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

느 한 상관계수가 1 혹은 -1 에 가까워지는 것 보다는 0 에 가까워지는 경우가 정상과 암을 더 잘 구분할 수 있음을 실험적으로 관측했기 때문이다.
우선 순위 큐는 pq_{normal}, pq_{tumor} 의 두 종류를 사용한다. 만약 $|r_{normal}| > |r_{tumor}|$이라면 이 유전자 쌍은 pq_{normal}에 저장되고, 그렇지 않다면 pq_{tumor}에 저장된다.
새로운 샘플을 분류하기 위한 방법은 다음과 같다.

1) pq_{normal}, pq_{tumor} 각각에서 상위 k 개의 유전자 쌍을 뽑는다.
2) r_{normal} 및 r_{tumor} 를 재계산한 값을 각각 r'_{normal} 및 r'_{tumor} 이라고 할 때, 모든 유전자 쌍에 대해서 상관계수를 재계산함으로써 $\sum(|r'_{normal}|-|r_{normal}|)$ 및 $\sum(|r'_{tumor}|-|r_{tumor}|)$를 구한다.
3) $\sum(|r'_{normal}|-|r_{normal}|) > \sum(|r'_{tumor}|-|r_{tumor}|)$ 라면 해당 샘플은 정상 샘플이며, 그렇지 않은 경우 해당 샘플은 암 샘플로 분류된다.

3)에서 $\sum(|r'_{normal}|-|r_{normal}|) = \sum(|r'_{tumor}|-|r_{tumor}|)$인 경우 해당 샘플을 암 샘플로 분류하는 이유는, 정상 샘플을 암 샘플로 오분류하는 것보다는 암 샘플을 정상 샘플로 오분류하는 것이 더욱 위험하기 때문이다.

3. 실험결과

본 실험에서는 전립선 암 조직에 대한 12600 개의 유전자 발현량을 측정한 마이크로어레이 데이터인 Singh[7], Welsh[8] 및 LaTulippe[9]를 사용했다. Singh 은 50 개의 정상 샘플과 52 개의 암 샘플을, Welsh 는 9 개의 정상 샘플과 24 개의 암 샘플을, LaTulippe 는 3 개의 정상 샘플과 23 개의 암 샘플을 가진다. Welsh 및 LaTulippe 가 적은 수의 정상 샘플을 가지고 있기 때문에 Singh 를 트레이닝 데이터로, Welsh 와 LaTulippe 를 합친 데이터를 테스트 데이터로 사용했다. 서로 그 스케일이 다른 데이터를 합치기 위해서 각 샘플별로 Z-transform[10]을 이용해서 정규화했다.
비교에 사용된 다른 기계 학습 알고리즘은 SVM (Support Vector Machine)과 random forest, Naïve Bayesian Network, k-TSP (k-Top Scoring Pair)이다. 이중에서 SVM 과 random forest, Naïve Bayesian Network 는 Weka 3.6[11]을 통해서 실험했고 k-TSP 는 Tan[6]에서 제공하는 실행파일을 사용하였다.

<표 1> 비교 실험 결과(%)

	Accuracy*	Sensitivity**	Specificity***
제시한 알고리즘	**98.31**	**97.87**	**100**
SVM	93.22	93.62	91.67
Random Forest	93.22	93.62	91.67
Naïve Bayesian Network	94.91	95.74	91.67
k-TSP	81.36	93.62	33.33

* accurary = 정확히 분류된 샘플의 개수/전체 샘플의 개수
** sensifivity = 정확히 분류된 암 샘플의 개수/전체 암 샘플의 개수
*** specificity = 정확히 분류된 정상 샘플의 개수/전체 정상 샘플의 개수

표 1 에서 본 연구에서 제시한 알고리즘은 다른 비교 알고리즘에 비해서 높은 sensitivity, specificity 및 accuracy 로 암을 분류하는 것을 확인할 수 있다.

4. 결론

본 연구에서 제안하는 방법의 핵심은 정상 클래스와 암 클래스에서의 상관관계가 현격한 차이를 보이는 유전자 쌍의 수를 최소로 선택하여 유의미한 분류자를 생성하는 것이다. 이 분류자는 실험 결과 비교적 높은 분류 정확도를 보였다. 또한, 본 연구에서 제안하는 분류 방법으로 찾아낸 암과 관련이 있는 유전자 상호관계는 분류자로서의 역할 뿐만이 아니라, 질병 특이적인 유전자 조절 네트워크(gene regulatory network) 혹은 단백질 네트워크(protein network) 등을 구축하는 중요한 단서가 될 수 있다는 의의가 있다.

참고문헌

[1] Duggan. D. J, Bittner. M, Chen Y, Meltzer. P, Trent. J. M, "Expression profiling using cDNA microarrays," Nature Genetics Supplement, vol. 21, pp.10-14, 1999.
[2] Pirooznia. M, Yang. J.Y, Yang. M.Q, and Deng. Y, "A comparative study of different machine learning methods on microarray gene expression data", BMC Genomics, 2008.
[3] Wang. Y, Tetko. L.V, Hall. M.A, Frankb. E, Facius. A, Mayer. K.F.X, and Mewes. F.W, "Gene selection from microarray data for cancer classification—a machine learning approach", Computational Biology and Chemistry, pp. 37-45, 2005.
[4] Duan. K.B, Rajapakse. J.C, Wang. H, and Azuaje. F, "Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data", IEEE TRANSACTIONS ON NANOBIOSCIENCE, vol. 4, no. 3, pp. 228-234, 2005.
[5] Diaz-Uriarte R, Alvarez de Andres S, "Gene selection and classification of microarray data using random forest", BMC Bioinformatics, vol. 7, no.13, 2006.
[6] Tan. A, Naiman. D, Xu. L, Winslow. R, Geman. D, "Simple decision rules for classifying human Cancers from gene expression profiles," Bioinformatics, vol. 21, pp. 3896-3904, 2005.
[7] Singh. D, Febbo. P. G, Ross. K, Jackson. D. G, Manola. J, Ladd. C, "Gene expression correlates of clinical prostate Cancer behavior," Cancer Cell, vol. 1, pp. 203-209, 2002.
[8] Welsh. J. B, Sapinoso. L. M, Su. A. I, Kern. S. G, Wang-Rodriguez. J, Moskaluk. C. A, "Analysis of gene expression identifies candidate markers and pharmacological targets in prostate Cancer," Cancer Research, vol. 61, pp. 5974-5978, 2001.
[9] LaTulippe. E, Satagopan. J, Smith. A, Scher. H, Scardino. P, Reuter. V, "Comprehensive gene expression analysis of prostate Cancer reveals distinct transcriptional programs associated with metastatic disease," Cancer Research, vol. 62, pp. 4499-4506, 2002.
[10] David, F. N. "The Moments of the z and F Distributions." Biometrika 36, 394-403, 1949.
[11] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.