다양한 종분화 진화 신경망을 결합한 대장암 분류

김경중,O 조성배
연세대학교 컴퓨터과학과
uribyul@sclab.yonsei.ac.kr sbcho@cs.yonsei.ac.kr

Classifying Colon Cancer by Integrating Diverse Speciated Evolutionary Neural Networks

Kyung-Joong Kim, Sung-Bae Cho
Dept. of Computer Science, Yonsei University

요 약

암의 발병을 조기에 예측하고 진단하는 것은 매우 중요하지만 그 과정이 매우 복잡하고 많은 노력이 필요하다. 암이 발생하는 원인은 매우 다양하지만 근본적으로 단백질을 형성하는 유전자에 변화가 오기 때문으로 생각해 볼 수 있다. 유전자 발현 정보로부터 기계적으로 암을 예측하기 위한 과정은 중요한 유전자의 선택, 모델의 학습, 모델을 이용한 예측과정으로 나뉘어 진다. 본 논문에서는 대장암 여부를 유전자 발현 데이터로부터 예측하기 위한 종분화 진화 신경망을 제안한다. 종분화 진화 신경망은 진화 알고리즘을 사용하여 신경망의 구조를 결정하고 종분화 알고리즘을 사용하여 다양한 개체의 생성을 유도한 후 모델의 앙상블을 통해 보다 높은 성능을 내는 방법이다. 실험 결과 제안하는 방법이 대장암 예측 cross validation 테스트에서 96.5%의 높은 성능을 보였다.

1. 서론

다양성은 앙상블 분류기를 설계하는데 매우 중요한 요소이다. 하나의 분류기만을 사용하는 것보다 다양한 특성을 지닌 여러 개의 분류기를 결합하는 것이 더 좋은 성능을 낸다는 것은 널리 알려져 있다. 하지만 어떻게 다양한 분류기를 생성할 것인가에 대한 방법은 다양한 측면에서 연구가 진행되어 오고 있다.

진화 신경망은 한 번에 여러 개의 해를 동시에 탐색하는 진화 알고리즘의 장점과 일반화 능력이 우수한 신경망의 장점을 결합하여 최근 활발히 연구되고 있다. 하지만 진화 알고리즘이 가지고 있는 genetic drift 현상으로 인해 다양한 해를 탐색하지 못하고 하나의 우수한 해만을 탐색하는 특성을 가지고 있다. 만약 집단에 속한 모든 개체를 이용할 수 있다면 진화 신경망의 유용성은 더욱 커질 것이다.

본 논문에서는 진화 알고리즘과 종분화 알고리즘을 결합하여 개체의 다양성을 높이는 방법을 사용한다. 적합도 공유는 개체 사이의 거리를 측정하여 공유반경 안에 존재하는 개체들과 적합도를 공유하여 밀집도가 높은 지역의 개체 밀도를 줄이고 상대적으로 밀집도가 낮은 지역의 개체를 살려 다양성을 높이는 방법이다. 적합도 공유를 이용하여 진화 신경망의 다양성을 높이기 위해서는 개체간 거리 측정을 위한 방법이 필요하다. 두 신경망 사이의 구조적인 거리를 측정하는 것은 어려움이 따르기 때문에 출력 값 사이의 유사도를 이용한다. 또한 종분화를 통해 생성된 다양한 개체들의 결과를 결합하기 위해 측정치 레벨의 결합, 추상 레벨의 결합, 순위 레벨의 결합을 이용한다.

실제 해 공간에는 가장 우수한 개체도 존재하지만 상대적으로 성능은 떨어지지만 다른 개체와 구별되는 좋은 특성을 지닌 개체가 많이 있다. 가장 우수한 개체 하나를 이용하는 것보다는 상호보완적인 여러 개의 개체를 결합하여 이용하는 것이 더 좋은 성능을 보인다. 이를 위해서는 탐색과정에 다양성이 유지되어야 하며 본 논문에서는 종분화 진화 방식을 이용한다.

제안하는 종분화 진화 신경망은 기존의 연구에서 높은 성능을 제시했으며[1] 구조결정을 위한 전문지식을 필요로 하지 않아 생물정보학 문제를 해결하는데 적합하다. 종분화 신경망을 이용하여 대장암 DNA 마이크로어레이 데이터를 분석하고 암의 발병 여부를 예측한다. 표 1은 대장암 분석과 관련된 기존의 연구를 보여준다.

표 1. 대장암 관련 연구

저자	방법		정확도 (%)
	특징 추출	분류기	
Furey *et al.*	Signal to noise ratio	SVM	90.3
Li *et al.*	Genetic algorithm	KNN	94.1
Ben-Dor *et al.*	All genes, TNoM score	Nearest neighbor	80.6
		SVM with quadratic kernel	74.2
		AdaBoost	72.6
Nguyen *et al.*	Principal component analysis	Logistic discriminant	87.1
		Quadratic discriminant	87.1
	Partial least square	Logistic discriminant	93.5
		Quadratic discriminant	91.9

2. 종분화 진화 신경망

마이크로어레이 데이터는 일반적으로 샘플의 개수가 적은 반면에 데이터의 차원이 높기 때문에 적절한 특징 추출 방법을 우선적으로 사용해야 한다. 샘플의 수가 적다는 점 때문에 10-fold cross validation을 이용하여 실험을 수행하였다. 진화 신경망은 검증 데이터를 이용하여 적합도를 평가하고 적합도 공유를 통해 종분화를 수행한다. 마지막 세대의 진화 신경망들을 클러스터링하여 종별로 분할하고 각 종의 대표 개체(가장 성능이 우수한 개체)를 선택하여 결합한 후 분류기로 사용한다.

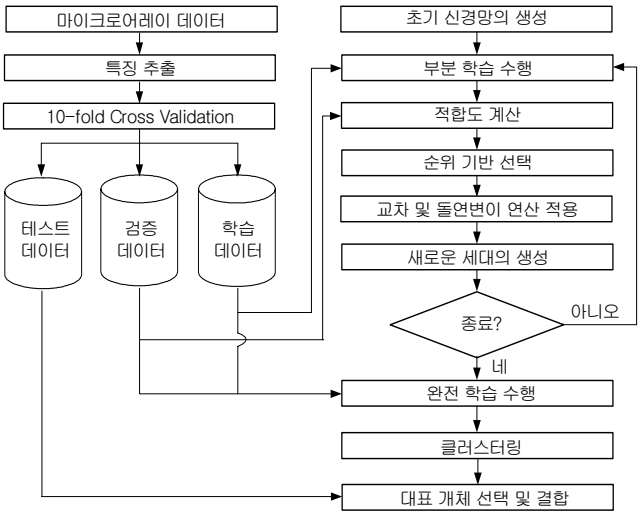

그림 1. 마이크로어레이 분석을 위한 종분화 진화 신경망

2.1 종분화

종분화는 집단의 다양성을 유지하기 위한 방법으로 적합도 공유, crowding algorithm, implicit sharing, clearing 등이 있다. 적합도 공유는 다양성을 유지하기 위해 적합도를 변형시켜주는 방법으로 개체 사이의 거리가 중요한 역할을 한다. 즉 가까운 거리에 있는 개체들은 공유를 많이 하도록 하여 밀집된 개체들이 상대적으로 선택확률이 낮아지도록 하는 방법이다. 신경망 i의 현재 적합도를 f_i라고 하면 공유 적합도 f_i은 현재의 적합도를 공유함수 값의 합인 s_i로 나누면 된다. M은 집단의 크기이고, σ는 공유반경이다. 적합도의 공유는 공유반경 안에 존재하는 개체들 사이에서만 이루어진다. d_{ij}는 신경망 i와 신경망 j사이의 거리를 나타낸다.

$$s_i = \sum_{j=1}^{M} sh(d_{ij}) \qquad sh(d_{ij}) = 1 - \frac{d_{ij}}{\sigma}$$

아직까지 두 신경망을 비교하는데 최적인 기준이 없기 때문에 각 신경망 출력값의 평균으로 유사도를 계산했다. 입력 데이터 t에 대한 신경망 i의 출력값 o_i^t는 다음과 같이 정의할 수 있다.

$$o_i^t = (o_{i1}^t, o_{i2}^t, \Lambda\ , o_{iN}^t)$$

이 때 N은 신경망 i의 출력 노드 수이다. 학습 데이터가 총 T개 있다면 학습 데이터에 대한 신경망 i의 p번째 출력 노드의 출력값 평균 $\overrightarrow{o_{ip}}$는 모든 데이터에 대한 p번째 출력 노드의 출력값을 합한 후 T로 나누면 된다. 신경망 i와 j간의 유사도는 각 신경망의 출력값 평균의 유클리드 거리로 측정한다.

신경망의 출력은 단순히 0과 1사이의 임의의 값이 아니라, 분류기의 베이시언 사후 확률값이라 할 수 있다. 이 성질을 이용하면 수정된 Kullback-Leibler 엔트로피값을 사용해서 두 신경의 차이를 측정할 수 있다. 이 엔트로피값은 두 분포 p와 q 사이의 거리를 다음과 같이 정의한다.

$$D(p,q) = \sum_{i=1}^{N} p_i \log \frac{p_i}{q_i}$$

여기서 N은 신경망의 출력 노드의 수를 나타낸다.
하지만, 이 엔트로피값은 대칭성이 없기 때문에 ($D(p, q) \neq D(q, p)$) 실제 거리를 나타내지 못한다. 이 문제를 해결하기 위해 대칭적인 엔트로피를 다음과 같이 정의할 수 있다.

$$D(p,q) = \frac{1}{2} \sum_{i=1}^{N} (p_i \log \frac{p_i}{q_i} + q_i \log \frac{q_i}{p_i})$$

p와 q를 두 신경망의 출력값 확률 분포라고 하고, 각 신경망이 N개의 출력 노드를 가지고, T개의 데이터로 학습된다고 할 때, 유사도는 다음과 같이 측정될 수 있다.

$$D(p,q) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{T} (p_{ij} \log \frac{p_{ij}}{q_{ij}} + q_{ij} \log \frac{q_{ij}}{p_{ij}})$$

이 때 p_{ij}는 j번째 학습 데이터에 대한 신경망의 i번째 출력값을 뜻한다. 두 신경망이 비슷할수록 엔트로피 값은 더 작아지게 된다.

2.2 결합

평균은 비교적 간단한 결합 방법으로 각 클래스마다 주어진 신경망의 출력값을 모두 합하여 평균을 구해서 값이 가장 큰 클래스를 다중 신경망의 결과로 사용하는 방법이다. 가중 평균은 평균 방법을 보안한 것으로, 각 신경망의 인식률 값을 이용해서 각 신경망마다 가중치를 부여한 후, 출력값에 가중치를 곱해서 평균을 구한다. 신경망의 인식률이 높을수록 더 높은 가중치를 부여하기 위해 신경망의 오류율을 이용해서 가중치를 결정하였다. 신경망 i의 오류율이 E_i 일 때 가중치 w_i는 다음과 같이 구한다.

$$w_i = \frac{1 - E_i}{\sum_k (1 - E_k)}$$

3. 실험 및 결과

본격적으로 대장암을 분류하는 실험을 수행하기 전에 기존 연구에서 널리 사용되던[1] 호주 신용카드 데이터에 대한 종분화 신경망의 특성을 분석해보았으며 이를 대장암 분류문제에 적용해 보았다.

3.1 종분화 신경망 특성 분석

총 1000세대 동안 진화를 수행해 본 결과, 종분화 결합의 성능은 800세대쯤에서 낮아지는 것을 볼 수 있다. 실제적으로 좋은 성능을 얻기 위해서는 800세대까지 가야 하나 많은 시간을 필요로 하기 때문에 200세대에서 멈추는 것으로도 만족스러운 성능을 얻을 수 있었다.

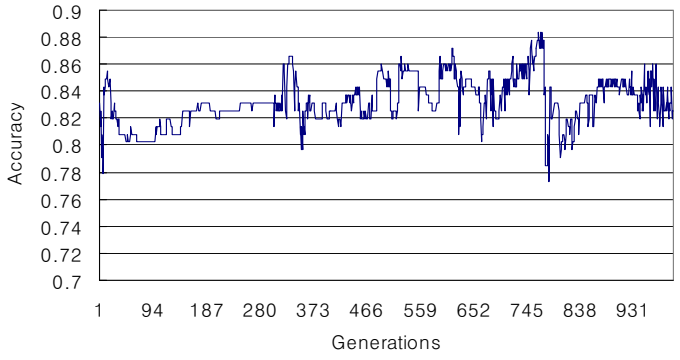

그림 2. 1000세대 동안의 성능 변화

그림 3은 교차율의 변화와 성능 사이의 관계를 보여준다. 신경망 사이의 거리를 측정하기 위해 엔트로피 거리 기준을 사용하였다. 이 경우 돌연변이율을 0.1로 고정하였으며 교차율이 0.3인 경우에 가장 좋은 성능을 얻었다.

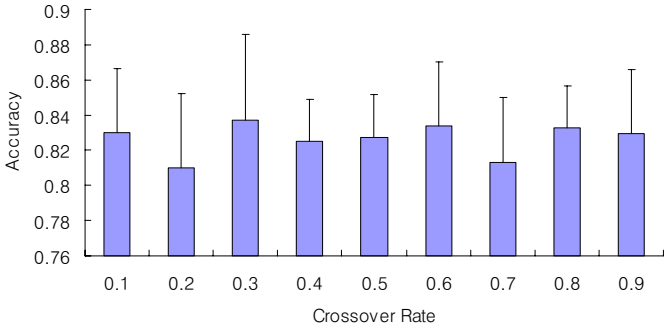

그림 3. 교차율의 변화 (엔트로피 거리기준, 클러스터=5)

그림 4는 클러스터의 개수를 몇 개로 하는 것이 좋은가에 대한 실험이다. 클러스터의 개수를 변화시켜 가면서 성능을 평가하였고 거리기준으로 Pearson Correlation을 사용하였다. 클러스터의 개수가 많아짐에 따라 성능의 향상이 있기는 하였으나 그 정도는 1%를 넘지 않았다.

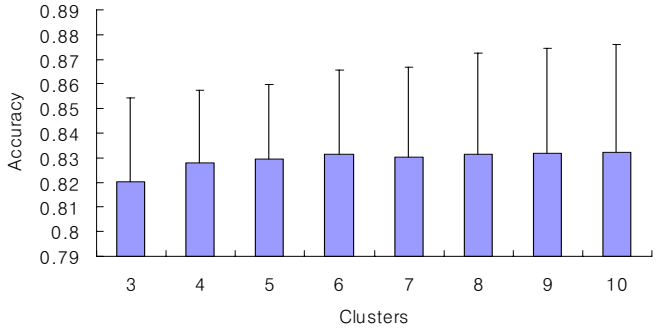

그림 4. 클러스터 개수의 변화 (Pearson Correlation 거리 기준)

그림 5은 세대의 변화에 따른 다양성을 분석한 것이다. 다양성을 평가하기 위해 각 세대마다 모든 개체들 사이의 거리를 엔트로피 방법으로 평가한 후 평균을 구하여 사용하였다. 종분화 하지 않은 것에 비해 종분화 한 것이 높은 다양성을 지속적으로 유지하는 것을 볼 수 있다.

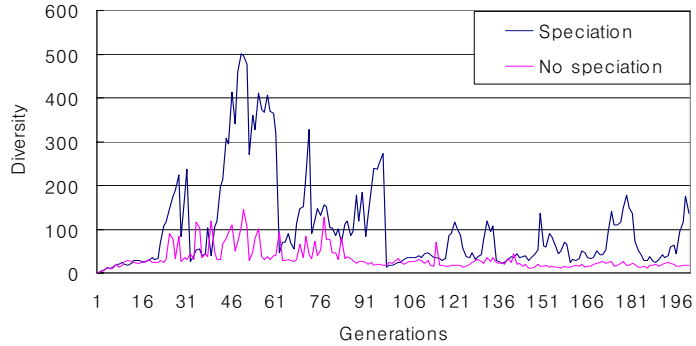

그림 5. 다양성의 분석

3.2 대장암 데이터의 분류

대장암 데이터는 62개의 샘플로 이루어져 있으며 각 샘플은 2000개의 유전자에 대한 발현정도를 가지고 있다. 62개의 샘플 중에서 40개는 대장암 세포이고 나머지는 정상 세포이다[2]. 차원의 감소를 위해 정보이득을 사용했으며 30개의 중요한 유전자를 진화 신경망의 입력으로 사용하였다. 62개의 샘플은 10개의 집단으로 나누어 각각 8개는 학습용, 1개는 검증용, 1개는 테스트용으로 사용하였으며, 10번의 반복실험을 하였다.

그림 6은 정보이득 방법으로 가장 분별력이 높은 30개의 유전자를 선택하고 이들을 계층적 클러스터링한 후 시각적으로 보여준 결과이다.

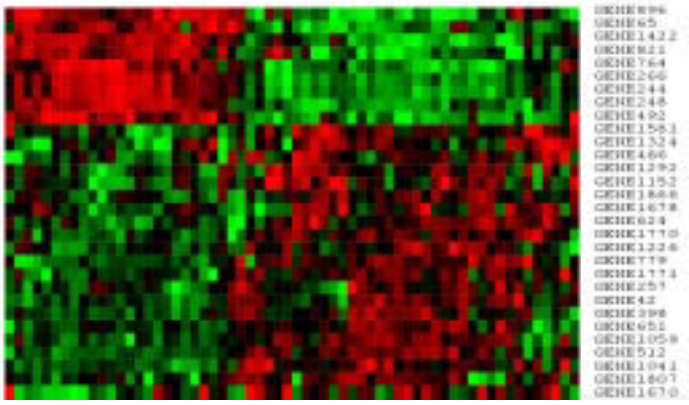

그림 6. 대장암 데이터의 시각화 결과

대장암 분류 문제를 위한 진화 신경망의 파라미터는 돌연변이율을 0.1, 교차율을 0.3으로 하고, 결합에 사용한 개체의 수를 3으로 하였다. 진화는 최대 200세대까지 반복한 후 멈추었다. 신경망의 입력 노드 수는 특징의 수에 맞게 30이며, 출력 노드의 수는 2이며 암인지 아닌지를 출력한다. 은닉 노드의 수는 최대 5개로 설정하였다. 신경망의 부분학습과 완전학습을 위해 BP를 사용하였으며 각각 데이터를 200번과 1000번씩 보여주었다. BP의 학습율은 0.1로 설정하였다.

실험결과 비교적 단순한 방법인 출력값 평균 거리 측정과 가중치 평균 결합 방법의 조합이 96.5%로 가장 높은 성능을 보였다. 데이터의 수가 작기 때문에 복잡한 방법보다는 단순한 결합 방법과 거리 측정 방법이 좋은 성능을 보인 것으로 생각된다(그림 7).

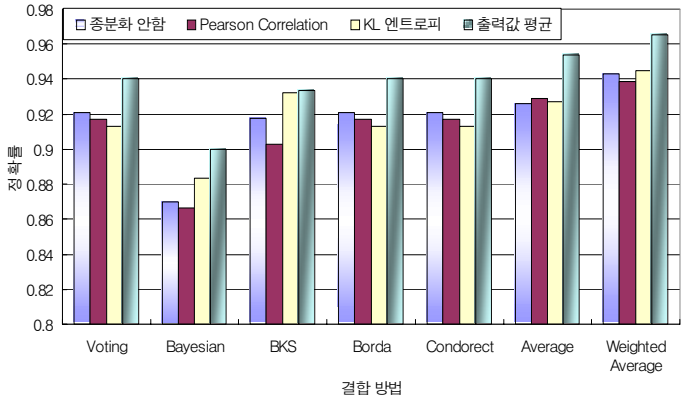

그림 7. 10-fold cross validation 평균

감사의 글

이 연구는 과학기술부가 지원하는 뇌과학연구 프로그램에 의하여 지원 받은 것임.

참고문헌

[1] S.-I. Lee, J.-H. Ahn, and S.-B. Cho, "Exploiting diversity of neural ensembles with speciated evolution," *Proc. Int. Joint Conf. on Neural Networks*, vol. 2, pp. 808-813, 2001.

[2] U. Alon, *et al.*, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Natl. Acad. Sci. USA*, vol. 96, pp. 6745-6750, June 1999.