제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

암 분류를 위한 기계학습 분류기의 성능평가

원홍희 조성배
연세대학교 컴퓨터과학과
cool@candy.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Performance Evaluation of Machine Learning Classifiers for Cancer Classification

Hong-Hee Won and Sung-Bae Cho
Dept of Computer Science, Yonsei University

요 약

Microarray 기술의 발전으로 많은 양의 유전자 정보를 얻게 되어 암의 정확한 분류와 진단에 대한 기대가 커지고 있다. 암을 정확하게 분류하기 위해서는 추출된 유전자에 많은 잡음이 들어가기 때문에 암과 관련이 있는 유전자만을 추출할 필요가 있다. 본 논문에서는 여러 가지 유전자 추출방법과 다양한 분류기의 성능을 체계적으로 평가하기 위하여, 세 가지 벤치마크 암 데이터에 대하여 실험하여 보았다. 또한 분류 성능을 향상시키기 위하여 분류기를 적절하게 결합한 결과, 결합된 분류기의 성능을 확인해볼 수 있었다.

1. 서론

복잡한 조절 기능을 갖는 생명 현상에 대한 분자 수준의 단편적인 이해는 한계가 있기 때문에 Human Genomic Project(HGP)와 같이 전체적인 이해를 위한 연구의 필요성이 대두되었다. 이 때, 염기서열의 기능을 이해하는 것이 필수적이기 때문에 이러한 과정에서 DNA 칩이 개발되었다. 최근의 cDNA microarray와 oligonucleotide microarray 기술의 발전은 엄청난 양의 유전자 정보를 제공하고 있다. DNA 칩 기술의 발전은 유전자 정보의 대량 생산을 가능하게 하였고, 특정한 실험 환경과 조건에 따른 수천 개의 유전자 발현 정도를 동시에 파악할 수 있고, 이를 대량으로 처리함으로써 수천 개의 유전자 정보를 굉장히 빠르고 정확하게 분석할 수 있게 되었다[1].

유전자 microarray 기술의 발전은 암의 정확한 예측과 진단 분야에도 적용되어 많은 도움을 줄 것으로 예상된다. 특히 암의 정확한 분류는 암의 치료에 있어서 매우 중요한 이슈가 되기 때문에 유전자 정보를 이용하여 암을 분류하는 문제에 관한 많은 연구가 진행되고 있다[2][3]. 분류에 필요한 정보를 주는 유전자를 선택하기 위한 다양한 특징 추출 방법과 여러 분류기를 이용하여 실험한 결과를 분석하고, 특징의 상관관계를 기준으로 각 분류기의 결과를 결합함으로써 분류 성능을 향상시키려는 연구가 있었다[4]. 하지만 하나의 벤치마크 데이터에 대하여 실험하였기 때문에 실험 결과에 대한 충분한 검증이 되지 않았다. 따라서 다양한 벤치마크 데이터를 이용하여 분류기의 성능을 체계적으로 분석해 볼 필요가 있다.

본 논문에서는 leukemia cancer, colon cancer, lymphoma cancer dataset 등 세 가지의 벤치마크 데이터 집합에 다양한 유전자 선택 방법과 분류 방법을 적용하여 그 성능을 비교 평가하고, 나아가 각 분류기의 분류 결과를 적절하게 결합하여 결합되어진 분류기의 분류 성능이 향상됨을 보이고자 한다.

2. 기계학습기반 분류 시스템

유전자 데이터로부터 분류에 필요한 정보만을 추출하여, 분류기를 학습시키고, 학습된 분류기를 이용하여 새로운 데이터를 분류하는 것을 분류 시스템이라 한다. 이를 위해서 필요한 정보만을 추출하는 특징 추출 방법과 이 정보를 이용하여 패턴을 분류하는 분류기를 적절하게 선택하여야 한다.

2.1 유전자 선택

Microarray로부터 얻어지는 유전자의 수는 대략 수천 개에서 수만 개이다. 하지만 얻어진 데이터에서 각 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 하며, 유전자 선택 과정이라고도 한다[2]. 유전자와 적은 수의 샘플을 포함하는 데이터이므로 많은 수의 유전자를 이용하여 적은 수의 샘플을 분류하는 것은 정확한 분류를 위하여 적절치 않다. 분류에 이용하기 적절한 유전자를 선택하는 과정이 필요하게 되는데, 주로 상관관계와 클러스터링 기법 등을 이용하여 클래스와의 상관관계가 높은 유전자를 선택한다.

M개의 샘플과 N개의 유전자를 갖는 $M \times N$ 행렬 데이터가 있고, M 개의 샘플은 클래스 A와 클래스 B로 나뉜다고 하자. 처음의 $M/2$개의 샘플을 암 세포의 샘플이라 하고, 나머지 $M/2$개의 샘플을 정상 세포의 샘플이라고 하면, 각 유전자 데이터는 수식 (1)과 같은 벡터로 표현할 수 있다.

$$G_i = (e_1, e_2, e_3, \ldots, e_M) \quad (1)$$

클래스에 대한 특징을 뚜렷하게 나타내는 이상적인 유전자를 G_{ideal} 이라고 했을 때, 암 세포의 특징을 1로 정의하고 정상 세포의 특징을 0으로 정의한다면, G_{ideal} = (1, 1, 1, ..., 1, 0, 0, ..., 0)으로 표현할 수 있다. 데이터의 각 유전자 G_i와 G_{ideal}의 상관관계를 측정할 수 있는 다양한 척도를 사용하여 분류에 있어 중요한 정보를 줄 수 있는 유전자를 선택할 수 있다. 본 논문에서 사용한 상관관계 척도는 표 1에 정리하였다.

표 1에서 각 수식은 순서대로 Pearson correlation coefficient(PC), Spearman correlation coefficient(SC), Euclidean distance(ED), cosine coefficient(CC), information gain(IG), mutual information(MI), signal to noise ratio(SN)를 의미한다.

표 1. 유전자 선택을 위한 상관관계 척도

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

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

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

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

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

$$IG(G_i, C_j) = P(G_i, C_j)\log\frac{P(G_i, C_j)}{P(C_j)P(G_i)} + P(\overline{G_i}, C_j)\log\frac{P(\overline{G_i}, C_j)}{P(C_j)P(\overline{G_i})}$$

$$MI(G_i, C_j) = \log\frac{P(G_i, C_j)}{P(G_i)P(C_j)}$$

$$SN(G_i) = \frac{\mu_1(G_i) - \mu_2(G_i)}{\sigma_1(G_i) - \sigma_2(G_i)}$$

2.2 분류

기계학습 분야에서 분류문제를 위해서 개발되어진 많은 알고리즘이 최근 유전자 정보를 이용하여 암을 예측하고, 분류하는 연구에 적용되어 왔다. 주어진 학습 데이터로부터 정확한 분류 결과를 낼 수 있도록 분류기를 학습하고 학습되어진 분류기에 실험 데이터로 분류를 하는 것이 일반적인 기계 학습에서의 분류 과정이다. 대표적인 분류 알고리즘으로 다층신경망(multi-layer perceptron, MLP), KNN (k-nearest neighbor), SVM (support vector machine) 등이 있으며, 클러스터링 기법인 SOM (self-organizing map, SOM) 등도 분류에 사용된다.

1) MLP

MLP는 인공신경망의 대표적인 기계 학습 알고리즘으로서, 일반적인 패턴 인식 문제에서 강하고 안정적인 성능을 보인다[6]. 역전파(back-propagation) 알고리즘을 사용하여 신경망의 결과 값이 분류 목표치에 가까워지도록 신경망의 가중치를 조절해나감으로써 주어진 패턴을 학습한다.

2) KNN

KNN은 샘플들의 상관관계나 유사도를 이용하여 비슷한 샘플들의 분류 결과를 참조하는 방식에 기반 한다. KNN 알고리즘은 분류하고자 하는 샘플을 입력받은 후에 Pearson correlation coefficient와 같은 상관관계 척도 혹은 Euclidean distance와 같은 유사도 척도를 이용하여 입력 샘플과 가장 유사한 k 개의 샘플을 찾는다. 선택된 k 개 샘플들의 분류 결과에 분류하고자 하는 샘플과의 유사도를 가중치로 곱하여 분류하고자 하는 샘플의 분류 결과를 결정한다.

3) SASOM

자기조직화지도는 Teuvo Kohonen에 의해 고안되었으며, 특정한 입력에 대해 가장 잘 반응하는 노드와 그 주위의 노드를 학습시킴으로써 자연스럽게 그 입력에 대한 대표자로서 역할하게 되는 것이다.

구조적응 자기조직화지도(structure adaptive SOM, SASOM)은 자기조직화지도의 위상이 학습을 통하여 자동으로 결정되는 분류 알고리즘이다[7]. 구조적응 자기조직화지도는 4×4로 지도의 크기를 초기화한 후에, 학습 중간에 하나의 노드에 여러 클래스가 있는 경우 그 노드를 2×2로 분화한다. 기본 알고리즘은 그림 1과 같다.

❑ SASOM (Structure Adaptive SOM)
(1) 학습에 필요한 변수를 초기화한다.
(2) SOM 알고리즘으로 학습시킨다.
(3) 지도의 노드들 중 여러 클래스의 데이터가 섞인 노드를 찾는다.
(4) 단계 (3)에서 발견된 노드들을 분화시킨다.
(5) 분화된 노드들을 LVQ알고리즘으로 학습시킨다.
(6) 학습에 참여하지 않는 노드를 삭제한다.
(7) 종료 조건이 만족될 때까지 위의 단계를 차례대로 반복한다.

그림 1. SASOM의 알고리즘

4) SVM

SVM은 1995년 Vladimir Vapnik에 의해 제안되었으며, 실제 문제에서 성능이 좋아 널리 사용되고 있다. 원래는 2 클래스 패턴 분류 문제를 해결하기 위해 제안되었다[8]. 선형 SVM은 기존의 경험적 리스크 최소화보다 월등한 성능을 갖고 있는 구조적 리스크 최소화의 통계적 학습 이론에 기반한다. 이차원 데이터 분류문제에서 가장 최적의 초평면(Hyperplane)을 구하여 이를 경계 결정면으로 선택한다. 최적의 초평면은 선형 분리가 가능한 두 집단에 대해 집단을 구분지으며, 마진을 최대로 한다.

하지만 실제 문제의 경우 선형적으로 구성되지 않기

때문에 커널 함수를 이용하여 비선형적 특징공간을 선형적 특징공간으로 매핑한 후에 선형 SVM으로 분류하게 된다. 특징공간에서의 결정 함수는 수식 (3)과 같다. $K(\vec{x}, \vec{x}') = \phi(\vec{x}) \cdot \phi(\vec{x}')$는 커널 함수를 의미하고, y_i는 각 x의 결과 값인 -1 이나 1을 의미한다.

$$f(x) = \sum_{i=1}^{N} y_i \lambda_i^* K(\vec{x}_i, \vec{x}) + b^* \qquad (3)$$

일반적으로 커널 함수의 특징에 따라 패턴 인식률의 차이가 있으므로, 문제에 맞게 적절하게 선택하여야 한다.

5) 분류기 결합 방법

하나의 분류기에 의해서 분류한 결과는 아무리 그 성능이 우수하다 하여도 신뢰적이지 못하다. 따라서 하나 이상의 분류기를 결합함으로써 더 안정적이고 좋은 성능을 기대할 수 있다.

더 좋은 분류 성능을 얻기 위하여 분류기를 결합하는 방법은 그 특징에 따라 다양한 방법으로 나누어 볼 수 있다. 하나의 인식 결과만을 출력하는 추상레벨(abstract level)과 순위의 형태로 출력하는 순위레벨(rank level), 그리고 순위와 더불어 신뢰 값까지 함께 출력하는 측정치레벨(measurement level) 등으로 분류된다.

본 논문에서는 추상 레벨의 대표적인 방법 중의 하나인 다수결 방법을 사용하였다. 다수결 결합은 결합된 분류기 집합에서 가장 많은 분류기가 내놓은 결과를 다중 분류기의 결과로서 사용하는 것이다.

3. 실험 및 결과

3.1 실험 데이터

1) Leukemia dataset

Leukemia dataset은 72개의 샘플 데이터로 구성되어 있으며, 백혈병의 두 가지 종류인 급성 골수성 백혈병(acute myeloid leukemia, AML) 환자 25명과 급성 림프성 백혈병(acute lymphoblastic leukemia, ALL) 환자 47명으로부터 얻어진 데이터이다. 72개의 샘플 데이터 중에서 63개는 골수로부터 채취하였고, 나머지 9개는 말초 혈액으로부터 채취하여 high density oligonucleotide microarray를 사용하여 만들어졌다.

72개의 샘플 중에서 38개를 학습 데이터로 사용하였고, 나머지 34개를 실험 데이터로 사용하였는데, 각 샘플은 7129개의 유전자 발현 정보를 갖고 있다.

2) Colon cancer dataset

Colon dataset은 결장암 환자의 결장 상피 세포로부터 추출한 62개의 샘플 데이터이며, 각 샘플은 2000개의 유전자 발현 정보를 갖고 있다. 원래의 데이터는 6000개의 유전자 정보를 갖고 있었지만, 정확하지 않은 정보를 갖고 있는 4000개를 제거한 것이다. 62개의 샘플 데이터 중에서 40개는 암 세포의 샘플이며, 다른 22개는 정상 세포의 샘플이다. 각 샘플은 같은 환자의 암 부위와 정상 부위의 세포에서 채취되었으며, high density oligonucleotide microarray를 사용하여 만들어졌다.

62개의 샘플 중에서 31개를 학습 데이터로 사용하였고, 나머지 31개를 실험 데이터로 사용하였다.

3) Lymphoma dataset

B cell diffuse large cell lymphoma (B-DLCL)은 형태학이나 임상적인 상태와 약물 반응에 있어 이질적인 두 가지 종류가 있다. 한 종류가 germinal center B cell-like DLCL이고 나머지가 activated B cell-like DLCL이다. Lymphoma dataset은 GC B-like 샘플 24개와 activated B-like 샘플 23개로 구성되어 있다.

47개의 샘플 중에서 22개를 학습 데이터로 사용하였고, 나머지 25개를 실험 데이터로 사용하였으며, 각 샘플은 4026개의 유전자 발현 정보를 갖고 있다.

3.2 실험 환경

MLP를 이용한 실험에서는 모멘텀은 0.9로 정하였고 총 레이어의 수는 3으로 고정한 후에 학습률을 0.01에서 0.50으로 변화시켜가며 실험을 하였다. 또한 학습 데이터에 과적합 되는 것을 막기 위하여 학습과정에서의 최대 반복은 100으로 고정하였다.

KNN의 실험에서는 가까운 인접노드의 개수 k를 1에서 8까지 변화하여 실험하였다. 그리고 샘플의 유사도를 측정하는 척도는 Pearson correlation coefficient에 의한 상관관계 척도와 Euclidean distance에 의한 유사도 척도를 이용하였다.

SASOM의 실험에서는 초기 학습률을 0.05, 최종 학습률을 0.02로 하고 초기 학습 반복수를 1000, 최종 학습 반복수를 10000으로 하였으며 초기 반경을 10, 최종 반경을 3으로 정하였다.

SVM의 실험에서는 커널 함수로 Linear 함수와 RBF 함수를 사용하였으며, RBF 커널 함수를 사용한 경우에 감마 변수를 0.1에서 0.5로 변화하여 실험하였다.

7개의 유전자 선택 방법과 6개의 분류기를 사용하여 42개의 분류 결과를 정리하였으며, 총 42개 중에서 3개의 분류기를 선택하여 모든 조합을 각각 다수결 결합으로 그 결과를 정리하였다.

3.3 실험 결과

Leukemia, colon, lymphoma dataset에 대하여 7 가지의 유전자 선택 방법과 6 가지의 분류기를 사용한 결과와 42개의 분류 결과 중에서 3개의 분류기를 선택하여 결합한 결과를 표 2, 3, 4에 각각 정리하였다.

Leukemia dataset의 경우에는 각 분류기의 가장 좋은 성능과 결합방법의 가장 좋은 성능이 같지만, colon dataset의 경우와 lymphoma dataset의 경우에는 결합방법의 결과가 더 나은 성능을 보이는 것을 알 수 있다. Colon dataset의 경우 분류기의 최고 인식률을 83.87%인데 반하여, 결합방법의 최고 인식률은 93.55%로 분류 성능이 향상된 것을 볼 수 있다. 또한 Lymphoma dataset의 경우에도 분류기의 최고 인식률은 92.00%인데 반해, 결합방법의 최고 인식률은 96.00%로 분류 성능이 향상되었음을 알 수 있다.

표 2. 각 유전자 선택 방법과 분류기에 따른 인식률과 분류기의 결합에 의한 인식률(Leukemia)

	MLP	KNN		SASOM	SVM		Ensemble
		Cosine	Pearson		Linear	RBF	
PC	97.06	97.06	94.12	76.47	79.41	79.41	97.06
SC	82.35	76.47	82.35	61.76	58.82	58.82	
ED	91.18	85.29	82.35	73.52	70.59	70.59	

CC	94.12	91.18	94.12	88.24	85.29	85.29	
IG	97.06	94.12	97.06	91.18	97.06	97.06	
MI	58.82	73.53	73.53	58.82	58.82	58.82	
SN	76.47	73.53	73.53	67.65	58.82	58.82	

표 3. 각 유전자 선택 방법과 분류기에 따른 인식률과 분류기의 결합에 의한 인식률(Colon)

	MLP	KNN		SASOM	SVM		Ensemble
		Cosine	Pearson		Linear	RBF	
PC	74.19	70.97	77.42	74.19	64.52	64.52	
SC	58.06	61.29	67.74	45.16	64.52	64.52	
ED	67.74	83.87	83.87	67.64	64.52	64.52	
CC	83.87	80.65	80.65	64.52	64.52	64.52	93.55
IG	70.97	74.19	80.65	70.97	70.97	70.97	
MI	70.97	74.19	80.65	70.97	70.97	70.97	
SN	64.52	64.52	70.97	45.16	64.52	64.52	

표 4. 각 유전자 선택 방법과 분류기에 따른 인식률과 분류기의 결합에 의한 인식률(Lymphoma)

	MLP	KNN		SASOM	SVM		Ensemble
		Cosine	Pearson		Linear	RBF	
PC	64.00	60.00	76.00	48.00	56.00	60.00	
SC	60.00	60.00	60.00	68.00	44.00	44.00	
ED	56.00	56.00	68.00	52.00	56.00	56.00	
CC	68.00	60.00	72.00	52.00	56.00	56.00	96.00
IG	92.00	92.00	92.00	84.00	92.00	92.00	
MI	72.00	80.00	64.00	64.00	64.00	64.00	
SN	76.00	76.00	80.00	76.00	72.00	76.00	

그림 2는 colon dataset의 결합방법에서 93.55%의 인식률을 보였을 때의 결합된 분류기들을 나타낸다. 가장 좋은 성능을 보인 결합방법으로 결합된 분류기를 살펴보면 결합된 분류기의 종류보다 특징 추출 방법이 분류 성능 향상에 더 많은 영향을 끼친 것을 볼 수 있다.

MLP : Cosine coefficient KNNcosine : Euclidean distance method KNNcosine : Pearson's correlation coefficient
MLP : Cosine coefficient KNNcosine : Euclidean distance method KNNpearson : Pearson's correlation coefficient
MLP : Cosine coefficient KNNcosine : Euclidean distance method SASOM : Pearson's correlation coefficient
MLP : Mutual information KNNcosine : Euclidean distance method KNNpearson : Pearson's correlation coefficient
MLP : Information gain KNNcosine : Euclidean distance method KNNpearson : Pearson's correlation coefficient
MLP : Cosine coefficient MLP : Pearson's correlation coefficient KNNpearson : Euclidean distance method
KNNpearson : Euclidean distance method KNNpearson : Mutual information SASOM : Pearson's correlation coefficient
KNNpearson : Euclidean distance method KNNpearson : Information gain SASOM : Pearson's correlation coefficient

그림 2. colon dataset의 ensemble에서 93.55%의 인식률을 보였을 때의 결합된 분류기

Colon dataset의 경우에 결합된 세 개의 분류기 중에서 두개는 반드시 Euclidean distance와 Pearson correlation coefficient에 의해 특징 추출된 분류기이고, 나머지 하나의 경우만 cosine coefficient이거나 mutual information, information gain에 의해 특징 추출된 분류기임을 알 수 있다. 하지만 나머지 분류기의 경우에도 cosine coefficient에 의해 특징 추출된 경우가 많았다.

이러한 경향은 lymphoma dataset의 경우에는 더 분명하게 나타났는데, 결합된 세 개의 분류기 중에서 대부분이 information gain, signal to noise ratio, Euclidean distance이거나 information gain, signal to noise ratio, Pearson correlation coefficient에 의해서 특징 추출된 분류기였다.

특징추출 방법이 결합방법에서 중요한 의미를 갖는 이유는 하나의 특징 추출 방법만으로는 분류하고자 하는 해 공간을 전부 포함하기 불가능하지만 다양한 특징 추출 방법을 결합하게 되면 다른 특징 추출 방법에 의해 포함되지 않은 해 공간 영역을 또 다른 특징 추출 방법이 포함해주기 때문이다. 해 공간 탐색의 범위를 넓혀 줌으로써 한쪽의 해 공간에 치우쳐진 분류가 일어나는 것을 막을 수 있을 것이다.

참고문헌

[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] J. W. Ryu and S. B. Cho, "Towards optimal feature and classifier for gene expression classification of cancer," *Lecture Note in Artificial Intelligence*, vol. 2275, pp. 310-317, 2002.

[5] A. Ben-Dor, L. Bruhn, N. Friedman, et al., "Tissue classification with gene expression profiles," *Proc. The Fourth Annual International Conf. on Computational Molecular Biology*, pp. 1-32, May 2000.

[6] R. P. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Magazine*, pp. 4-22, April 1987.

[7] S. B. Cho, "Ensemble of structure-adaptive self-organizing maps for high performance classification," *International Journal of Information Sciences*, vol. 123, pp. 103-114, March 2000.

[8] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.