암 분류를 목적으로 하는 기계 학습 분류기를 위한 효과적인 유전자 선택 방법

박형근[O] 이수정 이일병
연세대학교 컴퓨터정보공학부
{acrolein[O], crystal, yblee}@csai.yonsei.ac.kr

The Method of Gene Selection for Machine Learning Classifiers In Cancer Classification

Hyung-Keun Park[O] Soo-Jung Lee Yillbyung Lee
Division of Computer and Information Engineering, Yonsei University

요 약

유전자 발현 분석 시스템에 있어서 microarray 기술의 발전은 유전 질환 진단의 정확성과 신뢰도를 향상시키는 데에 큰 기여를 하였다. 다양한 microarray 기술을 통해 얻은 대량의 유전자 발현 정보는 기계 학습 분류기를 이용한 암의 분류와 진단, 예측 분야에도 효과적으로 이용될 수 있다. 이 과정에서 종류에 따른 암의 정확한 분류를 위해서는 되도록 해당 암 클래스와의 직접적인 연관이 있는 유전자만을 선택하여 활용하는 것이 효과적이다. 본 논문에서는 이러한 정보력 있는 유전자(informative gene)를 효과적으로 선택할 수 있는 유전자 선택 방법을 제시하고, 이를 이용하여 세 가지 벤치마크 암 데이터에 대하여 체계적인 실험을 하였다. 그 결과 향상된 분류 성능을 확인할 수 있었다.

1. 서론

유전자 발현 분석 시스템은 서열 분석 시스템과 더불어 Human Genome Project의 궁극적인 목적을 달성하기 위한 실질적 세부 응용 분야이다. 유사성 비교 등으로 유전자의 기능을 추측하고, 단백질의 분자 구조 및 기능을 추측하는 것은 유전자 지도에 새로운 현실적 의미를 부여할 수 있다는 점에서 매우 중요한 의의를 갖는다. 이러한 분야의 발전을 가속화시킨 것은 microarray 기술의 발전이다. Microarray는 슬라이드에 부착하는 검출 DNA(probe DNA)가 어떻게 만들어지는가에 따라 크게 cDNA microarray와 oligonucleotide microarray로 구분된다. 이러한 DNA chip 기술의 발전으로 특정 환경과 조건에 따른 유전자 발현 정보를 동시에 대량으로 획득하고 분석, 처리할 수 있게 되었다[1].

Microarray 기술은 기계 학습 분류기를 이용한 암의 분류와 진단, 예측 분야에도 효과적으로 적용될 수 있다[2][4]. 암의 종류에 따른 정확한 예측과 분류는 그 진단과 치료에 있어서 매우 중요한 부분을 차지하기 때문에, microarray를 통해 얻은 대량의 유전자 발현 정보를 이용하여 암을 종류별로 분류해 내는 문제에 대한 많은 연구가 진행되어 왔다[3]. 정확한 암의 분류를 위해 분류 성능을 향상시킬 수 있는 방법은 크게 두 가지가 있다. 그 하나는 분류기 자체의 성능을 향상 시키거나, 여러가지 분류기를 사용하여 각 분류기의 결과를 결합함으로써 분류 성능을 향상시키는 방법이다. 다른 하나는 기계 학습 기반 분류기를 사용할 경우, 암 분류에 직접적인 연관이 있는 정보력 있는 유전자 데이터를 분류기의 학습을 위한 데이터로 사용할 수 있도록 하여, 분류 성능을 향상시키는 방법이다. 전자의 경우, 결정적으로 큰 비용을 요구한다. 그러나 후자의 경우, 구체적인 연구가 미흡하고 요구 비용이 적으며 부가적인 장점도 내포하고 있다. 따라서 효과적으로 정보력 있는 유전자 데이터를 선택할 수 있는 방법을 모색해 볼 필요가 있다.

본 논문에서는 세 가지 종류의 벤치마크 데이터 집합에 대해 다양한 유전자 선택 방법을 적용하여 그에 따른 분류 성능을 비교 평가하고, 기존의 선택 방법을 조합(combination)함으로써 분류 성능을 향상시킬 수 있음을 보이고자 한다.

본 연구는 과기부 뇌신경정보학사업으로부터 부분적인 지원을 받아 수행되었음

2. 기계 학습 기반 암 분류 시스템

Microarray를 통해 얻어진 유전자 발현 정보 데이터들로부터 암 분류에 상대적으로 많은 연관성을 갖고 있는 유전자 정보만을 선택하여 분류기를 학습시키고, 학습된 분류기를 이용하여 새로운 암 데이터에 대해 그 종류를 예측하여 분류하는 체계를 기계 학습 기반 암 분류 시스템이라 한다. 기계 학습을 위한 양질의 유전자 데이터를 선택하는 과정까지를 시스템의 전단부(front-end), 그 이후 기계학습을 거쳐 예측 분류를 수행하는 과정까지를 시스템의 후단부(back-end)라 하겠다.

2.1 Gene Selection

암 분류를 위해 microarray를 통하여 얻게 되는 유전자 발현 정보 데이터는 기계 학습 기반 분류기를 사용하여 분류를 시도하던 기존 전통적 분야의 데이터 집합과는 매우 다른 형태를 지닌다. 직접 tumor 샘플에서부터 microarray 기술에 의해 데이터가 생성되기 때문에 샘플의 수는 적고 각 샘플 당 속성(attribute) 즉, 유전자의 수는 수천 개에서 수만 개 내외가 되는 것이 일반적이다. 이러한 특징은 기계 학습 기반 분류기를 이용하여 학습을 하기에 적절하지 않다. Microarray로부터 얻어지는 유전자 데이터 중에서 실제로 각 샘플의 특정 클래스, 즉 암의 특정 클래스에 연관이 큰 유전자는 그 수가 매우 적으며, 실험에 의한 잡음을 포함하고 있는 유전자 발현 데이터 또한 많다. 따라서 기계 학습 기반 분류기를 이용하여 현실적으로 효과적인 학습을 하기 위해서는 해당 클래스와의 연관성이 높은 유전자들을 시스템의 전단부인 전처리 과정에서 선택해야만 한다. 이것을 유전자 선택과정이라고 한다[2]. 주로 통계적 상관관계 척도나 클러스터링 기법 등을 이용하여 해당 클래스와 상대적으로 상관관계가 높은 유전자들을 선택한다. Microarray로부터 얻는 데이터 집합이 샘플 M개와 유전자 N개로 이루어져 있다고 하고, M개의 샘플은 클래스 A와 클래스 B 두 가지 종류로 구분된다고 가정한다. 처음의 K($0 \le K \le M$)개의 샘플을 암 세포의 샘플이라고 가정하고, 나머지 M−K개의 샘플을 정상 세포 혹은 처음의 K개의 샘플과는 다른 종류의 암 세포의 샘플이라고 가정하면, 각 유전자 데이터는 수식 (1)과 같은 형태의 벡터로 표현할 수 있다.

$$G_i = (e_1, e_2, e_3, \cdots, e_K, e_{K+1}, \cdots, e_M) \quad (1)$$

각 클래스에 대한 특징을 극단적으로 뚜렷하게 나타내는 이상적으로 발현하는 유전자를 G_{ideal}이라고 하면, 처음의 K개의 암 세포의 특징을 1로 정의하고 나머지 M−K개의 정상세포 혹은 처음의 K개와는 다른 암세포의 특징을 0로 정의하여 식 (2)와 같은 벡터로 표현할 수 있다[4].

$$G_{ideal} = (1, 1, 1, \cdots, 1, 0, 0, 0, \cdots, 0) \quad (2)$$

표 1. 유전자 선택을 위한 기존 상관관계 척도

- Pearson correlation coefficient (PC)

$$PC(G_i, G_{ideal}) = \frac{\sum G_i G_{ideal} - \frac{\sum G_i \sum G_{ideal}}{N}}{\sqrt{\left(\sum G_i^2 - \frac{(\sum G_i)^2}{N}\right)\left(\sum G_{ideal}^2 - \frac{(\sum G_{ideal})^2}{N}\right)}}$$

- Spearman correlation coefficient (SC)

$$SC(G_i, G_{ideal}) = 1 - \frac{6\sum (D_G - D_{ideal})^2}{N(N^2-1)}$$

- Euclidean distance (ED)

$$ED(G_i, G_{ideal}) = \sqrt{\sum (G_i - G_{ideal})^2}$$

- Cosine coefficient (CC)

$$CC(G_i, G_{ideal}) = \frac{\sum G_i G_{ideal}}{\sqrt{\sum G_i^2 \sum G_{ideal}^2}}$$

(D_G and D_{ideal} are the rank matrices of G_i and G_{ideal})

여러 개의 상관관계 척도를 사용하여 식 (2)와 각 유전자 사이의 상관관계를 측정한다. 각 상관관계 척도별로 상관관계가 높은 유전자들을 순차 정렬하고 상위의 유전자를 선택하여 기계학습기반 분류기의 학습 데이터로 사용한다. 기존 상관관계 척도는 표 1에 정리하였다.

그러나 정확한 암 분류에 있어서 중요한 정보를 줄 수 있다고 판단된 유전자가 각 상관관계 척도마다 상이하게 나타난다. 이것은 각 상관관계 척도마다 정보력 있는 유전자 데이터를 판단하는 기준이 다르기 때문이다. 그러므로 다수의 상관관계 척도로부터 informative하다고 평가된 유전자 데이터를 선택하기 위해, 각 상관관계 척도에 따라 계산된 값으로 이루어진 배열을 서로 적절하게 조합하도록 한다. 조합 방법을 모색하기 위해 사용할 수 있는 정보는 각 척도에 의해 계산된 유전자 각각의 G_{ideal}에 대한 상관관계 수치와 이를 기반으로 하는 상관관계 서열 번호(ranking)이다. 그러나 각각의 상관관계 척도는 그 판단 기준이 다르기 때문에 각각의 척도에 의해 계산된 상관관계 수치가 서로 의미하는 바가 다르고, 따라서 다른 상관관계 척도에 의한 상관관계 수치를 서로 연산하여 조합하는 것은 기대 이하의 결과를 초래할 수 있다. 따라서 상관관계 서열 번호를 이용하여 다수의 척도에서 정보력 있는 유전자 데이터로 인정받은 유전자들을 선택한다. 이를 위한 알고리즘은 표 2에 정리하였다.

표 2. 유전자 선택 방법 알고리즘

1) 조합하고자 하는 상관관계 척도를 k개 선택한다.
2) 각 척도에서 i번째 유전자 G_i가 평가받은 상관관계 서열 번호(ranking)를 모두 합산한다.
3) 2)에서 합산한 결과 값을 1)의 k로 나눈다.
4) 3)의 결과를 G_i의 상관관계 척도 k개에 대한 조합 값으로 정의한다.
5) G_i의 조합 값이 상대적으로 작은 유전자들을 선택하고 분류 성능 향상을 위해 이를 활용한다.

기계 학습 기반 암 분류 시스템 전단부의 마지막인 5)의 과정에서 만들어지는 정보력 있는 유전자 데이터 목록은 신약 제조와 같은 관련 분야에 직접적인 연구 동기를 부여할 수 있다.

2.2 Classification과 Prediction

Microarray를 통해 얻어진 유전자 발현 정보 데이터들로부터 암 분류에 상대적으로 많은 연관성을 갖고 있는 유전자 정보만을 선택하여 분류기를 학습시키고, 학습된 분류기를 이용하여 새로운 암 데이터에 대해 그 종류를 예측하여 분류하는 것이 일반적인 기계 학습 기반 암 분류 과정이다. 분류기의 구현을 위해 사용되는 대표적인 알고리즘은 다층 신경망(multi layer perceptron, MLP)이다.

인공 신경망의 대표적인 기계 학습 알고리즘인 MLP는 대부분의 패턴 인식 문제에 대해 안정적인 성능을 보이며, 일단 학습이 끝나면 응용 단계에서는 매우 빠르게 결과를 출력한다. MLP는 back propagation알고리즘을 사용하는데 이것은 출력층 오차 신호를 이용하여 은닉층과 출력층 간의 연결 강도를 변경하고 출력층 오차 신호를 은닉층에 역전파하여 입력층과 은닉층 간의 연결 강도를 변경하는 학습 방법이다.

3. 실험 및 결과

3.1 실험 데이터

3.1.1 Leukemia dataset

급성 골수성 백혈병(acute myeloid leukemia, AML) 환자 25명과 급성 림프성 백혈병(acute lymphoblastic leukemia, ALL) 환자 47명으로부터 얻은 데이터이다. Leukemia dataset은 72개의 샘플 데이터로 구성되어 있으며, 38개를 학습 데이터로 사용하였고, 34개를 테스트 데이터로 사용하였다. 각 샘플은 7129개의 유전자 발현 정보를 갖고 있다.

3.1.2 Colon cancer dataset

Colon dataset은 결장암 환자의 결장 상피 세포로부터 추출한 62개의 샘플 데이터이며, 40개는 암 세포의 샘플이고, 나머지 22개는 정상 세포의 샘플이다. 62개의 샘플 중에서 31개를 학습 데이터로 사용하였고, 나머지 31개를 테스트 데이터로 사용하였다. 각 샘플은 2000개의 유전자 발현 정보를 갖고 있다.

3.1.3 Lymphoma dataset

Lymphoma dataset은 GC B-like 샘플 24개와 activated B-like 샘플 23개로 구성되어 있다. 47개의 샘플 중에서 22개를 학습 데이터로 사용하였고, 나머지 25개를 테스트 데이터로 사용하였다. 각 샘플은 4026개의 유전자 발현 정보를 갖고 있다.

3.2 실험 환경

MLP를 이용하여 모멘텀은 0.9로, 총 레이어의 수는 3으로 고정한 후, 학습률을 0.01에서 0.50으로 변화시켜가며 실험 하였다. 학습과정에서의 최대 반복 수는 300번으로 고정하였다.

3.3 실험 결과

Leukemia, colon, lymphoma dataset에 대하여 기존의 상관관계 척도 4가지와 이들을 조합한 유전자 선택 방법 11가지를 MLP를 이용한 분류기와 함께 사용했을 때의 결과를 표 3, 4, 5에 각각 정리하였다. 표에 나타나 있는 수치는 해당 조건에서의 분류기의 인식률을 의미하며 단위는 퍼센트(%)이다.

Leukemia dataset의 경우 기존 상관관계 척도인 Pearson correlation coefficient(PC)만을 단일하게 사용하여 유전자 선택을 한 경우 분류기에서 88.24%의 인식률을 나타내었고, Euclidean distance(ED)만을 단일하게 사용하여 유전자 선택을 한 경우 분류기에서는 82.35%의 인식률을 나타내었다. 반면 이 두가지의 상관관계 척도를 조합(PC-ED)하여 두 척도 모두에서 상관관계가 높게 나타난 정보력이 있는 유전자 데이터들을 선택하여 사용한 결과 분류기에서 94.12%의 인식률을 나타내어 분류 성능이 향상되었음을 확인할 수 있다. 그러나 이러한 향상은 두 가지 척도를 조합하는 경우에서 가장 두드러지게 나타나고 세 가지 이상의 척도를 조합하는 경우에서는 크게 만족할만한 향상을 보이지는 않고 있다. 유전자 선택 방법 하나만으로는 분류하고자 하는 해 공간을 전부 포함하지 못했던 부분을 다양한 유전자 선택 방법의 조합 과정에서 포함해 주기 때문에 해 공간의 탐색 범위가 넓어져 분류기의 인식률이 향상될 수 있다. 그러나 많은 유전자 선택 방법의 조합은 오히려 포함하지 않아야 될 해 공간까지 포함하는 경우 분류기의 인식률을 저하시킬 수도 있으므로 조합 시 신중한 선택을 해야 함을 알 수 있다.

표 3. 조합을 통한 유전자 선택 방법과 분류기에 따른 인식률 (Leukemia)

PC		ED		SC		CC	
88.24		82.35		79.41		85.29	
PC-ED	PC-SC	PC-CC	ED-SC	ED-CC	SC-CC		
94.12	85.29	91.18	82.35	88.24	85.29		
PC-ED-SC	PC-ED-CC	PC-SC-CC	ED-SC-CC				
91.18	94.12	85.29	82.35				
PC-ED-SC-CC							
88.24							

표 4. 조합을 통한 유전자 선택 방법과 분류기에 따른 인식률 (Colon)

PC		ED		SC		CC	
74.19		67.74		58.06		80.65	
PC-ED	PC-SC	PC-CC	ED-SC	ED-CC	SC-CC		
80.65	67.74	87.10	67.74	83.87	80.65		
PC-ED-SC	PC-ED-CC	PC-SC-CC	ED-SC-CC				
80.65	87.10	80.65	80.65				
PC-ED-SC-CC							
74.19							

표 5. 조합을 통한 유전자 선택 방법과 분류기에 따른 인식률 (Lymphoma)

PC		ED		SC		CC	
68.00		56.00		60.00		68.00	
PC-ED	PC-SC	PC-CC	ED-SC	ED-CC	SC-CC		
76.00	68.00	84.00	64.00	72.00	72.00		
PC-ED-SC	PC-ED-CC	PC-SC-CC	ED-SC-CC				
72.00	80.00	76.00	68.00				
PC-ED-SC-CC							
64.00							

참 고 문 헌

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

[2] L. Li, 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, no. 12, pp. 1131-1142, 2001.

[3] S. Dudoit, J. Fridlyand, and T. P. Speed, " Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, pp. 77-87, 2002.

[4] T. R. Golub, D. K. Slonim, and P. Tamayo, " Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp.531-537, 1999.