종양 분류를 위한 특징 추출 및 분류 기법

박윤정[0], 이민수, 박승수
이화여자대학교 컴퓨터학과
cssstar@ewhain.net, ssue@ewhain.net, sspark@ewha.ac.kr

Feature Selection and Classification Methods for Tumor Classification

Yun Jung Park[0], Min Su Lee, Seung Soo Park
Department of Computer Science and Engineering, Ewha Womans University

요 약

현재 마이크로어레이 기술은 대량의 유전자 발현 데이터 특히 종양과 관련한 데이터들을 쏟아내고 있다. 이 데이터를 기반으로 종양의 종류에 따른 유전자들의 차별적 발현 양상을 분석하고 발현량의 변화가 두드러지는 유전자들에 기반하여 종양을 분별할 수 있는 분류 모델을 구축한 후, 이것을 종양을 진단하거나 예측하는데 이용할 수 있다. 대부분의 종양은 생성 매커니즘에 따라 세부 부류로 나눌 수 있고 세부 부류에 따라 치료 방법이나 예후가 달라지므로, 정확하게 종양의 세부 부류를 진단하는 것이 매우 중요하다.

본 논문에서는 종양의 종류에 따라 발현량이 민감하게 변화하는 유전자들을 뽑아내기 위한 특징 추출 방법들과 추출된 특징들에 기반해서 종양의 종류를 분별할 수 있는 기계학습 알고리즘들의 조합들의 성능을 비교분석 하였다.

1. 서 론

생명공학의 발전은 생명체 정보들을 대량으로 얻어내는데 큰 역할을 하고 있다. 마이크로어레이 (microarray) 기술은 처리 조건이나 환경에 따른 대량의 유전자 발현 정보를 정량적인 수치로 제공해 준다. 2000년대 초반부터 마이크로어레이를 이용하여 질병을 분류하고 진단하는데 이용하기 위한 연구가 활발히 진행되고 있다. 특히 종양 조직에 대한 마이크로어레이 데이터를 사용하여 종양 종류에 따라 유전자가 차별적으로 발현되는 양상을 분석함으로써, 종양의 분류에 유용한 유전자를 식별하고 정확한 분류 도구를 구축하고자 하는 연구도 이루어지고 있다. 이러한 분류 방법은 불확실성을 내포하고 있는 기존의 형태학적, 임상적 기반의 종양 분류 방법들의 대신할 수 있으면서도 육안으로는 구분하기 어려운 종양의 세부 분류들까지도 구분할 수 있을 것으로 기대되고 있다.

그런데, 수천 개에서 수만여 개의 유전자들이 박혀있는 마이크로어레이 데이터는 종양 샘플을 구하기가 쉽지 않을 뿐만 아니라 실험 비용도 매우 비싸 실제 표본의 개수에 비해 유전자의 개수가 훨씬 많다는 특성을 가지고 있다. 따라서 수많은 유전자들로부터 실제 종양들의 세부 부류에 따라 확연하게 발현량이 변하는 표본 분류에 유용한 유전자들을 추출하기 위한 특징 추출 (feature selection) 방법과 이 유전자들을 이용하여 보다 정확한 종양 분류 모델 (tumor classification model)을 구축하는 것이 매우 중요하다.

본 논문에서는 백혈병에 대한 마이크로어레이 데이터에 information gain, gini index 방법을 이용하여 질병의 클래스를 구분하는데 있어 분별력 있는 유전자 리스트를 선별한 후, 그 유전자들의 발현 데이터에 Naïve Bayes, KNN, Decision Tree, SVM 알고리즘을 적용하여 종양 분류 모델을 구축하고 각각의 실험 결과들을 비교 분석함으로써 성능평가를 하였다.

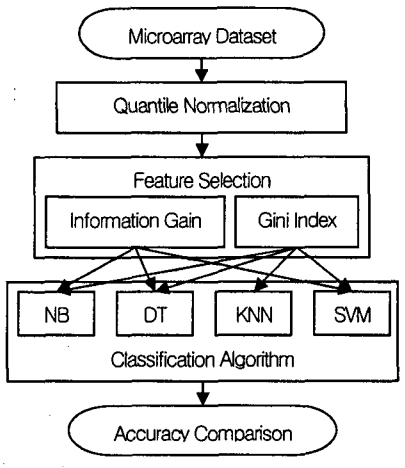

그림1. 시스템 구성도

표1. 데이터셋의 특징

데이터셋	표본의 수	유전자의 수	클래스 개수
ALLAML	72	7129	2
MLL	72	12582	3
SALL	327	12558	7

2. 데이터셋과 정규화

본 논문에서는 백혈병과 관련된 세 가지 마이크로어레이 데이터셋을 사용하였다 (표1). 단순한 이진 클래스 데이터셋 뿐만 아니라 클래스 개수가 3개, 7개인 데이터셋을 사용함으로써, 분별하기 힘든 종양의 세부 부류를 정확하게 분류하는 알고리즘을 찾고자 하였다. 사용한 첫 번째 데이터셋은 Golub의 실험에서 사용된 급성 림프구성 백혈병(ALL) 또는 급성 골수성 백혈병(AML)을 앓고 있는 환자에게서 얻은 것으로 47명의 ALL환자와 25명의 AML환자 데이터로 구성되어 있다[1]. 두 번째 복합형 백혈병 (mixed linage leukemia: MLL) 데이터셋은 24명의 ALL환자와 20명의 MLL환자와 28명의 AML환자 샘플로 구성되어 있다 [2]. 세 번째 데이터셋은 Subtypes of Acute Lymphoblastic Leukemia (SALL)으로 15명의 BCR-ABL환자, 27명의 E2A-PBX 환자, 64명의 Hyperdip50환자, 20명의 MLL환자, 43명의 T-ALL 환자, 79명의 TEL-AML1환자, 속하지 않는 79명의 OTHERS환자들의 데이터셋이다[3].

각 데이터셋 안의 여러 샘플들을 함께 분석하기 위해 Strand Genomics사의 Avadis를 이용하여 모든 슬라이드의 사분위 수를 똑같게 맞추는 quantile-normalization을 적용하였다 [4]. 정규화된 데이터셋은 따로 추상화 과정을 거치지 않고 실수 값 그대로 입력 데이터로 사용하였다.

3. 특징 추출

Microarray로부터 얻어지는 유전자의 수는 대략 수천 개에서 수만 개이다. 하지만 얻어진 데이터에서 각 종양 샘플의 특정 클래스와 연관이 있는 유전자의 수는 그 보다 훨씬 작다. 따라서 유전자 데이터를 이용하여 클래스를 분류하기 위해서는 클래스와의 연관성이 높은 유전자를 추출하는 과정이 필요하다. 이러한 과정을 일반적으로 특징 추출 과정이라고 한다. 이는 패턴분류 시 분류기(classifier)의 성능을 향상시키고 복잡도(complexity)를 감소시키는 등 효율적인 패턴분류를 가능하게 한다.

본 논문에서는 다음과 같은 정보공학적 방법으로 종양 분류에 유용한 유전자를 rank gene을 이용하여 선택하였다 [5].

(i) Information gain: Information gain 은 각 유전자에 대해서 특정 값을 기준으로 샘플들을 나눌 때 나눠진 그룹 내부의 entropy가 얼마나 낮아지는지를 측정하고, 가장 높은 information gain 값을 가지는 유전자들을 뽑는 방법이다.

$$\text{information gain} = \sum_{i=1}^{k}\left(\frac{l_i}{n}\log\frac{l_i}{n_l}+\frac{r_i}{n}\log\frac{r_i}{n_r}\right)-\sum_{i=1}^{k}\left(\frac{l_i+r_i}{n}\right)\log\left(\frac{l_i+r_i}{n}\right)$$

(ii) Gini Index: Gini index 는 동질성(다양성)의 높고 낮음의 척도에 의해서 분류하는 방법으로 0~1 의 값을 갖고, 0 으로 갈수록 균등의 의미를 가지며 1 로 갈수록 불균등의 의미를 갖는다

$$\text{gini index} = \frac{n_l}{n}\left(1-\sum_{i=1}^{k}\left(\frac{l_i}{n_l}\right)^2\right)+\frac{n_r}{n}\left(1-\sum_{i=1}^{k}\left(\frac{r_i}{n_l}\right)^2\right)$$

4. 분류 기법과 분석 결과

많은 기계학습 알고리즘은 최근 유전자 정보를 이용하여 종양을 예측하고, 분류하는 연구에 적용되어왔다. 본 논문에서는 분별력 있는 유전자에 기반하여 종양을 분류하는 모델을 구축하는데 있어 다음과 같은 4가지 기계학습 알고리즘을 사용하였다.

4.1. 분류기법

(i) Naïve Bayes (NB): Naïve Bayes는 베이지안 확률 모형에 기초한다. 사건 E 와 C_i가 있을 때, E 에 대해 C_i가 발생할 확률은 $P(C_i \mid E) = \frac{P(E \mid C_i) \times P(C_i)}{P(E)}$ 이다.

학습 데이터에 나타난 단어들이 특정 범주의 data에 나타날 확률을 계산하여 새로운 데이터의 범주를 예측하는 방법이며 자질들 사이의 독립성을 가정하여 입력 데이터에 대한 범주의 확률을 계산한다.

(ii) K-Nearest Neighborhood (KNN): KNN 알고리즘은 분류하고자 하는 샘플을 입력 받은 후에 Pearson correlation coefficient와 같은 상관관계 척도 혹은 Euclidean distance와 같은 유사도 척도를 이용하여 입력 샘플과 가장 유사한 k개의 샘플을 찾는다. 선택된 k개 샘플들의 분류 결과에 분류하고자 하는 샘플과의 유사도를 가중치로 곱하여 분류하고자 하는 샘플의 분류 결과를 결정한다.

(iii) Decision Tree (DT): Decision Tree는 순서도 같은 트리구조이다. 안쪽 노드는 속성에 대한 검사표시이고 가지는 검사의 결과를 나타내며 리프 노드는 클래스 레이블이나 클래스 분포를 나타낸다. Decision Tree는 두 가지 과정을 통해 생성 된다. 하나는 트리 생성으로 처음에 학습 데이터는 루트에 있고 선택된 속성을 기준으로 재귀적으로 분할한다. 다른 하나는 가지치기 과정으로 반사잡음이나 이상치의 가지를 식별하고 잘라낸다. 구축된 Decision Tree를 사용해서 트리에 대응하는 표본의 속성 값을 테스트함으로써 알려지지 않은 표본을 분류한다.

(iv) Support Vector Machine (SVM): Support Vector Machine은 이차원 데이터 분류문제에서 가장 최적의 초평면(Hyperplane)을 구하여 이를 경계 결정면으로 선택한다. 최적의 초평면은 선형 분리가 가능한 두 집단에 대해 마진을 최대로 하여 집단을 구분 짓는다. 하지만 실제 문제의 경우 선형적으로 구성되지 않기 때문에 커널 함수를 이용하여 비선형적 특징공간을 선형적 특징공간으로 매핑한 후에 선형 SVM으로 분류하게 된다.

4.2 분석결과

각 데이터셋에 세부 클래스와의 연관성이 높고 분별력 있는 유전자들을 information gain과 gini index에 기반해서 순위대로 나열한 후에, 유전자를 1위부터 50위, 100위, 150위, 200위까지 샘플링한 4개의 서브 데이터셋을 만들었다. 이렇게 선택된 각 특징 추출 방법 별 분별력 있는 유전자 개수에 따라 앞서 설명한 4가지의 기계학습 알고리즘으로 종양 클래스 분류 모델을 만들고 10 fold cross-validation을 사용하여 정확도를 측정하고 서로 비교하였다.

지면 관계상 각 데이터셋에 대해서 250개, 50개의 유전자로 만든 분류 모델의 정확도만 표로 정리하였다 (표2~4). ALL-AML과 MLL 데이터 셋의 경우에는 유전자를 250개 사용할 때보다 분별력 있는 유전자 50개만 사용했을 때, 대부분의 모델에서 더 정확도가 높아지는 경향을 나타내었다. 특히 MLL 데이터에서 information gain 방법을 사용하여 선택한 50개의 유전자로 SVM알고리즘을 사용해서 모델을 구축했을 때, 정확도가 100%를 나타내었다. 그러나 클래스 개수가 7개인 SALL 데이터의 경우에는 KNN과 SVM을 사용할 경우, 유전자 개수를 줄일 경우 오히려 정확도가 낮아지는 경향을 보였다. 표2~4에서 볼 수 있듯이 Decision Tree algorithm을 제외하면 ALL-AML과 MLL은 95~100%의 정확도를 보였고, SALL은 80~93%의 정확도를 보였다. SALL은 표본의 수가 327개나 됨에도 클래스의 개수도 7개나 된다. 이것은 클래스의 개수가 많아질수록 분별 정확도가 낮아지므로 더 많은 표본이 요구된다는 것을 보여준다. 클래스의 개수가 적을 경우에는 유전자 50개만 가지고도 높은 정확도로 분류할 수 있으나, 클래스의 개수가 많아지면 알고리즘에 따라 좀더 많은 유전자가 요구될 수 있다. 모든 경우에 있어서 성능이 높은 알고리즘은 없었으며, 대부분의 경우에 Naïve Bayes와 KNN은 비슷한 성능을 보였다. Decision Tree는 모든 경우에서 제일 낮은 성능을 나타냈다. Decision Tree는 분별 트리를 만드는 과정에서 각 단계마다 동적으로 해당 노드 안의 샘플들을 분류하는데 있어 가장 분별력 있는 유전자를 선택하므로, 초기에 미리 유전자 리스트를 50개 또는 250개로 제한하는 것이 오히려 모델의 성능을 저하시키는 요인이 되었던 것 같다.

표2. ALL-AML Leukemia 분석결과

Gene의 개수=250				
	NB	KNN	DT	SVM
Inforgation gain	97.2 %	97.2 %	90.3 %	98.6 %
Gini index	97.2 %	98.6 %	90.3 %	98.6 %
Gene의 개수=50				
	NB	KNN	DT	SVM
Inforgation gain	98.6 %	98.6 %	90.3 %	95.8 %
Gini index	98.6 %	98.6 %	93.1 %	97.2 %

표3. MLL Leukemia 분석결과

Gene의 개수=250				
	NB	KNN	DT	SVM
Inforgation gain	97.2 %	98.6 %	90.3 %	97.2%
Gini index	95.8%	97.2 %	87.5 %	97.2%
Gene의 개수=50				
	NB	KNN	DT	SVM
Inforgation gain	97.2 %	98.6 %	93.1 %	100 %
Gini index	97.2%	97.2%	83.3 %	95.8 %

표4. SALL분석결과

Gene의 개수=250				
	NB	KNN	DT	SVM
Inforgation gain	87.5 %	90.8 %	81.1 %	92.7 %
Gini index	86.2%	89.3 %	79.8 %	91.1 %
Gene의 개수=50				
	NB	KNN	DT	SVM
Inforgation gain	88.1 %	89.3%	81.1 %	85.3 %
Gini index	80.4 %	85.0 %	75.8 %	84.1%

5. 결론 및 향후 연구과제

본 논문에서는 백혈병에 대한 마이크로어레이 데이터를 사용하여 정보공학적 방법으로 분별력 있는 유전자들을 추출한 후, Naïve Bayes, KNN, Decision Tree, SVM 알고리즘을 이용하여 클래스 분류 모델을 구축하고, 성능을 비교분석 하였다. 실험결과들을 비교 분석한 결과 특징 추출 및 종양 부류 예측 모델의 최적의 조합 찾을 수는 없었지만, 클래스와 샘플 개수에 따른 대략적인 성능의 패턴은 추정할 수 있었다. 모든 경우에 있어서 최고의 정확도를 제공하기 위해서는 여러 알고리즘을 앙상블로 묶어서 구현해야 할 것이다.

참고 문헌

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

[2] Armstrong, S. A., *et al.* " "MLL Translocations Specify A Distinct Gene Expression Profile that Distinguishes A Unique Leukemia", *Nature Genetics*, 30:41-47, 2002

[3] Yeoh, E. J. *et al.* "Classification, Subtype Discovery,and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling". *Cancer Cell*, 1:133-143, 2002

[4] http://avadis.strandgenomics.com/

[5] Su, Y., *et al.* " RankGene: identification of diagnostic genes based on expression data' , *Bioinformatics*, 19(12):1578-1579, 2003