암 분류를 위한 분류기법의 성능비교

박윤정°, 박승수
이화여자대학교 컴퓨터학과
allypark@ewhain.net, sspark@ewha.ac.kr

Performance Comparison of Multiclass Classification Methods for cancer Classification

Yun Jung Park°, Seung Soo Park
Department of Computer Science and Engineering, Ewha Womans University

요 약

현재 마이크로어레이 기술은 대량의 유전자 발현 데이터 특히 암과 관련한 데이터들을 쏟아내고 있다. 이 데이터를 기반으로 암의 종류에 따른 유전자들의 차별적 발현 양상을 분석하고 발현량의 변화가 두드러지는 유전자들에 기반하여 암을 분별할 수 있는 분류 모델을 구축한 후, 이것을 암을 진단하거나 예측하는데 이용할 수 있다.

본 논문에서는 마이크로어레이 데이터를 사용해 특징추출방법과 분류를 위한 Naive Bayes, K-Nearest Neighborhood, Decision Tree, Support Vector Machine, Neural Network 알고리즘을 이용하여 최적의 조합을 찾고 어떤 알고리즘이 가장 효과적인지 실험을 통해 분석해보고 성능평가 하는 것을 목표로한다.

1. 서 론

마이크로어레이 (microarray) 기술은 처리 조건이나 환경에 따른 대량의 유전자 발현 정보를 정량적인 수치로 제공해 준다. 종양 조직에 대한 마이크로어레이 데이터를 사용하여 암 종류에 따라 유전자가 차별적으로 발현되는 양상을 분석함으로써, 암의 분류에 유용한 유전자를 식별하고 정확한 분류 도구를 구축하고자 하는 연구도 이루어지고 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 암들의 세부 부류에 따라 확연하게 발현량이 변하는 표본 분류에 유용한 유전자들을 추출하기 위한 특징 추출 (feature selection) 방법과 이 유전자들을 이용하여 보다 정확한 암 분류 모델 (cancer classification model)을 구축하는 것이 매우 중요하다.

본 논문에서는 클래스가 2개, 3개, 7개로 구성된 백혈병에 대한 마이크로어레이 데이터를 이용해 데이터의 정규화를 거쳐 특징추출방법인 Information Gain, Gini Index, One-dimensional Support Vector Machine, T-statistic 방법을 이용하여 분별력 있는 유전자 리스트를 선별하였다. 발현된 유전자 데이터에 Naive Bayes, K-Nearest Neighborhood, Decision Tree, Support Vector Machine, Neural Network 알고리즘을 적용하여 암 분류 모델을 구축하고 각각의 실험 결과들을 비교 분석함으로써 성능평가를 하였다.

2. 분류 기법

많은 기계학습 알고리즘은 최근 유전자 정보를 이용하여 암을 예측하고, 분류하는 연구에 적용되어왔다. 본 논문에서는 분별력 있는 유전자에 기반하여 암을 분류하는 모델을 구축하는데 있어 다음과 같은 5가지 기계학습 알고리즘을 사용하였다.

(i) Naive Bayes (NB): 베이지안 확률 모형에 기초한다. 사건 E 와 C_i가 있을 때, E 에 대해 C_i가 발생할 확률은

$$P(C_i \mid E) = \frac{P(E \mid C_i) \times P(C_i)}{P(E)} \text{ 이다.}$$

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): 분류하고자 하는 샘플을 입력 받은 후에 상관관계 척도 혹은 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): 순서도 같은 트리 구조이다. 안쪽 노드는 속성에 대한 검사표시이고 가지는 검사의 결과를 나타내며 리프 노드는 클래스 레이블이나 클래스 분포를 나타낸다.

(iv) Support Vector Machine (SVM): 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.

(v) Neural Network(신경망): 인간 두뇌의 신경세포를 모델링 하여 지능을 구현하고자 하는 기법이다. 신경망은 병행적으로 상호 작용하는 여러 개의 계산요소들로 이루어져 있으며, 각 계산요소는 가중치 합(weighted sum)과 같은 단순한 계산만을 수행한다.

3. 데이터셋과 정규화

본 논문에서 분석을 위해 사용한 데이터는 백혈병에 관한 세가지 마이크로어레이 데이터 셋 이다(표1).

표1. 데이터셋의 특징

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

단순한 이진 클래스 데이터 셋뿐만 아니라 클래스 개수가 3개, 7개인 데이터 셋을 사용함으로써, 분별하기 힘든 암의 세부 부류를 정확하게 분류하는 알고리즘을 찾고자 하였다. 사용한 첫 번째 데이터셋은 Golub의 실험에서 사용된 급성 림프구성 백혈병 또는 급성 골수성 백혈병을 앓고 있는 환자에게서 얻은 것으로 47명의 ALL환자와 25명의 AML 환자 데이터로 구성되어 있다[1]. 두번째 복합형 백혈병 데이터셋은 24명의 ALL환자와 20명의 MLL환자와 28명의 AML환자 샘플로 구성되어 있다[2]. 세 번째 데이터셋은 15명의 BCR-ABL환자, 27명의 E2A-PBX 환자, 64명의 Hyperdip50환자, 20명의 MLL환자, 43명의 T-ALL환자, 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.

4. 특징 추출

유전자 데이터를 이용하여 클래스를 분류하기 위해서는 클래스와 연관성이 높은 유전자를 추출하는 과정이 필요하다.

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

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

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

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

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

(iii) One-dimensional SVM: SVM은 기본적으로 두 범주를 갖는 객체들을 분류하는 방법이다. 두 범주를 구분하는 하이퍼플레인은 무수히 많을 수 있는데 어떤 것이 가장 적절한 것인지를 결정하고자 한다.

(iv) T-statistic : 각각의 유전자에 대해 t-통계량의 절대값을 감소하는 순으로 정렬해서 순위를 매긴다. 여기에서 t-통계량은 독립적인 두 집단의 평균차이를 나타내므로 class가 2개인 경우에만 해당한다.

5. 성능지표 와 분석 결과

5.1 성능지표

분류 알고리즘의 성능 지표를 살펴보기 위해서는 알고리즘의 정확성, 데이터를 구성하는 레코드와 속성의 가지 수가 충분한가에 대한 양적 척도와 속성이 취한 값들이 정확한가 또는 빠진 속성 값은 어느 정도인가에 대한 질적 척도에 관한 데이터의 충실도, 실제 측정된 값에 대한 신뢰성을 측정해야 한다. 성능 평가 기준(performance measures)은 얼마나 정확한 예측을 했는가를 평가하는 것이다. 즉, 예측된 모형에 대한 적합성에 대해서 평가한다. 각각의 분류 알고리즘에 대한 성능 지표를 측정하는 방법은 다음 표2와 같다.

표2. 평가기준

Accuracy =	(TN+TP)/(TP+FP+FN+TN)
TP-rate =	(TP)/(TP+FN)
FP-rate =	(FP)/(FP+FN)
Precision =	(TP)/(TP+ FP)
F-measure =	(2TP)/(2TP+FP+FN)

5.2 분석결과

각 데이터 셋에 세부 클래스와의 연관성이 높고 분별력 있는 유전자들을 information gain, gini index, 1:1 svm, t-statistic을 기반해서 순위대로 나열한 후에, 유전자를 1위부터 50위, 100위, 150위, 200위, 250위까지 샘플링한 4개의 서브 데이터 셋을 만들었다. 이렇게 선택된 각 특징추출 방법 별로 분별력 있는 유전자 개수에 따라 WEKA를 이용해 앞서 설명한 5가지의 기계학습 알고리즘으로 암 클래스 분류 모델을 만들고 10-fold cross-validation을 사용하여 정확도를 측정하고 서로 비교분석하였다. 그리고 특징추출을 하지 않았을 때와 특징추출을 했을 때의 성능을 비교하기 위해 전체 데이터 셋에도 알고리즘을 적용하여 분석하였다. 또한 알고리즘 별로 평균 걸리는 시간을 계산한 결과 ALL_AML 데이터 셋의 경우 각각의 특징추출 방법에 대해 적용한 알고리즘의 성능이 대부분 좋게 나왔다.

지면 관계상 각 데이터셋에 대해서 200개의 유전자로 만든 분류 모델의 정확도 와 평균시간, 데이터셋을 클래스 별로 성능평가한 내용을 정리하였다 (표3~5). ALL_AML 데이터 셋의 경우 각각의 특징추출 방법에 대해 적용한 알고리즘의 성능이 대부분 좋게 나왔다. 세부적으로 살펴보면 1:1svm을 사용해 특징 추출한 방법은 다른 알고리즘에 비해 SVM과 신경망 알고리즘에서 높은 성능을 보였다. 또한 Decision Tree는 모

든 경우에서 제일 낮은 성능을 나타냈다. 이는 분별 트리를 만드는 과정에서 각 단계마다 동적으로 해당 노드 안의 샘플들을 분류하는데 있어 가장 분별력 있는 유전자를 선택 하므로, 초기에 미리 유전자 리스트를 50개, 100개, 150개, 200개, 250개로 제한하는 것이 오히려 모델의 성능을 저하시키는 요인이 되었던 것 같다. MLL 데이터 셋은 클래스가 3개이다. 그런데 특징추출 방법 1:1svm 과 t-statistic는 클래스가 2개인 것만 가능하다. 이에 3개 클래스를 2개씩 조합하여 실험한 다음 각각의 value 값을 통해 우선순위를 다시 정해 분석하였는데 각각의 특징추출 방법에 대해 적용한 알고리즘의 성능이 대부분 좋게 나왔다. 특히 Information gain 방법을 사용하여 선택한 50개의 유전자로 SVM알고리즘을 사용해서 모델을 구축했을 때, 정확도가 100%를 나타내었다.

전체적으로 보면, 특징추출 방법 중 Information gain을 사용하여 5가지 알고리즘을 적용 했을 때 다른 특징추출 방법을 사용 했을 때 보다 성능이 좋게 나왔다. 클래스가 7개인 SALL 데이터의 경우 유전자의 개수를 줄일 경우 오히려 정확도가 낮아지는 경향을 보였다. 또한 특징추출 방법인 1:1svm을 사용 하였을 때가 모든 알고리즘에서 가장 낮은 성능을 보였다.

클래스가 2개인 ALL_AML 데이터 셋과 클래스가3개인 MLL데이터 셋은 전체적으로 대부분의 알고리즘이 결과가 좋게 나왔기 때문에 어떤 알고리즘을 사용 했을 때 더 효율적인지를 크게 구별 할 수 가 없었다. 그렇기 때문에 성능 대비 시간을 고려하여 좀더 효율적인 알고리즘을 구별 하려고 한다. 성능 대비 시간을 계산 하기 위해 평균 실행 시간을 계산하였는데 결과는 신경망을 제외한 나머지 알고리즘들은 대부분 시간이 짧게 걸렸다. 성능지표를 살펴보면 정확성(accuracy)은 MLL Leukemia가 가장 높았고 성능 이 좋다는 것을 알 수 있다.

표3. ALL-AML Leukemia 분석결과

평균실행시간					
	NB	KNN	DT	SVM	신경망
Gene=200	1초	7초	2초	10초	3시간
Gene의 개수=200					
	NB	KNN	DT	SVM	신경망
Inforgation gain	97.2%	98.6 %	90.3 %	98.6 %	98.6 %
Gini index	97.2 %	98.6 %	90.3 %	98.6 %	98.6 %
1:1 SVM	97.2 %	93.1 %	90.3 %	98.6 %	98.6 %
t-statistic	98.6 %	97.2 %	95.8 %	97.2 %	97.2 %
성능지표	**accuracy=98.2%**				
Class	TP Rate	FP Rate	Precision	Recall	F-M
ALL	1	0.04	1.979	1	0.989
AML	0.96	0	1	0.96	0.98

표4. MLL Leukemia 분석결과

평균실행시간					
	NB	KNN	DT	SVM	신경망
Gene=200	1초	7초	2초	10초	4시간
Gene의 개수=200					
	NB	KNN	DT	SVM	신경망
Inforgation gain	95.8 %	97.2 %	87.5 %	98.6 %	95.8 %
Gini index	98.6 %	97.2 %	93.1 %	97.2 %	97.2 %
1:1 SVM	97.2 %	97.2 %	88.9 %	95.8 %	97.2 %
t-statistic	95.8 %	97.2 %	91.7 %	98.6 %	97.2 %
성능지표	**accuracy=100%**				
Class	TP Rate	FP Rate	Precision	Recall	F-M
ALL	1	0	1	1	1
MLL	1	0	1	1	1
AML	1	0	1	1	1

표5. SALL분석결과

평균실행시간					
	NB	KNN	DT	SVM	신경망
Gene=200	17초	1분12초	40초	3분 20초	12시간
Gene의 개수=200					
	NB	KNN	DT	SVM	신경망
Inforgation gain	89.3 %	89.3 %	75.5 %	91.1 %	90.8 %
Gini index	90.2 %	91.7 %	81.3 %	92.4 %	90.8 %
1:1 SVM	73.1 %	75.2 %	74.8 %	75.2 %	73.1 %
t-statistic	90.2 %	91.7 %	81.7 %	93.1 %	92.1 %
성능지표	**accuracy=92.1%**				
Class	TP Rate	FP Rate	Precision	Recall	F-M
BCR-ABL	0.667	0.01	0.769	0.667	0.714
E2A-PBX1	1	0	1	1	1
Hyperdip>50	0.906	0.038	0.853	0.906	0.879
MLL	0.95	0.007	0.905	0.95	0.927
OTHERS	0.823	0.036	0.878	0.823	0.85
T-ALL	1	0	1	1	1
TEL-AML1	1	0.008	0.975	1	0.988

5. 결론

본 논문에서는 백혈병에 대한 마이크로어레이 데이터를 사용하여 정보공학적 방법으로 분별력 있는 유전자들을 추출한 후, Naïve Bayes, KNN, Decision Tree, SVM 알고리즘, 신경망 알고리즘을 이용하여 클래스 분류 모델을 구축하고, 성능을 비교분석 하였다. 실험결과 들을 비교분석한 결과 대략적인 성능의 패턴을 추정할 수 있었고 성능이 낮은 알고리즘에 대해서도 원인을 찾을 수 있었다. 실제 전체 데이터 셋을 사용하는 것 보다 분별력 있는 유전자들을 추출해 분석을 하는 것이 훨씬 더 좋은 성능을 내며 효율적이고 클래스가 적은 데이터 셋에서는 대부분의 성능이 비슷하지만 클래스가 많아질수록 SVM 과 Neural Network 알고리즘이 보다 성능이 좋다. 또한 모든 경우의 성능을 비교 분석했을 때 특징추출 방법으로 Information Gain을 사용하고 분류기법으로 SVM 알고리즘을 사용 했을 때가 가장 효율적이었다.

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