유전자 발현값 상관관계 분석을 통한 암분류자 생성방법

안재균*, 윤영미**, 신은지*, 박상현*
*연세대학교 컴퓨터과학과
**가천의과학대학교 IT 학과
e-mail : ajk@cs.yonsei.ac.kr

Tumor Classifier using Variation in Genes' Correlation

Jaegyoon Ahn*, Youngmi Yoon**, Eunji Shin*, Sanghyun Park
*Dept. of Computer Science, Yonsei University
**Dept. of Information Technology, Gachon University of Medicine and Science

요 약

본 논문에서는 이상 표식 유전자를 사용하는 기존 분석방법과 달리, 두 유전자 사이의 관계를 측정하여 정상 클래스와 암 클래스에서의 상관관계가 변화된 정도를 분석하여 차이가 두드러지는 유전자 쌍(gene pair)을 질병 분류자(classifier)로 선택하는 방법을 제시한다. 제안한 암 분류 방법의 실험 결과, 소수의 분류자를 선택하여 높은 정확도로 암을 분류함으로써 그 유용성을 검증하였다.

1. 서론

DNA 마이크로어레이[1]는 생체 조직 샘플들로부터 수만 개의 유전자와 EST(Expressed Sequence Tag)의 발현 양상을 동시에 관찰할 수 있는 도구이다. 마이크로어레이를 이용해서 특정 암에 따라 다르게 발현되는 유전자 양상을 통계적인 방법으로 발견함으로써 암 분류를 수행하는 방법이 많이 제시되어 왔다 [2-6]. 이러한 암 분류 방법들은 샘플들 사이에서 발현량이 큰 차이를 보이는 유전자(marker gene)를 선택함으로써 보다 정확하고 효과적인 암분류를 수행하고 있다.

하지만 유전자의 발현량뿐만 아니라, 유전자와 유전자 사이의 상관관계의 변화 또한 질병을 진단함에 있어서 지표가 될 수 있다. 예를 들어, 어떤 전사 인자(transcription factor)가 두 유전자 A 와 B 를 동시에 활성화(activate) 또는 억제(suppress)시킨다면 A 와 B 의 발현량은 높은 상관관계를 보이나, 질병으로 인해 유전자 B 가 변이 된다면 이 전사 인자는 A 를 그대로 활성화시키나 B 에는 더 이상 영향을 미치지 못하므로 A 와 B 사이의 상관관계는 없어진다.

따라서 표지 유전자뿐만이 아니라, 두 유전자 사이의 관계를 측정하여 정상 클래스와 암 클래스에서의 상관관계가 변화된 정도를 분석했을 때, 차이가 두드러지는 표지 유전자 쌍은 좋은 분류 기준이 될 수 있다. 본 논문에서는 이러한 표지 유전자 쌍을 찾아내어 분류자로 선택하는 새로운 분류 방법을 제안하고, 그 효용성을 입증하기 위하여 전립선 암환자의 마이크로어레이를 대상으로 실험을 수행했다.

2. 암 분류자의 구성 및 분류 방법

두 속성의 상관관계를 알 수 있는 방법 중 하나는 Pearson's Correlation Coefficient 을 통해서 두 속성의 상관계수를 측정하는 것이다. 상관계수 r 은 -1 에서 1 사이의 값을 가진다. 통상적으로 $r \geq 0.7$ 인 경우 X 와 Y 는 양의 상관관계를 가지고, $r \leq -0.7$ 인 경우 X 와 Y 는 음의 상관관계를 가지며, $r = 0$ 이면, X 와 Y 사이의 상관관계는 없다. X 와 Y 가 양 혹은 음의 상관관계를 가질 때, 두 속성은 유의한 상관관계를 가진다고 말한다. 본 논문에서 속성 X 및 Y 는 임의의 두 유전자이다. 마이크로어레이의 두 샘플 집합인 정상 및 암 샘플 집합 각각에 대해서 임의의 두 유전자 사이의 상관계수 r 을 구하고 각각을 r_{normal}, r_{tumor} 라 할 때, 다음 3 가지 조건을 동시에 만족한다면 이 두 유전자를 암 특이적인 유전자 쌍이라고 한다.

1) $| |r_{normal}| - |r_{tumor}| | > 0.5$
2) $\min(|r_{normal}|, |r_{tumor}|) < 0.7$
3) $\max(|r_{normal}|, |r_{tumor}|) \geq 0.7$

즉, 두 유전자의 상관관계가 정상 샘플에 대해서는 유의하고 암 샘플에 대해서는 유의하지 않거나, 그 반대인 경우라면 두 유전자는 암과 정상 샘플을 잘 구분해 줄 수 있는 암 특이적인 유전자 쌍이 된다.
이와 같이 마이크로어레이의 모든 유전자 쌍에 대해서 이 두 상관계수를 구해서 이 유전자 쌍이 암 특이적인 유전자 쌍일 경우, 이 유전자 쌍을 우선 순위 큐(priority queue)에 집어넣는다. 이 때 우선 순위를 정하기 위해서 다음 식을 사용한다.

$$p = \frac{\left| |r_{normal}| - |r_{tumor}| \right|}{\min(|r_{normal}|, |r_{tumor}|)}$$

p 가 클수록 우선 순위는 높아진다. 그리고 두 상관계수의 절대값의 차이가 크면서, 두 상관계수의 절대값 중 작은 것이 0 에 가까울수록 p 는 커진다. 그 이유는 두 상관계수의 절대값의 차이가 같다고 할 때, 어