신경회로망을 이용한 단백질 구조 예측

Protein Disorder Prediction Using Neural Networks

오 상 훈
목원대학교

Sang-Hoon Oh
Mokwon University

요약

단백질의 구조가 무질서한 것을 예측하는 문제는 단백질 시퀀스 구조의 비교 시간을 단축할 수 있으며 단백질 구조 분석 영역을 표시할 수 있기 때문에 중요하게 다루어진다. 이 논문에서는 단백질의 무질서한 구조 예측을 신경회로망을 이용하여 해결하고자 하였으며, 시뮬레이션 결과 일반적인 신경회로망 보다 심층신경회로망이 더 좋은 성능을 보임을 확인하였다.

Ⅰ. 서론

단백질은 생명체에 필수적인 기능을 수행하며 그 기능은 단백질의 3차원 구조에 의해서 결정된다. 이 경우 단백질이 ordered 인지 disordered 인지를 판별하는 것도 단백질 구조 판별에서 중요한 부분을 차지한다[1][2]. 이 논문에서는 신경회로망을 이용하여 이를 판별하는 문제를 학습시키고자 한다. 즉, 일반적인 구조의 신경회로망과 심층 신경회로망을 사용하여 protein disorder prediction을 학습시킨다.

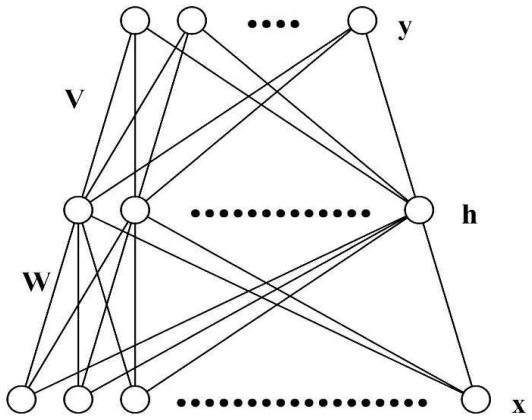
특히 protein disorder prediction은 클래스가 심하게 불균형을 이루고 있기 때문에 신경회로망의 학습에서 data imbalance를 염두에 둔 학습을 사용한다[3].

Ⅱ. 본론

1. 신경회로망의 오류 역전파 학습과 데이터 불균형 문제

신경회로망의 여러 형태 가운데 이 논문에서는 그림 1과 같은 구조의 전방향 신경회로망(Feed-Forward Neural Network)을 사용하였다. 여기서 입력벡터는 x, 중간층 노드 벡터는 h, 출력층 노드 벡터는 y, 중간층 가중치 행렬은 W, 그리고 출력층 가중치 행렬은 V로 표기하였다.

이를 소위 "다층 퍼셉트론(Multi-Layer Perceptron)"이라고도 하며, 중간층 노드(hidden node) 의 개수가 충분하면 임의의 함수를 근사화 시킬 수 있다는 수학적 증명을 토대로 전방향 신경회로망이 많은 분야에 응용되고 있다.



▶▶ 그림 1. 일반적인 전방향 신경회로망 구조

전방향 신경회로망의 학습은 다음과 같이 이루어진다. 먼저, 신경회로망이 지닌 가중치 값들(W, V)을 임의의 값으로 초기화 시킨다. 임의의 학습패턴이 입력되면 Wx에 의해 중간층 노드에 입력되는 가중치 합 $\hat{h}$이 결정되고, 이를 tanh(.)를 통과시켜 h가 결정된다. 마찬가지로 Vh로 출력층 노드에 입력되는 가중치 합 $\hat{y}$이 결정되면 이를 tanh(.)를 통과시켜 출력층 노드 y가 얻어진다. 이 과정을 전방향 계산이라 한다.

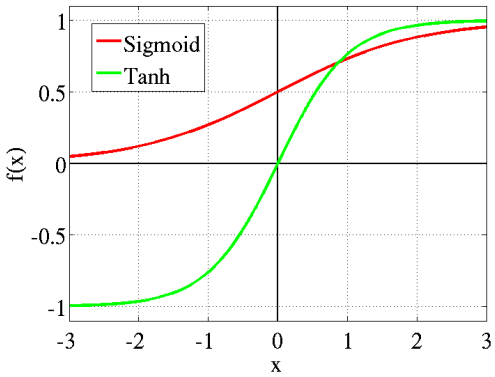
학습패턴 x의 목표벡터 t와 출력벡터 y의 오차를 측정하는 기준으로 보통 MSE(Mean-Squared Error) 함수를 사용한다. 신경회로망의 학습은 오차를 감소시키기 위하여 gradient descent 형식으로 W와 V를 변경하는 형태로 이루어진다. 즉, 오차함수를 출력층 가중치 V에 대하여 미분한 값의 반대 방향으로 V를 변경시키며, W도 마찬가지로 변경시킨다. 이때 출력층의 오류가 중간층으로 전파되는 과정이 필요하다. 이렇게 W와 V를 학습하는 과정을 오류 역전파 알고리즘이라 한다.

한편, 이 논문에서 학습시키고자 하는 protein

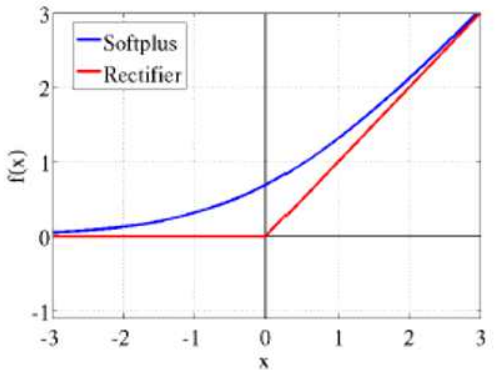
disorder prediction은 ordered class의 데이터 수가 disordered class 데이터 수 보다 아주 많은 경우이므로, 학습 시 data imbalance를 고려한 학습 방법을 사용하여야 한다. 따라서 신경회로망의 학습방법은 위에서 설명한 MSE를 감소시키는 학습방법이 아니라 참고문헌 [3]에서 제시한 방법을 사용한다.

2. Rectifier 심층 신경회로망

그림 1에 나타난 전방향 신경회로망은 중간층이 하나인 경우이다. 이 중간층의 수를 늘리면 신경회로망의 학습성능은 더 좋아질 것이며, 이와 같은 취지로 중간층의 수를 많이 지닌 신경회로망 구조를 심층 신경회로망(Deep Neural Networks)라고 한다. 이 심층 신경회로망은 학습을 잘 하기 위하여 학습 초기에 가중치들을 RBM(Restricted Boltzmann Machine)을 사용하여 정하는 복잡한 과정을 거쳐야 한다. 그렇지만 심층 신경회로망의 중간층 노드 함수를 그림 2에 주어진 tanh(.) 함수 대신 Rectifier 함수(그림 3)로 사용하면 RBM을 사용한 경우와 성능이 비슷하다고 밝혀졌으며[5], 이 논문에서는 이 방법을 사용하여 심층신경회로망을 학습시켰다. 한편 중간층 노드 함수를 Rectifier로 사용하면 tanh(.)를 사용한 경우 보다 중간층 노드 수가 많아져야 한다.



▶▶ 그림 2. 신경망의 중간층 노드 함수



▶▶ 그림 3. 심층신경망의 중간층 노드 함수

Ⅲ. 시뮬레이션

Protein Disorder Prediction의 시뮬레이션에 사용한 데이터 베이스는 KIAS(Korea Institute for Advanced Study)로부터 제공받았다. 723개의 단백질로부터 215,612개의 특징 벡터를 추출하였으며, 각 특징벡터는 330차원 sequence profile, 45차원 secondary structure profile, 16차원 solvent accessbility profile, 그리고 17차원 hydrophobicity profile로 구성된 408차원 벡터이다.

이 데이터를 우선적으로 입력 408, 중간층 노드 20, 출력층 노드 2(408-20-2) 구조의 신경회로망에 [3]의 방법(n=2,m=8)으로 학습률 0.001을 사용하여 학습시켰다. 그 다음으로 중간층 수를 3으로 증가시켜 408-20-20-20-2 구조 심층 신경회로망으로 n=2, m=8으로 학습시켰다. 마지막으로 408-30-30-30-2 구조 심층 신경회로망에 중간층 노드는 Retifier를 지니도록 하여 n=2, m=6으로 학습시켰다. 시뮬레이션 결과 Retifier를 사용하여 심층 신경회로망을 구성한 방법이 타 방법보다 향상된 결과를 보였다.

Ⅳ. 결론

이 논문에서는 Protein Disorder Prediction을 신경회로망을 이용하여 학습시켜 보았다. 학습 방법은 data imbalance를 고려한 방법을 사용하였으며, 중간층이 하나인 408-20-2 구조 신경회로망과 중간층 수를 3으로 늘린 408-20-20-20-2 구조 심층 신경회로망, 그리고 중간층 노드의 함수를 rectifier를 사용한 408-30-30-30-2 구조 심층 신경회로망을 이용하였다. 시뮬레이션 결과 Rectifier 심층 신경회로망이 좋은 성능을 보였다. 그렇지만, test 성능을 향상시키기 위하여 더 많은 노력이 필요하다.

■ 참 고 문 헌 ■

[1] F. Ferron, S. Longhi, B. Canard, and D. Karlin, "A practical overview of protein disorder prediction methods," PROTEINS: Structure, Function, and Bioinformatics, vol. 65, pp. 1-14, 2006.

[2] S.-H. Oh, "Protein disorder prediction using multilayer perceptrons," Int. J. Contents, Vol. 9, No. 4, pp 11-15, 2013.

[3] S.-H. Oh, "Error back-propagation algorithm for classification of imbalanced data", Neurocomputing, vol. 74, pp. 1058-1061, 2011.

[4] Y. Bengio, "Learning Deep Architecture for AI," Foundations and Trends in Machine Learning, vol.2, pp. 1-127, 2009.

[5] X. Glorot, A. Bordes, and Y. Bengio, "Dee sparse rectifier neural networks," Proc. AISTAS, Fort Lauderdale, FL, USA, Vol. 15, pp. 315-323, 2011.