vol. 10, pp. 119-142, 2003.

[2] L. Xu, A. Tan, D. Naiman, D. Geman and R. Winslow, "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data", Bioinformatics, vol. 21, pp. 3905-3911, 2005.

[3] J. K. Choi, U. Yu, S. Kim and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation", Bioinformatics, vol. 19, pp. 84-90, 2003.

[4] H. Jiang, Y. Deng, H.S. Chen, L. Tao, Q. Sha and J. Chen, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes," BMC Bioinformatics, vol. 5, pp. 81-92, 2004.

[5] J. Kang, J. Yang, W. Xu, and P. Chopra, "Integrating heterogeneous microarray data sources using correlation signatures," In International Workshop on Data Integration in the Life Sciences (DILS), 2005.

[6] S. Dudoit and J. Fridlyand, "Classication in microarray experiments," Statistical Analysis of Gene Expression Microarray Data, Chapman and Hall, 2003.
[7] C. Tang, A. Zhang and J. Pei, "Mining Phenotypes and Informative Genes from Gene Expression Data," ACM SIGKDD, pp. 24-27, Washington, DC, USA, August 2003.
[8] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer and Z. Yakhini, "Tissue classification with gene expression profiles," Journal of Computational Biology, vol. 7, pp. 559-583, 2000.
[9] C. Bishop, "Neural networks for pattern recognition," Oxford University Press, New York, 1995.
[10] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Collier, M. L. Loh, J. R. Downing and M. A. Caligiuri, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531-537, 1999.
[11] P. J. Park, M. Pagano and M. Bonetti, "A nonparametric scoring algorithm for identifying informative genes from microarray data," Pacific Symposium on Biocomputing, pp. 52-63, 2001.
[12] I. H. Witten, and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," Morgan Kaufmann, 1999.
[13] M. Robnik-Sikonja, and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," Machine Learning, vol. 53, pp.23-69, 2003.
[14] N. Bailey, "Statistical methods in biology," Cambridge university press, 1995.
[15] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machine," 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm.
[16] V. Vapnik, "Statistical Learning Theory," John Wiley & Sons, New York, 1999.
[17] B. Dasarathy, "Nearest Neighbor Norms: NN Pattern Classification Techniques," IEEE Computer Society Press, Los Alamitos, CA, USA. 1991.
[18] A. Tan, D. Naiman, L. Xu, R. Winslow and D. Geman. "Simple decision rules for classifying human cancers from gene expression profiles," Bioinformatics, vol. 21, pp. 3896-3904, 2005.
[19] D. Singh, P. G. Febbo, K. Ross,D. G. Jackson, J. Manola and C. Ladd, "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell, vol. 1, pp. 203-209, 2002.
[20] J. B. Welsh, L. M. Sapinoso, A. I. Su, S. G. Kern, J. Wang-Rodriguez and C. A. Moskaluk, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer", Cancer Res., vol. 61, pp. 5974-5978, 2001.
[21] E. LaTulippe, J. Satagopan, A. Smith, H. Scher, P. Scardino and V. Reuter, "Comprehensive gene expression analysis of prostate cancer reveals distinct transcriptional programs associated with metastatic disease," Cancer Res, vol. 62, pp. 4499-4506, 2002.
[22] L. Li, W. Leping, C. R. Weinberg, T. A. Darden and L. G. Pedersen, "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, pp. 1131-1142, 2001.

윤 영 미

1981년 2월 서울대학교 자연과학대학 졸업(학사). 1981년 6월~1983년 6월 오하이오 주립대학 수학과(학사수료). 1987년 3월 스탠포드대학교 컴퓨터과학과 졸업(석사). 1987년 5월~1993년 5월 IntelliGenetics Inc., Mountainview, California, Software Engineer. 1995년 2월~현재 가천의과학대학교 부교수. 2005년 3월~현재 연세대학교 컴퓨터과학과 박사과정. 관심분야는 데이타베이스 시스템, 데이타 마이닝, 바이오인포매틱스

이 종 찬

2004년 8월 연세대학교 컴퓨터과학과 졸업(학사). 2006년 8월 연세대학교 컴퓨터과학과 대학원 졸업(석사). 관심분야는 위치기반서비스, ITS, 데이타베이스 시스템, 데이타 마이닝, 바이오인포메틱스

박 상 현

1989년 2월 서울대학교 컴퓨터공학과 졸업(학사). 1991년 2월 서울대학교 컴퓨터공학과 졸업(석사). 2001년 2월 UCLA대학교 전산학과 졸업(박사). 1991년 3월~1996년 8월 대우통신 연구원. 2001년 2월~2002년 6월 IBM T. J Watson Research Center Post-Doctoral Fellow. 2002년 8월~2003년 8월 포항공과대학교 컴퓨터공학과 조교수. 2003년 9월~2006년 8월 연세대학교 컴퓨터과학과 조교수. 2006년 9월~현재 연세대학교 컴퓨터과학과 부교수. 관심분야는 데이타베이스 보안, 데이타 마이닝, 바이오인포매틱스, XML 등