생명정보학에서의 거대규모 특징추출을 위한 종분화 GA의 활용

황금성, 조성배
연세대학교 컴퓨터과학과
yellowg@cs.yonsei.ac.kr, sbcho@cs.yonsei.ac.kr

Applying Speciated GA to Huge-scale Feature Selection in Bioinformatics

Keum-Sung Hwang°, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

최근 생물 유전자 정보에 대한 관심이 커지면서 이를 위한 효과적인 분석 방법이 요구되고 있다. 특히, 분류기의 데이터로 사용하기 위해서 필요한 특징만을 뽑는 과정인 특징 추출은 대량의 유전자 정보에서 의미 있는 정보를 선별하는 중요한 과정이다. 그러나 유전자 정보는 사용되는 데이터의 특징규모가 매우 크기 때문에 일반적인 데이터 마이닝 기법으로는 분석이 힘들다. 본 논문에서는 효율적인 거대규모 특징 추출을 위해 유전자 알고리즘(GA)과 신경망을 사용한 특징추출 방법을 소개하고, 종분화 기법을 사용한 효과적인 특징추출 방법을 제시한다. 그리고, CAMDA 2000에 공개된 암 DNA Microarray로 암종류를 분류하는 문제에 대하여 성능을 평가하였다.

1. 서론

생명 공학과 분석화학의 발달로 생물의 유전자 정보를 대량으로 얻어내는 것이 가능하게 되었다. 그러나 이렇게 얻어진 정보는 단순한 숫자의 나열이므로 이를 분석하여 의미 있는 정보를 뽑아내는 연구와 생물의 방대한 유전자 정보의 분석을 지원해 주는 효과적인 분석도구의 필요성이 대두되고 있다. 이를 위한 연구 중의 하나가 유전자 정보를 통해 질병을 인식하는 분류기를 만드는 연구이다. 이 방법은 먼저 수많은 유전자 정보에서 가장 분류성향이 강한 특징을 뽑아내는 특징추출 과정과 추출된 특징 데이터를 분류기에 넣어 학습시키는 패턴분류 과정을 거친다[1,2].

본 논문에서는 이러한 과정에서 특징 추출을 좀더 효율적으로 하기 위한 방법을 제안한다. 앞선 연구에서는 데이터의 특징별 분류성향만을 고려하였으나, 비선형적인 특성도 고려하기 위해 유전자 알고리즘(GA: genetic algorithm)과 신경망을 이용한 방법을 사용한다. 또한, 이를 좀더 효과적으로 수행하기 위해 종분화 기법을 적용한 GA를 제시한다.

2. 배경

2.1. 유전자 발현정보 분류시스템

DNA Microarray는 용액이 투과하지 않는 고형 지지체 위에 두 개의 다른 환경에서 채집된 DNA를 고밀도로 고정시켜 놓고, 각각 다른 색상의 형광물질을 합성한 것을 똑같은 양만큼 보합하여 서로 다른 두 개의 환경 중 어떤 환경에서 유전물질이 많이 발현했느냐를 색상 정보를 통해 얻어낸 유전자 정보의 집합이다. 최근에는 이러한 과정이 자동화되어 많은 양의 데이터를 수집할 수 있게 되었으며 이를 분석하기 위한 여러 방법들이 연구되고 있다[1].

본 논문에서 사용된 CAMDA(Critical Assessment of Microarray Data Analysis, http://bioinformatics.duke.edu/camda) 2000년도 데이터는 이와 같은 과정을 거쳐 72명의 백혈병 환자의 골수 샘플로부터 얻어진 DNA Microarray 집합이다[2]. 이 데이터는 급성 림프구성 백혈병(ALL: acute lymphoblastic leukemia)과 급성 골수성 백혈병(AML: acute myeloid leukemia)의 두 가지 클래스로 구분되는데, 유전자 발현정보를 통해 환자가 가지고 있는 병이 어느 것인지를 알아내는 것이 분류시스템의 기능이다[3]. 그림 1은 분류시스템의 처리과정을 보여주고 있다.

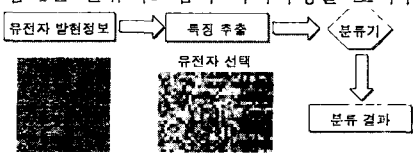

그림 1. 유전자 발현정보 분류시스템

2.2. 종분화 알고리즘

유전자 알고리즘은 다수의 개체를 동시에 진화시켜 가면서 최적의 해를 찾는 효과적인 탐색 알고리즘이다[4]. 여기에 종분화 기법을 적용하면 어떤 문제에 대한 여러 개의 좋은 해를 한 번에 얻을 수 있으며, 진화의 질적인 수준도 한층 더 끌어올릴 수 있다[4,5,7]. 본 논문에서는 종분화 기법으로 적합도 공유(Fitness Sharing) 방법을 사용하였다[6]. 이때 공유를 적용한 적합도 sf_i는 다음과 같이 계산된다.

$$sf_i = \frac{f_i}{\sum_{j=1}^{n} sh(d_{ij})}$$

여기서 f_i는 개체에 대한 적합도를 의미하며, n은 개체수를 의미한다. 그리고 $sh(d_{ij})$는 각 개체간 공유 상태를 나타내는데, 개체 i와 j의 거리(차이)를 나타내는 d_{ij}를 이용해서 각 개체가 공유거리 내에 들어와 있는 정도를 아래와 같이 합산한 값이다. 이때 공유거리 σ_s가 사용되는데 이는 개체간 공유를 적용하는 범위를 정하는 파라미터이다.

$$sh(d_{ij}) = \begin{cases} 1 - \dfrac{d_{ij}}{\sigma_s}, & for\ 0 \le d_{ij} < \sigma_s \\ 0, & for\ d_{ij} \ge \sigma_s \end{cases}$$

개체간 거리는 유전자형(Genotype) 거리의 하나인 해밍 거리(Hamming distance)를 이용하여 계산하였다[7].

2.3. 다층 신경망

신경망(neural network)은 패턴분류 분야에서 가장 활발하게 사용되는 분류기법의 하나이다. 그리고, 역전파(BP: backpropagation) 신경망은 그 중에서도 가장 잘 알려진 신경망의 일종으로, 학습 데이터를 입력하였을 때 계산되는 오류를 역으로 수정하는 과정을 반복적으로 수행하여 신경망을 학습한다. 그리고 다층 신경망은 하나의 작은 분류기와 같은 역할을 하는 퍼셉트론(perceptron)이 여러 층으로 결합되어 있어서, 선형적인 분류 뿐만 아니라 비선형적인 분류 문제도 해결할 수 있다[8]. 본 논문에서는 비선형적인 분류성향을 평가하기 위해 다층 신경망을 사용하였다.

3. 특징 추출

3.1. 기존 연구

생물 정보에 대한 기존의 특징 추출 연구에서 사용된 방법은 각각의 특징에 대해서 여러 가지 거리 측정법(피어슨 상관관계, 유클리드 거리, 신호 대 잡음비 등)을 사용하여 그룹별 거리 분포가 가장 크게 나타나는 특징을 원하는 수만큼 선택하는 것이었다[1]. 이 방법은 유전 정보 데이터와 같이 특징 규모가 큰 경우에 효율적으로 수행할 수 있다는 장점은 있으나, 각 특징별로 1차원적 평가를 내리기 때문에 특징의 조합에 의해 생기는 비선형적 분류 성향은 무시하게 된다. 따라서, 특징 조합에 의한 중요한 분류 정보를 놓칠 가능성이 있다. 그림 2의 비선형 분류 문제를 보면 선형 분류의 한계점을 알 수 있다.

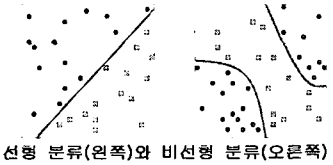

그림 2. 선형 분류(왼쪽)와 비선형 분류(오른쪽) 문제의 예

데이터 마이닝 분야에서 진행되고 있는 특징 추출에 관한 연구를 살펴보면 대부분 중간 규모(20~60개)와 큰 규모(60~수백개)의 특징 추출 연구에 제한되어 있다. 거대규모(1,000개 이상) 특징 추출에 관한 연구도 일부 있으나 특징 추출 전의 필터링 기법만을 소개할 뿐 특별한 대안은 없는 실정이다[9]. 본 논문에서 사용한 데이터의 경우 7,129개의 특징을 가지고 있는데 기존의 특징 추출 방법으로는 적용이 힘들다. 따라서 거대규모 특징 추출에 대한 새로운 방법이 요구된다.

3.2. 제안하는 방법

본 논문에서는 특징 조합에 의한 비선형적 분류 성향을 고려하기 위하여 다층 신경망을 이용한 특징 평가를 사용하였다. 다층 신경망은 비선형적인 문제에서 좋은 성능을 나타낸다.

이때 신경망을 통한 평가를 할 특징 조합을 선택하기 위해 GA를 사용하였다. 즉, 선택된 특징 조합을 GA 염색체를 통해 얻어서 이를 신경망으로 평가하여 적합도값으로 사용하였다. 특징 조합의 수는 아래 수식과 같이 계산되는데, 상당히 많은 경우의 수를 가지고 있기 때문에 모든 경우를 평가해보는 것은 어렵다. 따라서 최적화 문제에 적합한 GA를 사용한다. (n_f = 전체 특징의 수)

$$\text{특징 집합의 수} = \sum_{k=1}^{n_f} {}_{n_f}C_k = 2^{n_f}$$

GA를 이용한 특징 추출은 이미 Caballero 등에 의해 수행된 바 있지만[10], 중간 규모의 특징 추출에 대해서만 가능하다는 제약이 있다. Caballero 등이 사용한 염색체 표현법은 그림 3과 같이 특징의 수만큼 비트를 주고 비트값을 통해 사용여부를 결정하는 구조이다. 이런 구조는 특징의 수가 작을 경우에만 유효하다.

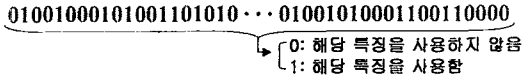

그림 3. Caballero가 사용한 염색체 구조

본 논문에서는 거대규모 특징의 GA표현에 적합하도록 그림 4와 같은 염색체 구조를 정의하여 염색체 크기를 줄였다. 전체 특징 중 선택된 특징을 가리키는 인덱스 f_k만을 포함하고 있는 염색체 구조이다. 이 구조는 전체 특징의 수에 큰 영향을 받지 않으며 상대적으로 상당히 짧은 염색체가 가능하다. 염색체 길이는 전체특징의 수 n_f와 선택 특징의 수 n_s의 크기에 따라 가변적인데, 본 논문에서는 n_s값을 25로 주어 25개의 특징 집합을 선택하였으며, 전체 특징의 크기 7,129를 표현하기 위해 13비트의 인덱스가 사용되었다.

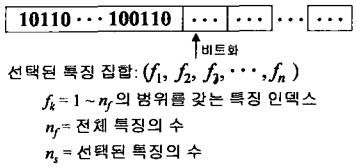

그림 4. 거대규모 특징 추출을 위한 염색체 구조

제시한 염색체 구조를 사용하면 기존 방법의 경우 7,129비트의 길이가 요구되는데 비해 325비트의 작은 길이만을 요구하여 효율적인 GA가 가능하다. 그러나, n_f와 n_s의 크기가 큰 경우 염색체의 길이가 길어질 수 있다. 따라서, 본 논문에서는 그림 5와 같이 종분화 기법을 적용하여 효과적으로 염색체를 표현하였다.

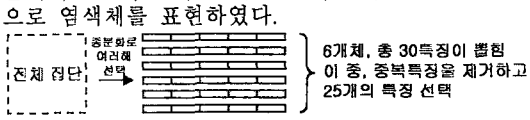

그림 5. 종분화된 개체를 통한 특징 수집

종분화 기법을 적용하면 한 번의 진화로 여러 해를 동시에 얻는다. 이 특성을 염색체에 적용하면 n_s개의 특징을 한 개의 염색체를 통해서가 아닌, 여러 개의 해를 통해 얻을 수 있다. 본 논문의 경우 5개의 특징을 가진 염색체 6개에서 중복 제거 후 총 25개의 특징을 얻도록 하였는데, 이 경우 염색체는 65비트의 길이만이 요구된다.

4. 실험 및 결과

CAMDA 데이터 72개의 샘플 중 38개는 신경망의 학습 데이터이고, 다른 34개는 분류기의 성능을 평가하기 위한 데이터이다. 학습 데이터 중 27개는 ALL 환자의 것이고, 11개는 AML 환자로부터 얻어진 것이다. 신경망을 학습시

키기 위해 ALL, AML 샘플을 각각 클래스 0, 1로 설정하였고, 실험의 입력값은 특징별로 0에서 1사이의 값을 가지도록 정규화하여 사용하였다.

4.1. 실험 환경

종분화 방법은 앞서 소개한 적합도 공유 기법을 사용하였으며 기본 환경 변수는 표 1과 같이 설정하였다. (단, 분류기는 별도로 튜닝하였다.) 유전연산자는 1점 교차와 룰렛휠 선택, 비트플립 돌연변이를 사용했고, 엘리트 유지 전략을 사용해서 최고해를 유지시켰다. BP 신경망은 오류를 줄이기 위해서 2번 학습한 뒤 좋은 값을 선택하도록 하였으며, 각 실험은 10번씩 수행한 뒤 결과를 평균하였다.

표 1. 실험 환경 변수

유전연산 변수	값	신경망 변수	값
집단의 크기	50	학습율	0.3
선택율	0.8	모멘텀	0.5
교차율	0.8	최대 반복수	1,500
돌연변이율	0.001	최소 오류	0.02

4.2. 결과

먼저 특징의 다차원적인 조합을 고려하였을 때 얻을 수 있는 성능향상을 알아보기 위하여 신경망의 학습 오류를 관찰해 보았다. 그림 6은 특징조합의 수를 1에서 4까지 주었을 때 진화에 따른 오류값의 변화를 나타낸 것이다. 특징 조합이 많이 허용될수록 오류값이 더 작아지는 경향을 보이고 있다. 따라서 여러 특징 조합의 비선형적 분류성향 고려는 유효한 선택이었음을 알 수 있다.

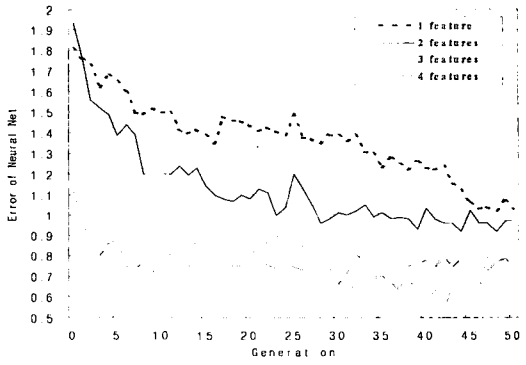

그림 6. 특징 조합의 수에 따른 신경망 오류 변화 (10회 평균)

이제 제안하는 방법의 효율성을 평가하기 위하여 특징 추출과 최종 분류 성능을 분석하여 본다. 표 2는 제안한 GA와 종분화를 적용한 특징추출 실험 결과를 보여주고 있다. GA의 경우 제안한 염색체 구조에 25개의 특징을 한 번에 선택하도록 설정하였고, 종분화의 경우에는 5개의 특징을 가진 염색체를 6개 선택하여 중복 특징을 빼고 25개의 특징을 선택하였다. 이때, 6개의 해를 찾기 위해서 학습 오류가 1.2이하인 개체가 6개에 도달할 때까지 진화를 수행하였다. 그리고 GA는 비교실험하기 위해 종분화와 비슷한 세대까지 수행하였다.

표 2. 특징 추출 결과 비교 (10회 평균)

과정	측정치	GA	종분화
특징 추출	학습오류 평균값	0.00080	0.71707
	평균 수행시간	157.4초	45.1초
	평균 수행 세대	5.0세대	4.4세대

실험 결과를 보면 예상대로 종분화의 경우 빠른 수행시간을 보이고 있다. 염색체 길이의 차이로 인한 결과이다. 그리고 학습 오류는 GA를 사용한 경우 훨씬 작게 나오는데, 그림 6에서 알 수 있듯이 많은 특징 조합을 허용했기 때문에 나온 결과이다. 표 3은 위에서 추출된 특징 집합을 이용하여 분류기를 구성한 결과이다.

표 3. 분류기 학습 및 평가 결과 비교 (10회 평균)

과정	측정치	GA	종분화
분류기	학습 오류 평균	0.00019	0.02403
	분류성공률	0.72353	0.80000
	평가 오류	4.28492	2.85500

실험 결과, 두 방법 모두 학습 데이터에 대해 낮은 학습 오류를 나타내었다. 그리고 분류 성공률은 오히려 종분화가 더 좋게 나왔다. 진화 시간의 단축을 위해 사용한 종분화 기법이 더불어 성능도 좋게 한 것이다. 종분화를 적용한 진화시 다양성이 잘 유지되기 때문에 전체 해공간의 다양한 특성이 고려된 특징 집합이 선택될 수 있었고, 이것이 분류기의 성능을 향상시킨 것으로 보인다.

5. 결론 및 향후 연구

본 논문에서는 기존의 거대규모 특징 추출 기법의 문제점을 알아보고 새로운 방법을 제시하였다. 특징 조합의 특성을 반영하기 위해서, 신경망으로 평가되는 GA를 사용하였으며, 추출된 특징이 학습 데이터에 대해 아주 낮은 오류값이 나타남을 확인하였다. 또한, 효율적인 진화가 수행되도록 염색체 크기를 크게 줄인 구조를 사용하였다. 그리고 종분화 기법을 이용한 다중 특징 추출을 통해 상당한 시간 단축 효과도 얻었다.

그러나 실험 결과 나온 분류성공률은 기존의 결과값에 미치지는 못했다[1]. 이는 학습데이터의 오류 감소가 평가 데이터에 대한 성능 향상을 항상 보장하지는 않기 때문에 생기는 현상이다. 학습데이터가 충분하지 않을 경우 신경망은 과적응(overfitting) 될 수 있다. 따라서 제시하는 방법에 대해서 향후 더 큰 데이터 집합을 이용한 추가적 검증이 필요하다.

6. 참고 문헌

[1] 권영준, 류중원, 조성배, "신경망 분류기를 이용한 암 관련 유전자 발현정보의 분류," *한국정보과학회 춘계 학술발표논문집(B)*, 28권, 1호, pp. 229-232, 2001.

[2] Simon M. Lin et al., *Methods of Microarray Data Analysis: Papers from CAMDA'00*, Kluwer Academic Publishers, 2001.

[3] T. R. Golub *et al.*, "Molecular classification of cancer: Class discovery and class prediction by gene-expression monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[4] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley, 1989.

[5] 황금성, 조성배, "종분화를 이용한 다품종 하드웨어의 진화," *한국정보과학회 춘계 학술발표 논문집(B)*, 28권, 1호, pp. 229-232, 2001.

[6] P. Darwen and X. Yao, "Every niching method has its niche: Fitness sharing and implicit sharing compared," *Proc. of Parallel Problem Solving from Nature (PPSN) IV*, Vol. 1141, Lecture Notes in Computer Science, Springer-Verlag, pp. 398-407, Berlin, 1996.

[7] S. W. Mahfoud, "Niching methods," *Evolutionary Computation 2: Advanced Algorithms and Operators*, Institute of Physics Publishing, pp. 87-92, 2000.

[8] S. Haykin, *Neural Networks*, Prentice Hall, 1999.

[9] J. Bins and B. Draper. "Feature selection from huge feature sets," *International Conference on Computer Vision*, Vol. 2, pp. 159-165, Vancouver, July 9-12, 2001.

[10] R. E. Caballero and P. A. Estevez, "Feature selection using a niching genetic algorithm: An experimental study," *VIII Latin American Congress of Automatic Control*, 1998.