단위 신경망을 이용한 단백질 기능 예측

황 두 성†

요 약

단백질의 기능 예측 모델은 guilt-by-association 개념을 바탕으로 단백질-단백질 상호작용 맵을 이용하고 있다. 이 방법은 목표 단백질이 기능이 알려진 단백질과 상호작용이 없는 경우 기능 예측이 불가능하다. 본 논문에서는 단백질 기능 예측 모델을 K-class 다중 분류 문제로 재정의하고 단백질-단백질 상호작용 데이터 및 단백질의 알려진 속성 등을 학습 모델에 이용한 단위신경망의 설계와 응용을 제안한다. 제안하는 모델은 Yeast 단백질 데이터의 기능 예측에서 단백질-단백질 상호작용 데이터를 이용하는 방법에 비해 분류 예측율에서 우수한 성능을 보였으며 또한 상호작용이 밝혀지지 않은 단백질의 기능 예측을 할 수 있다.

키워드 : 단위신경망, 다중 클래스 분류, 프로티오믹스, 단백질 기능 예측

Modular neural network in prediction of protein function

Doosung Hwang†

ABSTRACT

The prediction of protein function basically make use of a protein-protein interaction map based on the concept of guilt-by-association. The method however cannot determine the functions of proteins in case that the target protein does not interact with proteins with known functions directly. This paper studies protein function prediction considering the given problem as a K-class classification problem and proposes a predictive approach utilizing a modular neural network. The proposed method uses interaction data and protein related attributes as well. The experimental results demonstrate that the proposed approach can predict the functional roles of Yeast proteins whose interaction knowledge is not known and shows better performance than the graph-based models that use protein interaction data.

Key Words : Modular Neural Network, Multi-class Classification, Proteomics, Protein Function Prediction

1. 서 론

단백질 유전 정보학 또는 프로티오믹스 분야에서 생물학적 기능이 알려지지 않은 단백질의 기능 추론은 기본적으로 단백질-단백질 상호작용 데이터를 이용한다[1, 2, 3]. 특정 유기체의 단백질-단백질 상호작용 데이터는 단백질 기능 예측을 통한 단백질 복합체, 물질대사 경로, 물질대사 네트워크 등의 이해 및 분석에 효과적인 방법으로 알려져 있다[4, 5]. 상호작용 데이터를 그래프로 변환하여 단백질 기능을 예측하는 방법은 상이 단백질 간에 상호작용이 존재하는 경우에는 적합하지만, 대상 단백질이 기능이 밝혀진 단백질과 상호작용이 없는 경우에는 기능을 예측하는 것이 불가능하다.

최근 지능형 학습 알고리즘을 단백질 기능 분류 및 예측에 응용하는 사례가 증가하고 있는 추세이다. KDD Cup의 단백질 기능 예측을 위한 경쟁 데이터[6] 및 특정 유기체의 단백질 기능 추론에 지능형 학습 알고리즘들이 응용되고 있으며, 프로티오믹스 데이터의 분석에서도 그 효과가 있는 것으로 실험을 통하여 밝혀지고 있다. 학습 알고리즘을 단백질 기능 예측에 적용하는 방법론은 단백질-단백질 상호작용 데이터뿐만 아니라 단백질과 관련된 밝혀진 속성들까지 학습에 포함시키고 있다. 이러한 모델들은 그래프 기반 예측 기법과는 달리 기능이 알려진 단백질과 상호작용이 밝혀지지 않은 단백질의 기능 예측이 가능하다. 그러므로 학습 알고리즘을 단백질 기능 예측 모델에 응용하는 것은 단백질의 기능상 분류 및 분석에 따르는 실험 비용을 줄여 줄 것으로 기대된다.

본 논문에서는 단백질 기능 예측을 위한 단위신경망(modular neural network)의 응용을 제안한다. 이 연구에서 추구하는 바는 첫째, 신경망 모델과 같은 기계학습 알고리즘이 유기체 단백질의 기능 예측 및 분류에 적용될 때 발생될 수 있는 문제에 대해 논의하고자 한다. 둘째, 제안하는 단위신경망 예측 모델을 KDD Cup에서 우수한 결과를 보인 학습 모델과 상호작용 데이터를 이용하는 그래프 기반 모델

※ 이 연구는 2004년도 단국대학교 대학연구비의 지원으로 연구되었습니다.
† 종신회원 : 단국대학교 컴퓨터과학과 교수
논문접수 : 2004년 12월 11일, 심사완료 : 2005년 11월 30일

과 성능 면에서의 비교를 수행하는 것이다. 셋째, 제안하는 방법론을 KDD Cup과 같은 경쟁을 위하여 임의로 준비된 데이터가 아닌 실 유기체 단백질 데이터에서 실용성을 파악하는 것이다. MIPS 데이터베이스에서 제공된 Yeast 단백질 데이터에 대해 제안하는 예측 모델을 테스트하였다.

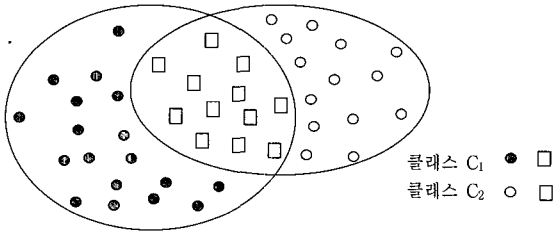
논문의 구성은 2절에서 단백질 기능 예측 모델에서 계산이론의 응용 측면에 대하여 살펴본다. 3절에서는 단위신경망 모델을 다중 분류 문제에 적용하는 방법을 제안하고, 4와 5절에서 KDD Cup 2001 및 MIPS Yeast 에 제안된 방법을 적용한 실험 결과를 보인다. 마지막으로 6절에서 제안하는 방법과 그래프 기반 모델의 비교, 단백질 기능 예측에 있어 해결해야할 어려움을 지적하고, 앞으로 연구방향에 대해 기술한다.

2. 관련연구

단백질 기능 예측의 응용에서 사용하는 데이터의 종류에는 단백질을 발현한 유전인자의 서열(sequence), 표현형(phenotype), 단백질 모티프(protein motif), 3차원 단백질 구조(structure) 및 단백질-단백질 상호작용 데이터 등이 있다. 단백질을 발현한 유전인자의 서열 유사성(sequence similarity) 및 서열 모티프(sequence motif)의 상동성(homology) 정보를 이용한 방법보다는 단백질 간의 상호작용 데이터를 사용하는 기능 예측 방법이 높은 신뢰성을 제공한다고 알려져 있다. 고 성능 실험 기법을 이용하게 됨에 따라 특정 유기체의 단백질-단백질 상호작용에 대한 대량의 데이터의 생성이 가능하게 되었으며 단백질의 기능 예측에 효과적인 계산 모델이 필요하였다.

단백질 상호작용 데이터를 이용한 단백질 기능 예측은 "guilt-by-association" 이론을 적용하여, 특정 단백질의 알려지지 않은 기능은 이미 기능이 알려진 다른 단백질과의 상호작용 여부에 따라 추론한다[4]. 따라서 단백질 기능 예측에 있어서 기초적인 방법은 기능이 알려진 단백질과 기능이 알려지지 않은 단백질의 상호작용을 확인하여, 상호 관계성을 그래프로 모델링하는 하는 것이다. 이러한 그래프를 이용하는 이론은 단백질의 기능 예측에 적합한 모델이 될 수 있으며, 각 단백질의 기능 예측은 직접적인 상호작용 데이터를 이용하기 때문에 기계학습의 관점에서 인스턴트 기반 학습[7] 이라고 할 수 있다. 그래프를 이용한 예측 모델의 적용은 neighbor-counting 모델[1], χ^2-통계치를 이용한 모델[3], markov random field 모델[8], simulated annealing[9] 등이 응용되었다. 그러나 이러한 접근방법은 단백질이 기능이 알려진 단백질과 상호작용이 없는 경우 기능 예측이 불가능하며 이미 밝혀진 단백질의 속성 데이터를 사용하지 못하고 있다.

학습 이론 기반 예측 모델은 단백질 상호작용 데이터, 서열 및 모티프 유사성, 또는 기능에 관련 있는 알려진 속성 등이 이용한다. 서열 유사성과 상호작용 데이터의 단백질 기능에 대한 계층적 군집화의 응용[10]. 단백질 서열 모티프와 기능에 대한 decision tree C4.5의 응용[11], gene ontology와 서열 모티프와의 logistic regression 분류기의 적용[12] 그리고 단백질 상호작용 데이터 및 단백질 속성 등을 학습에 포함하는 연상규칙의 학습 모델 응용[13], 그리고 KDD Cup 경쟁에서 응용된 support vector machine(SVM), inductive logic programming(ILP), Bayesian 모델 등이 높은 기능 예측 결과를 보인 것으로 보고되었다[6]. 살펴본 학습 모델들의 응용성에 대한 검증은 관련 웹 데이터베이스로부터 수집된 데이터 또는 실험을 통해 수 데이터 수집된 데이터를 가지고 테스트되었다.



(그림 1) 2-class 분류 문제의 예

3. 단위신경망 학습 분류 모델

지능형 학습 모델에 K-class(K>2) 다중 분류 문제를 응용하는 것은 각 학습 데이터가 K개중 하나의 클래스로 구분하여 분류를 구성하는 것이 일반적이다[14, 15]. 그러나 KDD Cup의 Yeast 단백질 데이터 분류에 있어서의 문제는 단백질은 한 개 이상의 생물학적 기능을 가질 수 있다는 점이다. 따라서 주어진 기능 예측 문제는 하나 이상의 클래스에 학습 데이터가 포함되어야하는 K-class 다중 분류 문제로 귀착된다. 이러한 K-class 다중 분류 문제를 해결하기 위한 단일 분류기를 사용할 경우 여러 클래스에 속하는 특정 데이터의 학습방해 효과 때문에 분리경계의 학습이 어려워 우수한 성능을 기대하기 어렵다. 그러므로 주어진 학습 데이터로부터 K-class 다중 분류 문제를 식별하고 단위 접근방법(modular approach)을 적용한 클래스 분리를 사용하여 K개의 단순 이진 분류기(2-class classifier)들을 설계하는 것이 효과적이다.

각 이진 분류기는 one-against-the-rest를 기반으로 하여 준비된 데이터에 대해 학습을 수행한다. 이러한 단위 접근방법을 이용한 학습은 K개의 클래스를 단일 분류기로 학습시키는 경우에 비해 학습이 단순하기 때문에 일반적으로 빠르게 진행되며 병렬 학습이 가능하다. (그림 1)과 같은 2-class 다중 분류 문제의 학습 모델 구성 시 단일 분류기의 설계는 새 클래스 $C_1 \cap C_2$ 에 속하는 학습 데이터를 다른 새 클래스로 구분하여 학습시켜야 한다. 그러나 단위 접근방법의 클래스 분리를 이용하여 두 개의 이진 분류기 M_1와 M_2를 사용하면 단일 분류기의 이용보다 학습의 일반화를 높일 수 있다. M_1의 학습 데이터는 해당 클래스에 속하는 positive 데이터 C^+와 속하지 않는 negative 데이터 C^-의 합

〈표 1〉 2-class 분류 문제를 학습데이터의 구성

분류기	학습 데이터	
	C^+	C^-
M_1	C_1	$C_2 - C_1 \cap C_2$
M_2	C_2	$C_1 - C_2 \cap C_1$

집합이 된다. 2-class 분류 문제를 위한 두 분류기 M_1, M_2의 학습데이터의 구성을 보이고 있다.

K개의 기능을 갖은 단백질 기능 예측 문제에서 각 기능 클래스의 예측 모델은 단위신경망(modular neural network)을 선택하였고, 학습은 역전파 알고리즘을 사용하였다. 단백질 기능별 학습 데이터가 K개의 클래스 데이터의 집합 $C_i(i=1,2,...,K)$으로부터 구성될 때, 각 단위신경망 M_i의 학습 데이터는 다음과 같이 C_i^+와 C_i^-의 집합이다.

$$C_i^+ = C_i$$

$$C_i^- = \bigcup_{j=1}^{K} \{C_j - C_i \cap C_j\}, j \neq i$$

단백질 기능 학습을 위한 단위신경망 M_i의 학습 데이터 크기 $|C_i^+|$와 $|C_i^-|$의 차이가 현저할 때, 역전파 학습 알고리즘은 클래스 불균형 효과로 인하여 학습 데이터가 많은 클래스의 학습이 빨리 발생하게 된다. 따라서 학습 데이터가 부족한 클래스는 학습이 잘 되지 않아 일반화 성능에 영향을 주게 된다[16]. 일반적으로 다중 분류 문제에 대한 단위 접근방법을 응용할 때는 one-against-the-rest 방법을 사용하기 때문에 클래스 분류기의 학습 데이터 준비는 높은 클래스 불균형비율을 나타낸다. 이러한 단점을 보완하기 위해 SMOTE(Synthetic Minority Over-sampling Techniques, [17]) 샘플링 기법을 이용하여 새로운 학습 데이터를 학습 데이터가 부족한 클래스에 추가 시켜 분류기 M_i의 학습을 진행시킨다. SMOTE 기법은 데이터의 수가 적은 클래스의 데이터를 nearest neighbor 규칙을 사용하여 데이터를 선택 후 아주 작은 임의 값을 더하여 새로운 학습 데이터를 생성시킨다.

(그림 2)는 제안하는 K개의 2-class 단위신경망의 학습 과정을 보인다. 여기서 Sample(E,κ)는 학습데이터 E에 κ개의 새 학습 데이터를 SMOTE 기법을 이용하여 생성시키는 함수이다. 단위신경망 M_k의 학습은 학습 데이터 E_k를 준비 후 SMOTE 샘플링을 이용하여 새 학습 데이터를 데이터 수가 적은 클래스에 추가를 하여 학습시킨다. $E_k(\approx E_k^+ \cup E_k^-)$의 준비 시 $(x,1) \in E_k^+$ 그리고 $(x,0) \in E_k^-$인 x는 학습을 방해하는 학습 데이터가 됨으로 E_k^-로부터 삭제한다.

단백질 기능 예측에서 단위 접근방법을 이용한 K개의 2-class 신경망의 적용 단계는 학습 방해 문제를 위한 클래스 데이터 분석, 클래스 분리 및 one-against-the-rest 방식의 K개 학습 데이터 준비, 데이터 코딩 및 교차검증(cross-validation)을 위한 데이터 준비, 클래스 분류기의 학습 및 교차검증, 마지막으로 테스트단계로 세분화된다.

```
Build_Modular_Network( TrainSet, K, a )
// TrainSet 는 학습 데이터 (x,t) 집합. x는 입력 데이터이며
   t={1,...,K}는 클래스
// K는 클래스의 총 수
// a는 subordinate class에 대한 오버샘플링 비율
for k=1,2,3,...,K do
    1. Prepare the training data E_k for classifier M_k
       • add (x,1) to E_k^+ if (x,k) ∈ TrainSet and t=k
       • add (x,0) to E_k^- if (x,k) ∉ TrainSet and t≠k
       • delete (x,0) from if (x,1)∈E_k^+ and (x,0)∈ E_k^-
    2. Oversample a set of synthetic data from the subordinate
       class
       • if |E_k^+| ≪ |E_k^-| then add Sample(E_k^+, a×|E_k^+|) to E_k^+
       • if |E_k^-| ≪ |E_k^+| then add Sample(E_k^-, a×|E_k^-|) to E_k^-
       • update E_k = E_k^+ ∪ E_k^+
    3. learn a modular neural network M_k
       • construct M_k with hidden neurons
       • train M_k with E_k
```

(그림 2) 단위신경망을 이용한 K개의 2-class 분류기의 학습 방법

4. 실험 데이터의 준비

제안된 단백질 기능 예측 방법론의 검증을 위해 전처리 과정을 통해 KDD Cup과 MIPS Yeast 데이터베이스[18]로부터 준비된 학습 데이터에 대해 시뮬레이션을 수행하였다. Yeast의 KDD Cup 문제 데이터는 6개의 단백질 속성, 13개의 생물학적 단백질 기능 및 단백질-단백질 상호작용 데이터로 구성된다. 단백질 속성에는 4개의 essential, 23개의 단백질 클래스, 54개의 복합체, 11개의 표현형, 285개의 모티프, 16개의 염색체에 대한 속성 데이터가 있으며, 일부 속성 값은 아직까지 밝혀지지 않은 경우도 있다. 두 단백질의 상호작용 데이터는 발현 상관관계 계수와 물리적 또는 유전적 상호작용으로 표현된다.

KDD Cup 학습 데이터에 나타나는 유일한 단백질들의 수는 862개이며, 테스트 데이터에 나타나는 유일한 단백질의 개수는 381개이다. 단백질 이름을 고려하지 않고 속성 및 기능을 갖은 단백질의 수는 4,345개의 학습 단백질 및 1,928개의 테스트 단백질들이 속해 있다. 주어진 단백질의 수보다 데이터 수가 많은 이유는 하나의 단백질은 여러 개의 속성 및 기능들을 가질 수 있기 때문이다. 주어진 상호작용 데이터의 수는 학습에서 910개이며, 테스트에서 896개가 제공되었다. 하나의 단백질이 여러 개의 기능을 가질 수 있다. 학습 데이터로부터 각 단백질은 평균 2.5개의 기능을 보이고 있어, 제안하고 있는 K-class 다중 분류 문제를 위한 단위 접근방법의 이용이 가능하다. 학습 데이터의 준비에 있어 각 속성에 속하는 데이터를 0을 포함하는 자연수로 대치하고 digit-to-binary 벡터를 생성시켰다. 단백질 속성 클래스 데이터의 경우 23개의 기술 데이터의 벡터 코딩은 첫 번째 digit은 3개와 두 번째 digit은 9개의 비트로 코드화 되어 12차원의 이진벡터가 필요하며 속성 값이 알려지지 않은 경우에는 벡터값이 모두 0이다.

단백질을 발현한 유전자 패턴간의 유사도인 발현 상관계

〈표 2〉 단일 신경망의 KDD Cup 데이터에서 테스트 결과

시도	1	2	3
학습(%)	91.7	91.2	91.4
CV(%)	91.8	91.1	87.3
테스트(%)	87.3	87.0	87.8

수(expression correlation) 데이터를 이용하여 학습 데이터의 단백질에 기능을 추가하였다. 단백질 A, B, C의 상호작용 (A, B)와 (B, C) 데이터가 그들의 발현 상관관계의 곱이 0.5보다 크면 (A, C)상호작용은 상동성에 의해 단백질 A에 나타나지 않은 C의 기능을 "guilt-by-association"에 의해 추가하였다.

MIPS Yeast 데이터베이스로부터 KDD Cup 데이터에서 고려된 단백질 기능 계층에 나타나는 단백질의 속성 및 기능을 추출하여 19 기능 클래스와 같은 6개의 속성에서 기술되는 데이터 집합을 생성시켰다. 생성된 데이터에는 4개의 essential, 24개의 단백질 클래스, 71개의 복합체, 12개의 표현형, 641개의 모티프, 16개의 염색체에 대한 기술 데이터가 있다. 유일한 단백질의 수는 3,510개이며, 구성된 단백질 기능 예측을 위하여 만든 관계형 데이터베이스로 부터 3개의 테스트 집합을 만들었다. 각 테스트 집합에 속하는 데이터는 다른 두 집합에는 나타나지 않으며, 해당 학습 데이터에는 포함되지 않도록 생성시켰다. 3개의 테스트 데이터 집합은 기능 클래스의 수가 100보다 크면 10%의 임의 선택된 데이터를, 그렇지 않으면, 25%의 단백질 데이터를 테스트 데이터로 포함시켰다. 테스트에 나타난 단백질의 수는 561(ES-1), 581(ES-2), 574(ES-3)이다.

5. 실험 및 결과

학습 데이터의 전처리 과정을 통해 준비된 KDD Cup 학습 데이터에 대해 단일 신경망과 제안하고 있는 단위 접근방법의 성능 비교를 수행하였다. 신경망 분류기의 구성에서 뉴런의 활성함수는 logistic 함수이며 학습에 이용된 알고리즘은 역전파 알고리즘의 학습 속도를 개선한 RPROP(resilient backpropagation)이다. 학습율은 0.1, 최대 학습 반복수는 20,000, 최소 학습 율의 변화는 0.001, 최대 가중치 변화는 50번, 가중치 변화 증가 및 감소 비율은 1.2이다. 학습에 사용된 시스템은 IBM PC Pentium III 800MHz이며 적은 데이터 클래스에 대한 오버샘플링 비율은 100%와 200%로 하였다.

<표 2>는 13개의 단백질 기능에 대한 단일 신경망의 KDD Cup 데이터에서 실험 결과를 보여주고 있다. 단일 신경망의 구조는 65×78×13으로 설정하였고, 삼원(3-way) 교차검증(CV)을 이용하여 테스트되었다. 단일 신경망의 경우 학습에서 91.4%, 교차검증은 91.0%의 학습 예측율, 그리고 테스트에서 87.3% 예측율을 보였다. 학습 및 테스트 예측율이 낮은 이유는 여러 클래스에 속하는 데이터의 학습이 잘 이루어지지 못하기 때문이다.

〈표 3〉 KDD Cup 데이터에 대한 제안하는 방법의 학습 및 교차검증 예측 결과

M	구분	Original	100%	200%
1	학습	99.60±0.10	99.71±0.10	99.67±0.06
	CV	94.39±1.43	95.57±0.73	96.89±1.35
2	학습	99.70±0.09	99.79±0.06	99.72±0.09
	CV	96.99±0.83	97.33±0.94	96.82±1.17
3	학습	99.70±0.10	99.67±0.12	99.78±0.07
	CV	97.71±0.64	98.11±0.69	98.45±0.61
4	학습	99.23±0.20	99.59±0.09	99.80±0.05
	CV	97.48±2.99	97.81±0.98	97.99±0.95
5	학습	99.23±0.20	99.37±0.24	99.23±0.29
	CV	97.48±2.99	97.86±0.95	97.06±1.59
6	학습	99.46±0.09	99.60±0.11	99.62±0.06
	CV	97.56±0.90	98.03±0.70	98.13±0.76
7	학습	99.78±0.02	99.81±0.03	99.75±0.06
	CV	99.16±0.45	99.33±0.45	99.42±0.37
8	학습	99.71±0.05	99.63±0.09	99.62±0.10
	CV	99.16±0.27	99.23±0.55	99.21±0.44
9	학습	99.69±0.09	99.79±0.08	99.82±0.06
	CV	95.78±0.86	96.88±1.23	96.99±0.62
10	학습	99.61±0.12	99.71±0.08	99.71±0.08
	CV	96.68±1.02	97.69±0.85	97.43±1.01
11	학습	99.71±0.05	99.73±0.09	99.65±0.10
	CV	99.42±0.24	99.26±0.23	99.18±0.32
12	학습	99.63±0.09	99.74±0.08	99.79±0.07
	CV	96.55±0.81	97.71±0.80	97.14±1.17
13	학습	99.68±0.05	99.70±0.10	99.78±0.05
	CV	99.13±0.22	99.30±0.35	99.46±0.47

단위 접근방법을 이용한 실험에서 이진 신경망의 구조는 65×16×1로 설정하였으며, 각 단백질 기능 클래스 학습 데이터의 10%를 교차검증 데이터로 선택하여 십원(10-way) 교차검증으로 테스트되었다. 표 3의 단위 접근방법의 실험에서 대부분 클래스 분류기의 학습과 교차검증은 99.0%와 97.0% 이상의 학습 예측율을 나타내고 있다. 표 2의 결과와 비교할 때 학습 및 교차검증의 예측율로부터 단위 접근방법이 단일 신경망에 비해 우수한 성능을 보이고 있다. 학습 성능에서 샘플링 전략을 사용하는 경우가 사용하지 않는 경우에 비해 근소한 차의 분류 성능이 증가되었다. 그러나 교차검증 테스트에서 샘플링 전략의 적용은 대부분 단위 신경망의 학습에 성능 향상을 보이고 있다. 데이터 수가 적은 클래스에 100%와 200%의 샘플링 전략의 이용은 학습 성능을 높이는데 효과가 있었다. 테스트 데이터 실험에서 평균 90.0% 이상의 예측율을 보였으며 200%의 샘플링을 사용한 경우 93.0%의 높은 예측율을 얻었다.

<표 4>는 KDD Cup 데이터를 가지고 높은 예측율을 보인 SVM와 ILP의 응용과 그래프 기반 학습인 2-NC(neighbor-counting), χ^2-통계치를 이용한 방법과 예측 성능 비교를 오분류행렬[1]로 보여주고 있다. 그래프 기반 방법이 336개의 테스트 단백질에 대해 기능 예측이 가능하였으며

1) TP:True Positive, TN:True Negative, FP:False Positive, FN:False Negative

〈표 4〉 KDD Cup 데이터의 성능 비교

	TP	TN	FP	FN	Acc(%)
SVM	690	4,304	58	282	93.6
ILP	654	4,264	90	326	92.2
MNN	668	4,295	84	287	93.0
2-NC	641	3,265	248	213	89.4
χ^2	432	3,234	428	284	83.9

2-NC의 89.4%와 χ^2-통계치의 83.9%의 분류 예측율을 나타내었다[20]. 학습 기반 모델은 381개의 모든 테스트 단백질에 대해 예측이 가능하며 우수한 예측율을 보여주고 있다. KDD Cup 데이터에 대한 예측율에서 SVM의 93.6%, ILP의 92.2%와 비교할 때 제안된 단위 접근방법의 응용은 경쟁력이 있다.

제안하고 있는 단위신경망 예측 모델이 SVM과 성능 비교에서 나타난 근소한 차는 단순신경망 구조 선택의 어려움과 학습 데이터 자체가 많은 의미 없는 속성 값을 가지고 있기 때문이라 예측된다. 신경망 구조의 중간층의 뉴런 수는 선 실험을 통해 선택하는 것이 일반적이다. 제안된 방법론의 실험에서 단순신경망 구조는 4 배수만큼의 중간층 뉴런 수를 설정하고 선 실험하여 최종 뉴런 수를 결정하였다. 중간층 뉴런의 수는 입력층의 뉴런 수보다 작게 제한하였다. 보고된 응용 연구로부터 SVM은 고차원 데이터 학습 문제, 잡음이 심한 데이터의 학습 문제, 가변 길이 데이터로 구성된 학습 문제 등에서도 높은 일반화 성능을 보였다. 그러나 kernel의 선택에 따른 계산비용이 너무 높다[19].

<표 5>는 단위신경망을 19개의 기능을 갖는 실제 MIPS Yeast 데이터의 실험에서 대부분 단백질 기능 분류기의 학습 데이터와 교차검증 데이터의 예측율은 KDD Cup데이터와 비슷한 97.0%와 95.0% 이상의 예측율을 보였으나 준비된 3개의 테스트 데이터에서 EQ-1의 85.16%, EQ-2의 85.11%, EQ-3의 84.0%의 결과를 내었다. 클래스 학습에 사용된 신경망의 구조는 75×24×1이며, SMOTE 샘플링은 200%로 설정하였고, 학습 알고리즘 및 설정 파라미터는 KDD Cup의 실험과 같다. 학습 및 교차검증의 결과보다 테스트의 성능 결과가 낮은 이유는 단백질의 기능 수는 약 4-5개이며, 단백질의 기능이 증가되었으나 충분한 학습데이터가 준비되지 않았으며, 속성 값이 알려지지 않은 데이터의 비율이 높기 때문이라 추측된다.

〈표 5〉 MIPS Yeast 데이터의 실험 결과

테스트	EQ-1	EQ-2	EQ-3
TP	1,258	1,273	1,242
TN	7,819	8,122	7,917
FP	94	97	101
FN	1,488	1,547	1,647
Total	10,659	101,039	10,906
Acc(%)	85.16	85.11	83.98

6. 결론 및 향후 과제

본 논문에서는 단위 접근방법을 이용한 신경망을 단백질 기능 예측에 적용하는 방법론을 제안하였다. 제안하고 있는 단위신경망은 Yeast 단백질 데이터의 실험에서 SVM과 ILP 등과 경쟁할 수 있는 학습 모델임을 실험으로 보였다. 제안된 방법은 대규모의 2차원 단백질-단백질 상호작용 맵을 이용한 그래프 이론 기반 방법들과 비교할 때 단백질 분자의 관련 지식을 포함시킬 수 있는 모델링을 제공한다. 그러므로 학습 모델의 응용은 상호작용 데이터로부터 예측 결과와 생물분자의 알려진 속성으로부터 예측을 하는 경험적 지식과 생물적 지식 예측 모델링이 가능하다는 장점을 제공한다.

실험을 통하여 단위신경망과 같은 학습 알고리즘을 이용한 단백질 기능예측은 그래프 기반 모델 기법보다 예측 성능이 우수함을 보였다. 최근 다양한 응용문제에서 우수한 결과가 보고되고 있는 SVM 모델의 주어진 문제에 적용은 quadratic programming의 해를 얻는데 학습 데이터의 크기가 증가함에 따라 높은 계산비용이 필요하게 되지만, 제안하고 있는 단위신경망을 이용한 방법은 비교적 적은 계산 비용으로 학습이 가능하다. 또한 SVM, ILP 보다는 신경망의 구현이 비교적 쉽다. 마지막으로 단백질 유전 정보학에 종사하는 전문가가 하나의 예측 모델을 이용하여 기능 예측을 하는 것 보다 여러 학습 모델의 예측 결과를 확보하여 그 결과를 이용하는 것이 보다 신뢰할 수 있을 것으로 기대된다. 앞으로 기능 예측 모델의 평가를 위한 데이터의 준비가 필요하며, 생물 정보학 전문가들과 협력을 통한 표준화된 데이터의 준비가 필요하며, 계산모델의 객관적 평가에 대한 연구 및 다른 유기체에 대한 모델 응용 연구가 필요하다.

참 고 문 헌

[1] B. Schwikowski et al., "A network of protein-protein interactions in yeast," Nature Biotechnology, Vol.18, No.3, pp.1257-1261, 2000.

[2] M. Fellenberg et al., "Integrative Analysis of Protein Interaction Data," Vol.8, Intelligent Systems for Molecular Biology, AAAI Press, pp.152-161, 2000.

[3] H. Hishigaki et al., "Assessment of prediction accuracy of protein function from protein-protein interaction data," Yeast, Vol.18, pp.523-531, 2001.

[4] S. Oliver, "Guilt-by-association goes global," Nature, Vol.403, pp.601-603, 2000.

[5] C. L. Tucker et al., "Towards an understanding of complex protein networks," TRENDS in cell biology, Vol.11, No.3, pp.102-106, 2001.

[6] J. Cheng et al, "KDD Cup 2001 Report," SIGKDD Exploration, Vol.3, No.2, pp.47-64, 2002.

[7] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[8] M. Deng et al., "Prediction of protein function using protein-protein interaction data," Proceedings of the IEEE Computer Society Bioinformatics Conferences, 2002.

[9] A. Vazquez, et al.,"Global protein function prediction in protein-protein interaction networks," Nature Biotechnology, Vol.21, No.6, pp.697-700, 2003.

[10] Yonata Bilu and Michal Linial, "The Advantage of Functional Prediction Based on Clustering of Yeat Genes and Its Correlation with Non-Sequence Based Classifications," Journal of Computational Biology, Vol.9, No.2, pp.193-210, 2002.

[11] Xiangyun Wang et al., "Automated data-driven discovery of motif-based function classifiers," Information Science, Vol.155, pp.1-18, 2003.

[12] Xinghus Lu et al., "Automatic annotation of protein motif function with Gene Ontology terms," BMC Bioinformatics, Vol.5, No.122, 2004.

[13] T. Oyama et al., "Extraction of knowledge on protein-protein interaction by association rule discovery," Bioinformatics, Vol.18, No.5, pp.705-714, 2002.

[14] A. J. C. Sharkey, Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems, Springer, 1999.

[15] S. Haykin, Neural Network: A Comprehensive Foundation, Prentice Hall, 1998.

[16] N. Japkowicz, "The Class Imbalanced Problem: Significance and Strategies," Proceedings of the 2000 International Conference on Artificial Intelligence(IC-AI'2000), 2000.

[17] N. V. Chawlar et al., "SMOTE: Synthetic Minority Over-sampling Techniques," Journal of Artificial Intelligence Research, Vol.16, pp.321-357, 2002.

[18] MIPS Yeast Database, http://mips.gsf.de/proj/yeast/

[19] John Shawe-Taylor and Nello Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, pp.47-82, 2004.

[20] 황두성, 정재영, "단백질 기능 예측을 위한 그래프 기반 모델링," 정보처리학회논문지 B, 제 12-B권, 제 2호, 2005.

황 두 성

e-mail : dshwang@dankook.ac.kr

1985년 충남대학교 계산통계학과(학사)

1990년 충남대학교 대학원 계산통계학과(석사)

2003년 Wayne State University, Computer Science(박사)

1990년~1991년 국토개발연구원 연구원

1991년~1998년 전자통신연구소 연구원

2003년~현재 단국대학교 컴퓨터과학과 교수

관심분야 : 데이터 마이닝(data mining), 머신 학습(machine learning), 바이오인포매틱스(bioinformatics)