DNA 마이크로어레이 데이터의 분류를 위한 종분화 진화 기반의 최적 다중 분류기

박찬호°, 조성배
연세대학교 컴퓨터과학과
cpark@sclab.yonsei.ac.kr°, sbcho@cs.yonsei.ac.kr

Multiple Optimal Classifiers based on Speciated Evolution for Classifying DNA Microarray Data

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

요 약

DNA 마이크로어레이 기술의 발전은 암의 조기 발견 및 예후 예측을 가능하게 해주었으며, 이와 관련된 많은 연구가 진행 중이다. 마이크로어레이 데이터의 분류에서 관련 유전자들의 선택은 필수적이며, 유전자 선택방법은 분류기와 짝을 이루어 특징-분류기를 형성한다. 이제까지 여러가지 특징-분류기를 사용하여 마이크로어레이 데이터를 분류해 왔지만, 알고리즘의 한계와 데이터의 결함 등으로 인하여 최적의 특징-분류기를 찾기 어려웠다. 따라서 앙상블 분류기를 이용하여 높은 분류성능을 얻는 방법이 시도되어왔으며, 최적의 것을 찾기 위하여 유전자 알고리즘이 사용되기도 했다. 본 논문에서는 이를 발전시켜 다양한 최적의 앙상블을 생성하기 위해 종분화 방법을 사용한다. 림프종 암 데이터에 대하여 leave-one-out cross-validation을 적용한 결과, 제안한 방법으로 다양한 최적해를 탐색하는 것을 확인할 수 있었다.

1. 서론

암은 인간에게 많은 피해를 주고 있는 무서운 질병이지만, 빨리 발견하여 조기에 적절한 조치를 취한다면 치료가 가능하다. 최근 DNA 마이크로어레이 기술의 발전은 수천 개 유전자의 발현 정보를 한번에 획득할 수 있게 해 주어, 암을 조기에 발견하고 부형태(subtype)를 분류할 수 있는 가능성을 제시해주었다. 지난 수년간 많은 곳에서 이와 관련된 연구가 진행되어 왔으며, 다양한 방법들이 개발되고 적용되어 왔다[1].

마이크로어레이 데이터의 분류에서는 우수한 분류기를 선택하여 분류하는 것 못지않게 관련된 유전자(특징)들을 찾아내는 것이 중요하다. 관련 유전자 자체만을 연구하는 것도 의미가 있고, 또한 이들을 이용하여 분류한다면 계산 시간도 줄이고 분류 성능도 높일 수 있기 때문이다. 특징선택 방법은 분류기와 결합하여 특징-분류기를 형성하며, 특징선택이나 분류기의 종류가 많을수록 다양한 특징-분류기를 얻을 수 있다. 그러나 많은 특징-분류기들이 제시되어 왔지만, 알고리즘 자체의 불완전성과 데이터의 결함, 매개변수 설정의 까다로움 등으로 인하여 그 중 완벽한 것을 찾을 수가 없었다.

분류기의 조합으로 새로운 분류기를 만들어내는 앙상블은 이런 상황에서 높은 분류 성능을 얻기 위하여 사용되어 왔으며, 이전 연구에서도 마이크로어레이 데이터의 분류에 유전자 알고리즘(GA)으로 얻은 앙상블 분류기를 이용하여 높은 분류성능을 얻을 수 있는 가능성을 확인하였다[2]. 그러나 이전연구에서는 일반적인 GA를 사용하여 동시에 다양한 앙상블을 탐색할 수 없었다. 따라서 본 논문에서는 GA에 종분화 방법을 적용시켜 다양한 최적 앙상블을 얻는 방법을 제시하고, leave-one-out cross-validation (LOOCV) 방법을 시행하여 제안한 방법의 일반화된 성능을 보여주고자 한다.

2. DNA 마이크로어레이 데이터

DNA 마이크로어레이는 용액이 투과하지 않는 딱딱한 지지체 위에 고밀도로 cDNA를 고정시켜 놓은 것이다. 어레이를 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[3].

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

그림 1은 유전자 발현 데이터를 얻는 과정으로, 최종 데이터는 행렬로 표현되며, 각 행은 유전자, 열은 샘플을 의미한다.

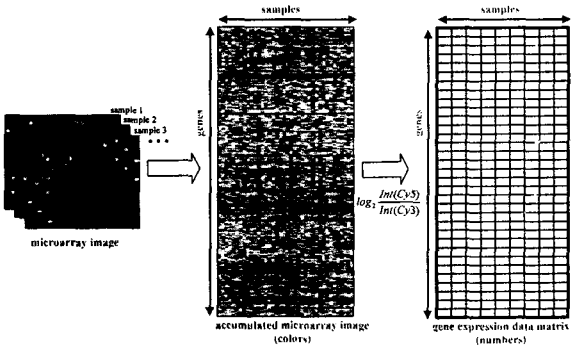

그림 1. 유전자 발현 데이터의 취득

3. 종분화 GA 기반 최적 앙상블 분류기

본 논문에서는 마이크로어레이 데이터를 분류하기 위

한 최적의 앙상블을 찾기 위한 방법으로 그림 2와 같은 방법을 제안한다. 먼저 유전자 발현 데이터로부터 8개의 특징선택과 6개의 분류 과정을 거쳐 총 48개의 특징-분류기를 구성한 후, 종분화 GA를 이용하여 이들의 조합 중 최고의 성능을 보이는 것들을 탐색해나간다. 제안한 방법의 검증을 위하여 매번 다른 테스트 데이터에 적용시키는 LOOCV 방법을 시행한다.

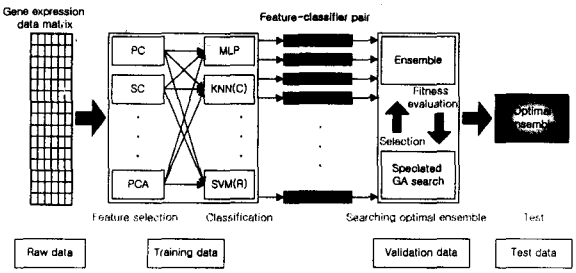

그림 2. 제안하는 방법

3.1 특징선택 및 분류기

본 논문에서는 여러 가지 분야에서 사용되고 있는 총 8가지 특징선택 방법과 6가지 분류기를 사용하였다. 특징선택 방법으로는 상관계수 기반의 피어슨 상관계수(**PC**)와 스피어맨 상관계수(**SC**), 유사도 기반의 유클리드 거리(**ED**)와 코사인 계수(**CC**), 정보이론 기반의 정보이득(**IG**), 상호정보(**MI**), 신호 대 잡음 비(**SN**), 그리고 특징들의 선형조합으로 새로운 특징을 구성하는 주성분분석(**PCA**)를 사용하였고, 분류기로는 다층신경망(**MLP**), 구조적응자기조직화지도(**SASOM**), 선형 커널함수를 사용한 SVM(**SVM(L)**), RBF 커널함수를 사용한 SVM(**SVM(R)**), 코사인 계수를 사용한 k-최근접 이웃(**KNN(C)**), 피어슨 계수를 사용한 k-최근접 이웃(**KNN(P)**)을 사용하였다.

3.2 앙상블 분류기

일반적으로 앙상블의 분류 결과는 참여하는 분류기들에 비하여 높은 정확률을 내는 것으로 알려져 있다. 참여하는 분류기들의 성능이 양호하며 서로 다른 샘플에 대하여 잘못 분류를 하는 경우 더욱 좋은 앙상블 분류기를 얻을 수 있다는 것이 이론적으로나 실험적으로 증명되었다[4]. 따라서 분류기간에 서로 보완적인 것들을 찾아내어 그들을 결합하는 것이 필요하다. 48개의 특징-분류기의 조합을 통하여 얻을 수 있는 앙상블의 수는 2^{48}가지나 되기 때문에, 개별 특징-분류기 자체에 비하여 좋은 분류기가 존재할 가능성이 크다.

다양한 결합방법들이 존재하지만, 본 논문에서는 이진 분류 문제에 간단하게 적용시킬 수 있는 투표결합 방법을 이용하여 앙상블 분류기를 구성하였다. 투표결합은 앙상블에 참여하는 분류기들의 출력의 다수결로 분류결과를 결정하는 방법이다.

3.3 종분화 GA를 이용한 앙상블 탐색

단일해가 존재하고 해공간이 복잡하지 않은 경우 일반적인 GA는 해를 빠르고 정확하게 찾아주는 장점이 있다. 하지만 다중해 문제에서 표준 GA는 쉽게 지역해에 빠지는 단점이 있다. 그러한 문제를 해결하기 위하여 GA에 종분화 방법을 적용시키는 방법이 고안되었다[5]. 마이크로어레이 데이터의 분류를 위한 최적앙상블의 탐색문제도, 최적의 앙상블이 하나만 존재하는 것이 아니므로 다중해 문제라고 볼 수 있기 때문에 표준 GA 보다는 종분화 방법을 적용시켜 동시에 다양한 해를 탐색하도록 하는 것이 더욱 효과적이다. 본 논문에서는 두 가지 대표적인 종분화 기법인 적합도 공유와 결정적 크라우딩 방법을 사용하여 최적의 앙상블을 탐색하였다[5].

적합도 공유는 공유반경 안쪽의 유사한 개체들의 적합도를 낮추어 해가 한쪽으로 쏠리는 현상을 방지해주는 기법이며, 공유 적합도 sf_i는 원래의 적합도 f_i와 근처 개체들의 영향을 나타내는 m_i에 의하여 식 (2)와 같이 표현된다.

$$sf_i = \frac{f_i}{m_i} \tag{2}$$

m_i는 공유함수 sh의 합으로 표현되며, sh는 공유 반경이 σ_s인 경우, 다음과 같이 구할 수 있다.

$$sh(d_{ij}) = \begin{cases} 1-(\frac{d_{ij}}{\sigma_s})^{\alpha}, & \text{for } 0 \le d_{ij} < \sigma_s \\ 0, & \text{for } d_{ij} \ge \sigma_s \end{cases} \tag{3}$$

결정적 크라우딩은 높은 적합도와 개체의 다양성을 동시에 유지시키는 전략이며, 그림 3의 과정으로 진행된다.

```
Input: g - number of generations to run, s - population size
Output: P(g) - the final population

P(0) ← initialize()
for t ← 1 to g do
    P(t) ← shuffle(P(t-1))
    for i ← 0 to s/2 -1 do
        p1 ← a2i+1(t)
        p2 ← a2i+2(t)
        {c1, c2} ← recombination(p1, p2)
        c1' ← mutate(c1)
        c2' ← mutate(c2)
        if[d(p1,c1')+d(p2,c2')] ≤ [d(p1,c2')+d(p2,c1')] then
            if F(c1') > F(p1) then a2i+1(t) ← c1' fi
            if F(c2') > F(p2) then a2i+2(t) ← c2' fi
        else
            if F(c2') > F(p1) then a2i+1(t) ← c2' fi
            if F(c1') > F(p2) then a2i+1(t) ← c1' fi
        fi
    od
od
```

그림 3. 결정적 크라우딩의 의사코드

GA에서 염색체는 48비트로 구성되어 있으며, 각 비트는 대응하는 특징-분류기의 앙상블 참여여부를 나타낸다. 초기 집단은 각각의 종분화 방법에 의하여 원하는 적합도에 도달하거나 일정 세대가 지날 때까지 진화과정을 반복하며 최적의 앙상블을 탐색해 나간다.

4. 실험 및 결과

4.1 실험 환경

본 논문에서는 제안한 방법의 성능을 알아보기 위하여 림프종 데이터(http://genome-www.stanford.edu/lymphoma)를 대상으로 적용시켰다. 이 데이터는 47개의 샘플로 구성

되어있고, 각 샘플은 4026개의 유전자로 이루어져 있다. 선택하는 유전자의 수는 25개로 고정하였으며, 몇 가지 특징선택 방법에서는 특징의 값을 0과 1사이로 정규화시켰다. GA의 경우 100부터 2000사이의 다양한 집단 크기, 0.3부터 0.9사이의 교차율, 0.01과 0.05의 돌연변이율을 사용하였다. 적합도 공유는 유전자형에 대해서 적용하였고, 공유반경의 크기는 5, 상수 α값은 2를 사용하였다.

4.2 실험 결과

표 1은 개별 특징-분류기들의 평균 정확률을 나타낸다. PCA와 IG가 특징선택 방법으로 좋은 성능을 보여주었으며 분류기로는 MLP와 KNN이 좋았다.

표 1. 개별 특징-분류기의 분류결과(단위 : %)

	MLP	SASOM	SVM(L)	SVM(R)	KNN(C)	KNN(P)	평균
PC	77.6	67.6	66.4	55.6	78.4	78.0	70.6
SC	78.8	67.2	68.0	57.6	78.4	76.8	71.1
ED	75.2	62.8	66.4	64.0	76.0	77.6	70.3
CC	80.0	64.4	72.4	56.4	78.0	78.4	71.6
IG	85.2	75.2	77.6	66.8	81.6	83.2	78.3
MI	80.0	67.6	67.2	58.4	76.4	77.2	71.2
SN	81.2	70.8	68.0	58.4	78.8	79.2	72.7
PCA	87.2	84.0	**88.4**	58.4	86.0	86.4	81.7
평균	80.7	70.0	71.8	59.5	79.2	79.7	73.5

한편, 종분화 GA를 이용하여 앙상블을 탐색한 결과 validation set(그림 2)에 대하여 모든 샘플을 정확히 분류하는 앙상블을 찾아주었다. 그림 4는 여러 번 실험에서 각 방법의 분류율을 의미하며, 종분화 GA가 찾은 것의 우수성을 보여준다.

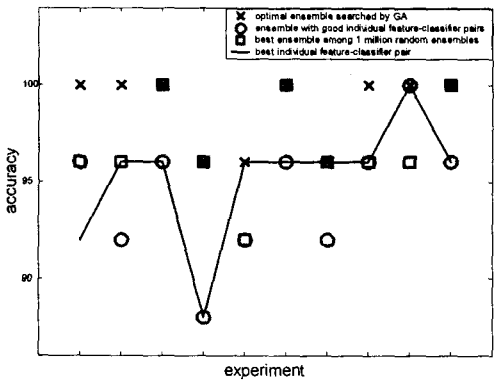

그림 4. 각 앙상블의 성능 비교

표준 GA와 종분화 GA를 비교하기 위하여 동일한 집단 크기로 같은 세대만큼 진화시킨 결과, 표준 GA와 적합도 공유간에는 큰 차이가 나지 않았으나, 결정적 크라우딩은 5배에서 150배정도 다양한 최적의 앙상블들을 찾아주었다. 그림 5는 표준 GA와 적합도 공유, 결정적 크라우딩에 대한 평균 적합도의 추이를 보여준다. 표준 GA는 처음에 증가하다가 약 150세대 이후 수렴하는 모습을 보여 주었고, 적합도 공유는 증가속도는 느리지만 꾸준히 증가하는 모습을 보여주었으며, 결정적 크라우딩의 경우는 초기에 급격하게 상승한 후 소강상태에 접어드는 모습을 볼 수 있었다. 두 가지 종분화 방법에서 평균적합도의 변화는 서로 다른 양상을 보이기 때문에, 해를 탐색하는 전략에 따라 다른 종분화 방법을 적용시킬 수 있다.

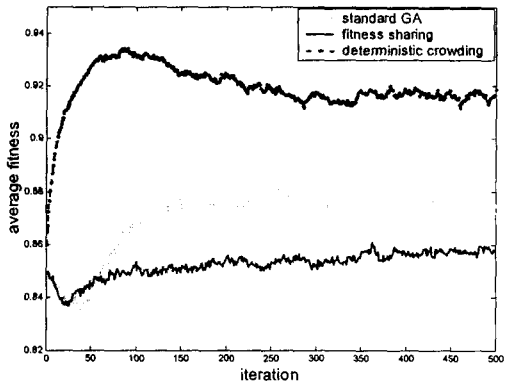

그림 5. 각 방법별 평균적합도 비교

한편, 제안한 방법의 일반성을 보이기 위하여 결정적 크라우딩에 LOOCV 방법을 시행하였다. 47번의 실험에서 각각 다른 테스트 데이터에 대하여 최적의 앙상블을 적용시킨 결과 오직 3개의 샘플만 잘못 분류하였고, 나머지 44개의 샘플에 대해서는 올바르게 분류하여 93.6%의 분류 정확률을 기록하였고, 이는 개별 특징-분류기의 최고성능보다 뛰어난 결과이다.

5. 결론

본 논문에서는 종분화된 GA를 이용하여 DNA 마이크로어레이 데이터의 분류를 위한 특징-분류기들의 최적 앙상블을 탐색하였고, 몇 가지 실험을 통하여 제안한 방법의 우수함을 보였다. 향후에는 이렇게 찾은 최적의 앙상블을 체계적으로 분석할 것이다.

감사의 글

본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임.

참고문헌

[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] 박찬호, 조성배, "유전자 알고리즘을 이용한 림프종암의 최적 분류기 앙상블," 한국정보과학회 춘계학술대회 발표논문집, pp. 356-358, 2003.

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

[4] D. Opitz and R. Maclin, "Popular ensemble metohds: An empirical study," *Journal of Artificial Intelligence Research*, 11. pp. 169-198, 1999.

[5] K. Deb and D. E. Goldberg, "An investigation of niche and species formation in genetic function optimization," *Proc. 3rd Int. Conf. Genetic Algorithms*, pp. 42-50, 1989.