유전자 알고리즘을 이용한 림프종 암의 최적 분류기 앙상블

박찬호°, 조성배
연세대학교 컴퓨터과학과
cpark@candy.yonsei.ac.kr°, sbcho@csai.yonsei.ac.kr

Optimal Classifier Ensemble for Lymphoma Cancer Using Genetic Algorithm

Chanho Park° and Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

DNA microarray기술의 발달로 한꺼번에 수천 개 유전자의 발현 정보를 얻는 것이 가능해졌는데, 이렇게 얻어진 데이터를 효과적으로 분류하는 시스템을 만들어놓으면 새로운 샘플이 정상상태인지, 질병을 가진 상태인지 예측할 수 있다. 분류 시스템을 위하여 여러 가지 특징선택방법들과 분류기법들을 사용할 수 있는데, 모든 상황에서 항상 뛰어난 성능을 보이는 특징선택법이나 분류기를 찾기는 힘들다. 안정되고 개선된 성능을 내기 위해서 특징-분류기의 앙상블을 이용할 수 있는데, 앙상블에 이용될 수 있는 특징선택방법이나 분류기의 수가 많다면, 앙상블을 만들 수 있는 조합이 많아지기 때문에, 모든 조합에 대하여 앙상블 결과를 구하기는 거의 불가능하다. 이를 해결하기 위하여 본 논문에서는 유전자알고리즘을 이용하여 모든 앙상블 결과를 계산하지 않으면서 최적의 앙상블을 찾아내는 방법을 제안하였으며, 실제로 림프종 암 데이터에 적용한 결과 100%의 결합결과를 보이는 최적의 앙상블을 효과적으로 찾아내었다.

1. 서 론

지난 몇 년간 암의 조기 발견과 정확한 분류를 위한 연구가 활발하게 진행되어 왔지만, 완벽한 방법을 제시한 연구는 아직 없었다. 이는 암의 원인이 되는 경로가 다양할 뿐만 아니라, 다양한 변이가 존재하며, 또한 대량의 데이터를 얻기도 힘들었기 때문이다. 그러나 최근 DNA microarray기술의 발달로 생명체에 관한 대량의 유전정보를 얻는 것이 가능해졌다. 이렇게 얻어진 정보의 형태는 단순한 숫자들의 나열이므로, 그 속에서 의미를 발견할 수 있도록 효과적인 분석이 필요하다. 이와 관련하여 많은 연구 결과가 발표되어 왔으며 현재도 활발하게 진행 중이다[1].

DNA microarray로부터 나온 데이터는 특징선택을 통하여 그 중 의미 있는 유전자들이 선택되고, 분류기를 통하여 분류 결과를 내어 놓는다. 다양한 특징선택법과 분류기가 존재하며, 둘을 곱한 수만큼의 특징-분류기가 존재한다. 본 논문에서는 7가지 특징선택법과 6가지 분류기를 사용하였으므로 42가지 특징-분류기가 존재한다. 한편 그 중 항상 좋은 결과를 보장하는 특징-분류기가 없으므로 이들의 앙상블을 통하여 개선된 성능을 기대할 수 있는데 42개의 특징-분류기로 만들 수 있는 앙상블 조합은 무려 2^{42}가지가 존재하며 이를 모두 구하기는 사실상 불가능하다. 따라서 본 논문에서는 유전자알고리즘을 이용하여 효과적으로 앙상블 공간을 탐색하는 방법을 제안한다. 그리고 잘 알려진 벤치마크 데이터인 림프종 데이터를 이용하여 제안한 방법의 타당성을 평가한다.

2. DNA microarray

DNA microarray는 용액이 투과하지 않는 딱딱한 지지체 위에 고밀도로 cDNA를 고정시켜 놓은 것이다. Array를 구성하는 각 셀은 두 개의 다른 환경에서 채집된 유전물질에 녹색의 Cy3와 빨간색의 Cy5라는 형광물질을 합성한 것을 동일한 양으로 보합한 것이다. 이것을 스캐너로 읽어 들이면 녹색부터 빨간색에 이르는 발현정도를 얻을 수 있는데, Cy5/Cy3의 비율에 로그를 취한 값을 그 셀의 발현정보 값으로 얻게 된다[2].

$$gene\_expression = log_2 \frac{Int(Cy5)}{Int(Cy3)} \quad (1)$$

암 샘플과 정상 샘플에서 관련 유전자의 발현 정도는 서로 다른 양상을 보이게 되므로, DNA microarray상에서 얻은 유전발현 정보를 이용하여 분류를 할 수 있다.

3. 최적의 특징-분류기 쌍 앙상블 탐색

본 논문에서는 그림 1과 같은 분류구조를 제안한다. 림프종 데이터에 대하여 다수의 특징-분류기가 만들어지며, GA는 그들을 이용한 수많은 앙상블 중 최적의 것을 찾아 탐색하게 된다.

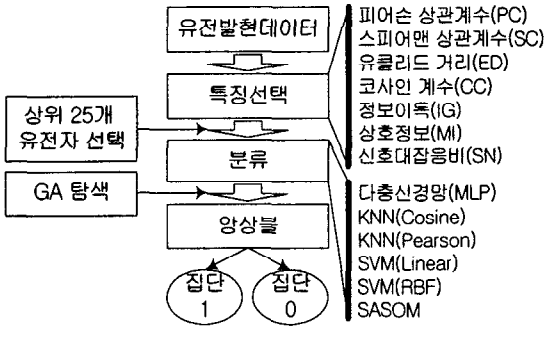

그림 1. 앙상블 탐색 시스템

3.1 유전자 선택과 분류기

유전자 선택 과정에서는 분류에 도움이 될 것이라 예상되는 유전자들을 선택하게 된다. 본 논문에서는 통계

적 상관계수에 기반을 둔 피어슨 상관계수(PC)와 스피어맨 상관계수(SC), 유사도에 기반을 둔 유클리드 기리(ED)와 코사인 계수(CC), 마지막으로 정보이론에 기반을 둔 정보이득(IG), 상호정보(MI), 신호 대 잡음 비(SN)를 유전자 선택 방법으로 사용하여, 각 방법별로 순위를 매긴 후 상위 25개의 유전자를 의미 있는 유전자로 정하여 분류에 이용하였다.

각 방법은 표 1의 식들을 이용하여 구할 수 있다. X, Y는 비교하고자 하는 두 벡터를 의미하며, 특정 유전자의 i번째 샘플이 특정 집단 c에 속하는 가의 여부와 그 유전자가 발현했는가의 여부에 의하여 A, B, C, D를 구분할 수 있다. 신호 대 잡음 비에서는 주어진 유전자 g를 집단 c에 속하는 것들과 그렇지 않은 것들로 나누고, 각각에 대하여 구한 평균과 표준편차를 이용한다.

표 1. 특징선택방법들

$$PC = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (2)$$

$$SC = 1 - \frac{6\sum (Dx - Dy)^2}{N(N^2-1)} \quad (3)$$

$$ED = \sqrt{\sum (X - Y)^2} \quad (4)$$

$$CC = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}} \quad (5)$$

$$IG = A \cdot \log \frac{A}{(A+B)\cdot(A+C)} + B \cdot \log \frac{B}{(A+B)\cdot(B+D)} \quad (6)$$

$$MI = \log \frac{A}{(A+B)\cdot(A+C)} \quad (7)$$

$$SN = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)} \quad (8)$$

분류기로는 패턴인식분야에서 널리 사용되는 다층신경망(MLP)과 kNN 및 근래에 많이 사용되는 SVM과 SOM의 단점을 개선한 SASOM을 사용하였다. MLP는 노드간의 가중치를 조절하여 입력패턴이 목표출력에 최대한 가까운 값을 낼 수 있도록 해주며, 학습방법으로는 오류 역전파 알고리즘을 사용하였다. KNN은 기억기반 추론방법으로서 테스트 샘플과 학습 샘플들과의 거리를 계산하여 가장 가까운 k개에 대하여 어느 집단이 가까운지를 알아보는 분류기이며, 거리를 측정하는 방법으로는 피어슨 상관계수(KNN(P))와 코사인 계수(KNN(C))의 두 가지를 사용하였다. SVM은 학습 샘플들에 대하여 최적의 초평면을 찾아주는 방법인데, 선형 분리가 불가능한 경우 커널함수를 사용하여 선형의 특징공간으로 사상시킨다. 선형커널(SVM(L))과 RBF커널(SVM(R))을 커널함수로 사용하였다. SASOM은 학습이 시작되기 전에 구조를 결정해야 하는 SOM의 단점을 보완하기 위하여 고안된 방법이다.

3.2 분류기 앙상블

특징선택방법과 분류기의 조합을 통하여 다양한 특징-분류기들이 만들어지지만, 데이터와 환경에 상관없이 항상 좋은 결과를 내는 것을 찾기는 매우 힘들다. 분류기 앙상블은 특징-분류기의 결합을 통해 이뤄지는데, 이런 상황에서 유용하게 사용된다[3].

앙상블의 다른 장점으로는 단일 특징-분류기에 비하여 비교할 수 없을 정도의 다양한 결과를 낼 수 있다는 것이다. 본 논문에서는 42가지의 특징-분류기가 사용되었고 이는 42가지의 분류 결과만을 의미하지만, 이들을 사용하여 만들 수 있는 앙상블은 무려 2^{42}가지나 된다. 앙상블의 방법에 따라 더욱 다양한 결과를 얻을 수 있다. 본 논문에서는 투표 방법을 사용하였으며, 이는 앙상블에 참여한 특징-분류기 출력의 다수결을 구해서 테스트 샘플에 대한 출력으로 결정하는 방식이다.

3.3 GA를 이용한 앙상블 탐색

분류기 앙상블이 가능한 모든 앙상블에 대하여 그 결과를 구하려면 2^{42}가지에 대하여 모두 시도해 보아야 한다. 이는 사실상 불가능에 가까우며 모든 앙상블 결과가 좋은 것만은 아니다. 따라서 효과적으로 앙상블 공간을 탐색하는 방법이 필요하며, 본 논문에서는 GA를 이용한 방법을 제안한다.

최적의 앙상블을 탐색하기 위한 GA의 염색체는 그림 2와 같이 42비트로 이루어져 있으며, 각 비트는 사전에 정의된 서로 다른 특징-분류기에 대응된다. 각 비트의 의미는 대응하는 특징-분류기를 앙상블에 이용할지 안할지의 여부이다. 그림 2의 염색체는 2, 3, 6번째 비트가 1이므로 그 비트들에 대응하는 3개의 특징-분류기가 앙상블에 이용된다.

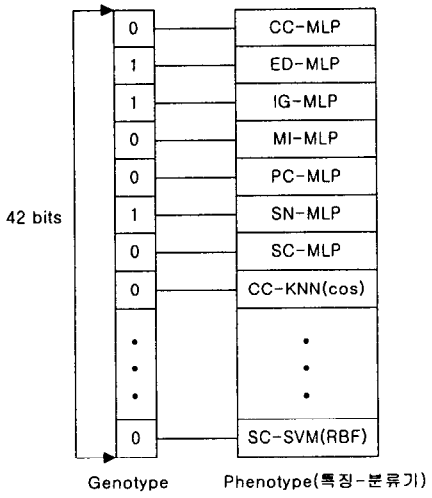

그림 2. GA의 염색체 구조

초기집단을 이루는 염색체들이 임의로 생성되면, 그들은 각각 앙상블 결과에 대하여 적합도를 평가받는다. 그 후 적합도에 걸맞은 선택확률을 부여받고 선택과정에서 선택된다. 선택된 염색체들은 둘씩 짝을 이루어 교차율에 의하여 교차여부를 결정하고, 마지막으로 각 비트는 돌연변이 확률에 의하여 돌연변이를 결정한다. 유전연산이 끝난 염색체들은 다음 세대의 부모가 되어 적합도를 평가받는 부분부터 다시 반복된다.

4. 실험 및 결과

4.1 실험 환경

실험 데이터로는 암 관련 유전발현 데이터인 림프종 데이터(http://llmpp.nih.gov/lymphoma)를 사용하였다. 이 데이터는 4026개의 유전자로 구성되어 있고, 두 집단으로 구분되는 47개의 샘플을 가지고 있다. 이중 22개를 학습에 이용하고, 25개를 테스트에 이용하였다.

MLP의 경우 목표인식률 98%, 최대 반복회수 500회, 은닉노드 8개를 두었으며, 학습률은 0.01~0.5, 모멘텀은 0.9로 정하였다. SASOM은 4×4 크기의 지도를 초기에 사용하였다. GA는 집단의 크기를 100, 200, 500, 1000, 1500, 2000으로 변화시켜가며 실험하였고, 각 집단 크기에 대하여 0.3, 0.5, 0.7, 0.9의 교차율과 0.01, 0.05의 돌연변이율을 적용시켜보았으며, 교차방법으로는 일점교차, 선택방법으로는 룰렛 휠 방법을 사용하였다. 돌연변이는 빠른 수렴을 위하여 0에서 1로 바뀌는 비율을 1에서 0으로 바뀌는 비율의 반으로 정하였다.

4.2 단일 특징-분류기 실험 결과

림프종 데이터에 대하여 7가지 특징선택법과 6가지 분류기의 조합을 적용시킨 결과가 표 2에 있다. 전체적으로 특징선택법은 IG가, 분류기는 *K*NN(C)이 상대적으로 우수하였으며, 전체 평균 인식률은 67.1%였다.

표 2. 림프종 데이터에 대한 단일 특징-분류기 결과

	MLP	KNN(P)	KNN(C)	SASOM	SVM(L)	SVM(R)	평균
CC	68.0	60.0	72.0	52.0	56.0	56.0	60.7
ED	56.0	56.0	68.0	52.0	56.0	56.0	57.3
IG	92.0	920.	92.0	84.0	92.0	92.0	90.7
MI	72.0	80.0	64.0	64.0	64.0	64.0	68.0
PC	64.0	60.0	76.0	48.0	56.0	60.0	60.7
SN	76.0	76.0	80.0	76.0	72.0	76.0	76.0
SC	60.0	60.0	60.0	68.0	44.0	44.0	56.0
평균	69.7	69.1	73.1	63.4	62.9	64.0	67.1

4.3 GA를 이용한 최적 앙상블 탐색 결과

최적의 앙상블을 찾기 전에 GA가 옳은 방향으로 진화하고 있는지의 여부를 알아보기 위하여 평균 적합도의 추이를 살펴보았다. 그 결과 그림 3과 같이 세대가 지날수록 평균적합도가 상승하는 모습을 보여주었고, 따라서 GA가 해에 다가가고 있음을 알 수 있었다.

GA를 이용하여 최적 앙상블을 찾아본 결과, GA는 100%의 인식률을 보이는 앙상블을 찾아주었다. 비록 단일 특징-분류기의 성능이 상대적으로 좋지 않은 것들도 많이 있었지만, GA는 유전연산을 통하여 그들의 상보적 결합을 발견하여 최적해를 찾아주었다. 한편 GA가 최적해를 발견하는 세대는 그림 4와 같이 집단을 이루는 염색체 수에 반비례하는 경향을 보여주었다.

5. 결론

분류기의 성능을 높이기 위하여 단일 특징-분류기의 결합을 이용하는 앙상블을 시도할 수 있는데, 사용되는 특징-분류기의 수가 많을수록 소요되는 시간은 그 수에 지수적으로 비례하여 증가한다. 이때 GA를 이용함으로써 적은 시간을 들여 효과적으로 최적 앙상블을 찾아낼 수 있었다.

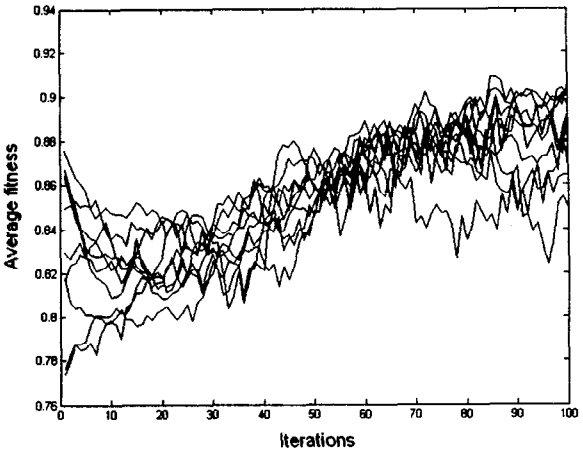

그림 3. 세대에 따른 평균 적합도 추이

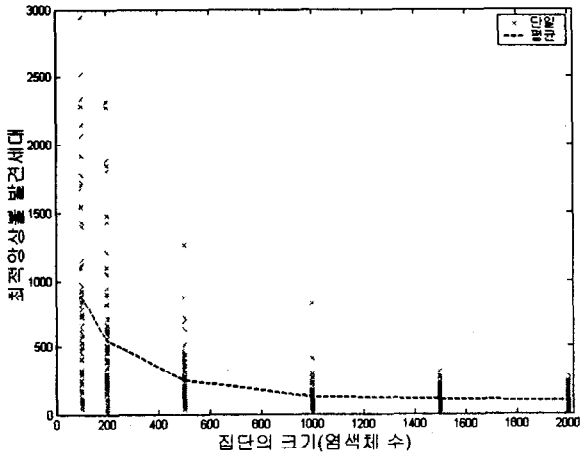

그림 4. 염색체 수와 최적해 발견 세대와의 관계

감사의 글

본 연구는 보건복지부 보건의료기술 진흥사업의 지원에 의하여 이루어진 것임.

참고 문헌

[1] T. R. Golub, et al., "Molecular classification of cancer class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, no. 15, pp. 531-537, October 1999.

[2] J. Quackenbush, "Computational analysis of microarray data," *Nature Reviews Genetics*, vol. 2, pp. 418-427, June 2001.

[3] S.-B. Cho, and J.-W. Ryu, "Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features," *Proc. of the IEEE*, vol. 90, no. 11, pp. 1744-1753, 2002.