2003년도 대한전기학회 하계학술대회 논문집 2003.7.21-23

난수발생기를 이용한 일반화된 회귀신경망 분류기와 프로모터 분류에의 응용

김근호[1], 김병환[1,*], 김경남[2], 홍진한[3]
[1,*]세종대학교 전자공학과, [2]세종대학교 분자생물학과, [3]마크로젠 DNA 칩 부서

A GRNN classifier using random generator and application to classifying promoters

Kunho Kim[1], Byungwhan Kim[1,*], Kyungnam Kim[2], Jin Han Hong[3]
[1,*]Sejong University , Electronic Engineering, [2]Molecular Biology, [3]Macrogen, DNA Chip Division

Abstract - 난수발생기 (Random generator-RG)와 GRNN을 이용한 분류기 설계방식을 제안하며, 이를 프로모터 염기서열의 분류에 적용한다. 주어진 난수범위에서 다중 분류기를 발생하였으며, 그 성능을 예측정확도와 분류민감도 측면에서 평가하였고, 분류민감도는 다시 전체와 개별적 프로모터에 대해서 세분화하여 평가하였다. 최적화된 분류기 상호간의 비교에서, 제안된 기법은 모든 임계점에 대해서, 전체 분류민감도와 전체 예측정확도를 향상시키었으며, 이는 전체 분류 민감도에서 더 두드러졌다. 한편, 개별적 프로모터에 대한 분류민감도와 예측정확도도 평균적으로 향상되었다. 이 같은 결과로 제안된 기법이 분류와 예측성능을 동시에 증진하는데 매우 효과적임을 알 수 있었다.

1. 서 론

DNA 칩 정보로부터 질병진단과 신약개발을 위한 유용한 생물학 정보를 추출하기 위한 연구가 활발히 진행이 되고 있다. 인공신경망은 무정형의 DNA염기서열상의 전사초기점과 종점등과 같은 프로모터 확인을 위한 중요한 정보를 예측하고 분류하는데 이용되고 있다 [1-3]. 염기서열분석에는 다양한 종류의 신경망이 응용되고 있다. 이 중, 일반화된 회귀신경망 (Generalized regression neural network-GRNN) [4]은 구조가 간단하고 성능최적화를 위한 제어인자도 하나밖에 없어, 많은 응용이 기대되고 있다. 일반적으로, GRNN의 성능은 spread 변수값을 실험적으로 결정하여 최적화하며, 패턴층의 뉴런은 동일한 최적화된 spread를 가지게 된다. 이에 따라 GRNN의 성능을 개선하는데에는 한계가 있었다.

본 연구에서는 난수발생기 (RG)를 이용하여 GRNN의 성능이 다중값을 가지는 spred에서 최적화가 되게하는 기법을 제안하며, 본 기법을 프로모터 염기서열의 분류에 적용한다. 고정된 난수범위에서 다중 분류기를 발생하며, 이중 결정된 최적의 분류기의 성능을 분류민감도와 예측정확도 측면에서 평가한다. 분류민감도는 임계점(Threshold) 값을 변화시켜 평가한다. 제안된 분류기를 종래의 분류기와 그 성능을 비교평가한다.

2. 본 론

2.1 실험데이터 수집

프로모터 데이터는 Oriza Sativa (OS), Arabidopsis Thaliana(AT), Escherichia Coli(EC), Zymomonas Mobils(ZM)의 각기 다른 4가지 유형으로 구성된다. 이중에서 OS와 AT는 eukaryotic 프로모터, 그리고 EC와 ZM은 prokaryotic 프로모터에 각기 속한다. AT 프로모터는 전체 cDNAs [http://signal.salk.edu/cgi-bin/tdnaexpress]를 genomic DNA [hppt://arabidopsis.org]와 비교하여 추출하였고, OS 프로모터는 rice 데이터 베이스 [http://www.ncbi.nlm.nih.gov]에서 수집하였다. EC와 ZM 프로모터는 NCBI와 마크로젠 (Macrogen)데이터 베이스에서 각기수집하였다. 학습데이터는 115개의 프로모터 염기서열패턴으로 구성되며, 이는 다시 OS 20, AT 25, EC 35, ZM 35개로 구성된다. 테스트 데이터는 58개의 패턴으로 구성되며, 이는 다시 OS 13, AT 15, EC 15, ZM 15개로 구성된다. 각 염기서열 패턴은 146 base pair로 이루어졌다.

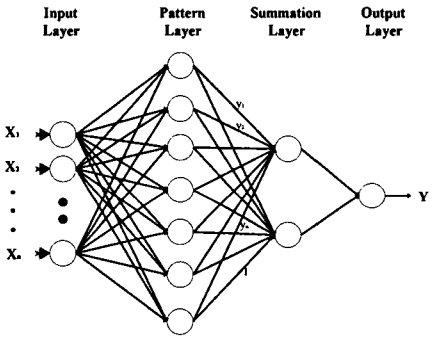

그림 1. GRNN 구조

2.2 일반화된 회귀 신경망

그림 1에서와 같이, GRNN은 총 4개의 층, 즉 입력층, 패턴층, 합층, 그리고 출력층으로 구성된다. 입력층의 뉴런수는 독립 변수 (예 공정 변수)의 수와 일치하며, 패턴층의 뉴런 수는 학습패턴의 수와 일치한다. 합층은 두 개의 뉴런으로 구성된다. 입력층과 패턴층간의 하중치(W_p)는 입력패턴 (X)에 의해 결정되며, 이를 표현하면,

$$W_p = X^t \tag{1}$$

패턴층의 하나의 뉴런은 합층의 두 개의 뉴런에 연결되며, 패턴층의 i번째의 뉴런과 합층 첫 뉴런간의 연결 하중치는 y_i가 된다. 이 i번째의 뉴런과 합의 다른 하나의 뉴런과의 연결 하중치는 1이 된다. 합측과 출력층간의 하중치 (W_s)는 y_i와 1에 의해 다음과 같이 결정된다.

$$W_s = [\text{Y ones}] \tag{2}$$

출력층에서는, 단순히 합층의 두 뉴런의 출력을 나누어 예측치를 출력한다. 임의의 입력패턴 x에 대한 예측치는 (3)으로 구해진다.

$$\hat{y}(x) = \frac{\sum_{i=1}^{n} y_i \exp[-D(x, x_i)]}{\sum_{i=1}^{n} \exp[-D(x, x_i)]} \quad (3)$$

여기서 x_i는 저장된 i번 째의 학습패턴을 지칭하며, n은 전체 학습데이터의 수를 의미한다.
(3)에서 함수 D는

$$D(x, x_i) = \sum_{j=1}^{p} \left(\frac{x_j - x_{ij}}{\sigma_j} \right)^2 \quad (4)$$

여기서 p는 각 입력패턴을 구성하는 전체 독립변수의 수를 지칭한다. x_j와 x_{ij}는 x와 x_i의 j번 째의 요소를 의미한다.그리고 변수 σ는 폭 spread라 불리며, GRNN의 성능을 결정하는 유일한 학습인자이다. 일반적으로 spread는 실험적으로 일정한 범위에서 결정하며, 결정된 spread는 그림 1의 패턴층을 구성하는 모든 뉴런에 대해서 동일하다. spread 값을 다중값으로 결정할 때, GRNN의 성능이 향상될 수 있으며, 이과 관련한 연구는 미미한 상황이다.

2.3 분류기의 성능 분석

분류기성능을 예측정확도와 분류민감도 측면에서 평가한다. 예측정확도는 (5)로 정의된 RMSE에 의해 계산한다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{q}(d_{ij} - out_{ij})^2}{pq}} \quad (5)$$

여기서 p와 q는 출력층 뉴런의 수와 테스트 패턴의 수를 의미한다. d_{ij}와 out_{ij}는 j번 째의 테스트 입력에 대한 i번 째 출력뉴런에 주어지는 실제치와 그 뉴런으로부터의 예측치를 의미한다. 분류민감도는 임의의 class로 정확히 분류되는 테스트 입력패턴의 수로 정의 된다. 분류민감도는 (6)의 임계점 (threshold)을 기준으로 더욱 세분화하며 평가한다.

$$|d_{ij} - out_{ij}| < Threshold \quad (6)$$

임계점은 0.6-0.9로 변화시켰으며, spred는 0.4-1.4범위에서 0.1간격으로 증가시켜 주었다. 한편, RG-기초 분류기의 경우 주어진 난수 범위에서 200개의 모델을 발생하였다.

2.3.1 고정 spread를 가지는 분류기

우선 종래의 방식으로 분류기를 설계한다. 주어진 spread에서 구성한 분류기의 전체 RMSE가 그림 2에 도시되어 있다.

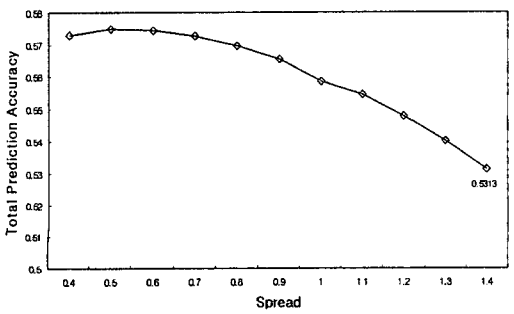

그림 2. 전체 RMSE의 spread에 따른 변화

여기서 RMSE는 테스트 패턴 전체에 대해서 계산된 값이다, 그림 2에서와 같이 RMSE는 spread의 증가에 따라 거의 선형적으로 증가하고 있다. 그림 2에서, 분류기는 spread가 1.4에서 최적의 RMSE (0.5313)을 가진다.

다음에는 각 spread에서 계산된 RMSE와 (6)을 이용하여 전체 분류민감도를 임계점의 함수로 계산하였으며, 그 결과가 그림 3에 도시되어 있다.

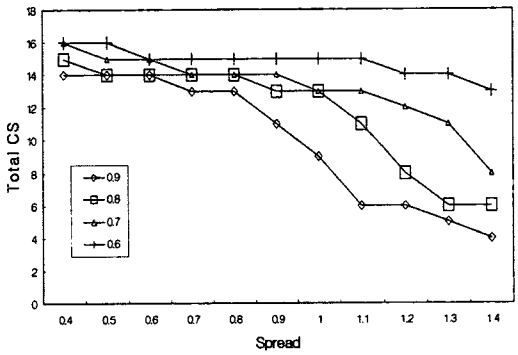

그림 3. 전체 분류민감도의 spread에 따른 변화

그림 3에서와 같이, 전체 분류민감도는 spread의 증가에 따라 일반적으로 감소하고 있다. 모든 임계점에 대해서, 분류기는 0.4에서 가장 우수한 전체 분류민감도를 보이고 있다. 각 임계점에서 결정된 전체 분류민감도를 각 프로모터 별로 다시 계산하였으며, 그 결과가 표 1에 정리되어 있다.

표 1. 전체 분류 민감도의 프로모터 별 분류

Threshold	OS	AT	EC	ZM	Total CS
0.9	6	0	4	4	14
0.8	6	0	4	5	15
0.7	6	0	5	5	16
0.6	6	0	5	5	16

표 1에서와 같이, 분류기는 OS인 경우에 대해서 가장 우수한 분류성능을 보이고 있으며, AT 프로모터인 경우, 임계점에 상관없이 전혀 분류를 하지 못하고 있다. 평균적으로 OS와 AT에 비해, EC와 ZM에 대한 분류성능이 더 나았으며, 이는 prokaryotic 유형의 프로모터의 분류가 더 용이하다라는 것을 암시한다.

표 2. 전체 분류 민감도의 프로모터 별 분류

Threshold	OS	AT	EC	ZM	Total CS
0.9	1	8	4	8	21
0.8	1	9	6	8	24
0.7	1	9	6	8	24
0.6	1	9	6	8	24

2.3.2 다중 spread를 가지는 분류기와 비교평가

RG를 이용하여 패턴층 뉴런이 다중값을 갖게 한다음, 그 성능을 평가한다. 임의의 난수범위에서, 발생된 200개의 각 분류기에 대해서 주어진 임계점에 대한 전체 분류민감도를 계산하였으며, 그 중 전체 분류민감도가 가장 큰 분류기에 대한 전체 분류민감도 값을 그림 4에 도시되어 있다. 그림에서와 같이, 전체 분류민감도는 그림 3에서와 같이 일반적으로 난수범위의 증가에 따라 감소하고 있다. 모든 임계점에 대해서 난수범위 0.4에서 전체 분류민감도는 가장 우수하였으며, 이는 그림 3에서 얻은 결과와 동일하다. 다음에는 각 임계점에 대해서 전체 분류민감도가 가장 우수한 분류기 성능을 각 프

로모터 유형별로 세분화하였으며, 그 결과가 표 2에 정리되어 있다. 표 1과 비교할 때, OS의 경우 분류성능이 매우 저하되었다. 그러나 AT의 경우 표 1에서의 결과와는 달리 상당한 분류성능을 보이고 있으며, 이 같은 결과는 매우 주목할 만하다. EC와 ZM에 대해서도 분류성능이 향상되었다. 평균적으로 prokaryotic 유형의 프로모터에 대한 분류성능이 우수하며, 이는 종래의 분류기에서 관측된 결과와 동일하다. 표 1과 비교할 때, 전체 분류성능은 임계점 0.9에 대해서 7개 향상되었고, 0.8에 대해서는 9개 향상되었다. 나머지 0.6과 0.7에 대해서는 8개가 향상이 되어, RG를 이용한 분류기의 성능이 더 우수하다라는 것을 알 수 있다.

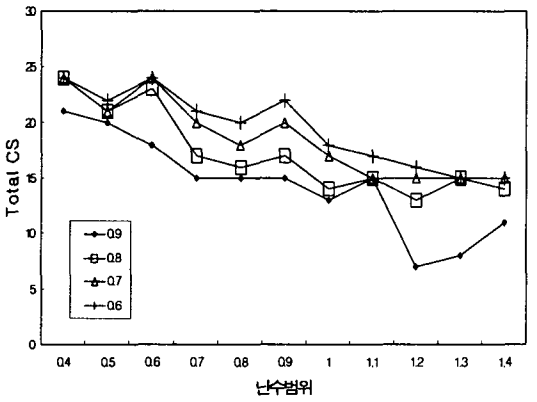

그림 4. RG-다중 spread 모델의 분류민감도

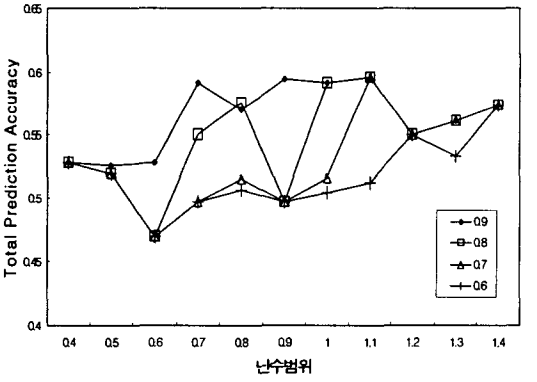

그림 5. RG-다중 spread 모델의 예측정확도

그림 5는 그림 4의 각 분류기에 대한 전체 RMSE를 도시하고 있다. 임계점이 0.9의 경우 난수 범위 0.4-0.5에서 전체 RMSE는 가장 작으며, 나머지 임계점에 대해서는 난수범위 0.6에서 가장 작다. 앞에서 알 수 있었듯이, 전체 분류민감도는 난수범위 0.4에서 최적화되었다. 이는, 전체 분류민감도와 전체 RMSE는 다른 난수 범위에서 최적화된다는 것을 알 수 있다.

다음에는 전체 분류민감도가 가장 우수한 분류기의 RMSE 성능을 전체 프로모터와 개별적인 프로모터에 대해서 비교 평가한다. RG-기초의 분류기는 난수범위 0.4에서 임계점에 관계없이 가장우수한 전체 분류민감도를 보였으며, 이에 해당하는 전체 RMSE는 공교롭게도 0.528에서 모든 임계점에 대해서 동일한 값을 가진다. 종래의 분류기도 동일한 spread에서 전체 분류민감도가 최적화 되었으며, 이에 해당하는 전체 RMSE는 0.573이 된다. 결국, 기존 분류기에 비해 RG-기초의 분류기가 7.8%정도 향상된 전체 RMSE를 보이고 있다. 다음에는 결정된 최적의 분류기의 성능을 각 프로모터 별로 비교 평가한다. 그 결과가 표 3에 정리되어 있다. 표 3은 임계점이 0.9에 대한 결과만을 포함하고 있으며, 이는 다른 임계점에 대해서도 동일한 결과를 얻었기 때문이다. 표 3에서와 같이, 기존 분류기에 대해서 AT, EC, ZM의 경우 그 RMSE가 향상되었으며, 단지 OS의 경우에 대해서만, 저하된 RMSE를 보이고 있다. 표 3에서와 같이, 일반적으로 RG-기초의 분류기 설계방식이 개별적 프로모터의 RMSE를 향상시키는데에도 매우 효과적임을 알 수 있다.

표 3. 최적의 분류민감도에 대한 예측정확도 비교

예측정확도	RG-GRNN	GRNN	Improvement
OS	0.675	0.518	-30.3%
AT	0.419	0.701	40.2%
EC	0.538	0.577	6.7%
ZM	0.480	0.574	16.3%

3. 결 론

RG-기초의 분류기 설계방식을 제안하였으며, 이를 프로모터 염기서열의 분류에 적용하였다. 분류기의 성능을 예측정확도와 분류민감도 측면에서 평가하였으며, 분류민감도는 임계점의 함수로 더 세분화하였다. 비교결과, 제안된 분류기는 전체 분류민감도를 획기적으로 증진시켰고, 각 프로모터에 대한 분류민감도는 평균적으로 증진시켰다. 또한 최적화된 분류기의 비교에서 알 수 있었듯이, 제안된 기법은 전체 RMSE와 개별적 프로모터에 대한 RMSE도 증진시켰다. 이 같은 결과는 제안된 RG-기초의 분류기 설계방식이 DNA 칩 데이터의 해석에 효과적으로 응용될 수 있음을 말한다.

감사의 글

본 연구는 IMT 2000 연구비에 의해 지원 받았으며, 한국보건산업진흥원에게 감사를 드립니다.

[참 고 문 헌]

[1] Gils M V, Jansen H, Nieminen K, Summers R, Weller P R, Using artificial neural networks for classifying ICU patient states, *IEEE EMB Mag.* 41-47, 1997.

[2] Knudsen S, Promoter 2.0: for the recognition of Pol II promoter sequences, *Bioinformatics* 15:356-361, 1999.

[3] Matis S, Xu Y, Shah M, Guan X, Einstein J R, Mural R, Uberhacher E, Detection of RNA polymerase II promoters and polyadenylation sites in human DNA sequence. *Comp. Chem.* 135-140, 1996.

[4] Specht D F, A generalized regression neural networks. *IEEE Trans Neural Networks* 2:568-576, 1991.