# 유전자 알고리즘과 일반화된 회귀 신경망을 이용한 프로모터 서열 분류

論 文
53D-7-9

# Promoter Classification Using Genetic Algorithm Controlled Generalized Regression Neural Network

金 性 模* · 金 根 鎬** · 金 秉 桓***
(Sungmo Kim · Kunho Kim · Byungwhan Kim)

**Abstract** - A new method is presented to construct a classifier. This was accomplished by combining a generalized regression neural network (GRNN) and a genetic algorithm (GA). The classifier constructed in this way is referred to as a GA-GRNN. The GA played a role of controlling training factors simultaneously. The GA-GRNN was applied to classify 4 different promoter sequences. The training and test data were composed of 115 and 58 sequence patterns, respectively. The classifier performance was investigated in terms of the classification sensitivity and prediction accuracy. Compared to conventional GRNN, GA-GRNN significantly improved the total classification sensitivity as well as the total prediction accuracy. As a result, the proposed GA-GRNN demonstrated improved classification sensitivity and prediction accuracy over the conventional GRNN.

**Key Words** : Promoter, Generalized Regression Neural Network, Genetic Algorithm, Classification

## 1. 서 론

Deoxyribonucleic acid (DNA) 칩 정보로부터 질병진단과 신약개발에 유용한 생물학 정보를 추출하기 위한 연구가 활발히 진행이 되고 있다. 인공신경망은 무정형의 DNA 염기서열상의 전사초기점 (transcription start site)과 같은 프로모터 확인을 위한 중요한 정보를 예측하고 분류하는데 이용되고 있다 [1-3]. 염기서열분석에는 다양한 종류의 신경망이 응용되고 있으며, 그 대표적인 구조가 역전파 신경망 (backpropagation neural network-BPNN) [4]이다. BPNN의 최적화에는 다수의 학습인자가 관여하고 있으며 [5-6], 초기 웨이트의 불규칙성으로 인해 그 최적화가 매우 어렵다. 특히, 다 변수, 대용량의 바이오 데이터 처리에 소요되는 계산이 상당하여, 그 응용이 제한을 받고 있다. 이에 따라, 구조가 간단하고, 성능 최적화가 용이한 신경망이 요구되고 있으며, 이에 적합한 신경망으로 일반화된 회귀신경망(generalized regression neural network-GRNN) [7]이 있다. GRNN은 학습이 용이하고, 학습제어인자도 가우시안 (gaussian) 함수의 spread 변수 하나밖에 없어, 분류기 내지 예측기의 설계가 용이하다. 바이오데이터 처리와 관련하여, 다른 신경망구조와 비교된 사례가 극히 드물어 GRNN의 성능 파악이 어려우나, 플라즈마 데이터에의 응용사례 [8]를 비추어 볼 때, GRNN의 예측성능은 통계적 회귀모델과 비슷하였다. 한편 BPNN은 플라즈마 데이터에 대한 예측모델개발에서, 통계적인 회귀모델에 비해 월등한 예측성능을 보였으며 [6], BPNN에 비해 상대적으로 낮은 GRNN의 성능은 패턴층 뉴런의 가우시안 함수가 동일한 값에서 최적화되는 것에 주로 기인한다. 따라서, GRNN의 예측성능을 증진하기 위해서는 가우시안 함수가 단일 spread가 아닌, 여러 값 (multi-valued)에서 최적화되게 하는 기법이 요구된다.

본 연구에서는 유전자 알고리즘(genetic algorithm-GA) [9]을 이용하여 GRNN의 성능이 multi-valued spread에서 최적화가 되게 하는 기법을 제안한다. 편의상, 제안된 분류기를 "GA-GRNN"이라 칭한다. 본 기법을 프로모터 염기서열의 분류에 적용하며, 그 성능을 예측정확도와 분류민감도 측면에서 평가한다. 한편, 제안된 분류기를 종래의 분류기와 그 성능을 비교 평가한다.

## 2. 프로모터 데이터

프로모터 데이터는 Oriza Sativa (OS), Arabidopsis Thaliana (AT), Escherichia Coli (EC), 그리고 Zymomonas Mobils (ZM)의 서로 다른 4가지 프로모터 서열들로 구성된다. 이중에서 OS와 AT는 eukaryotic 프로모터, 그리고 EC와 ZM은 prokaryotic 프로모터에 속한다. AT 프로모터 서열은 전체 cDNAs [10]를 genomic DNA [11]과 비교하여 추출하였고, OS 프로모터 서열은 rice 데이터베이스 [12]에서 수집하였다. EC 프로모터 서열은 NCBI로 부터, 그리고 ZM 프로모터 서열은 마크로젠 (Macrogen) 데이터베이스에서 수집하였다. 학습데이터는 115개의 프로모터 염기서열 패턴으로 구성되며, 이는 다시 OS 20, AT 25, EC 35, ZM 35개로 구성된다. 테스트 데이터는 58개의 패턴으로 구성되

* 學生會員 : 世宗大 電子工學科 碩士課程
** 學生會員 : 世宗大 電子工學科 博士課程
*** 正 會 員 : 世宗大 電子工學科 副敎授
接受日字 : 2003年 11月 10日
最終完了 : 2004年 4月 19日

며, 이는 다시 OS 13, AT 15, EC 15, ZM 15개로 구성된다. 각 염기서열 패턴은 146개의 base pair로 이루어졌다.

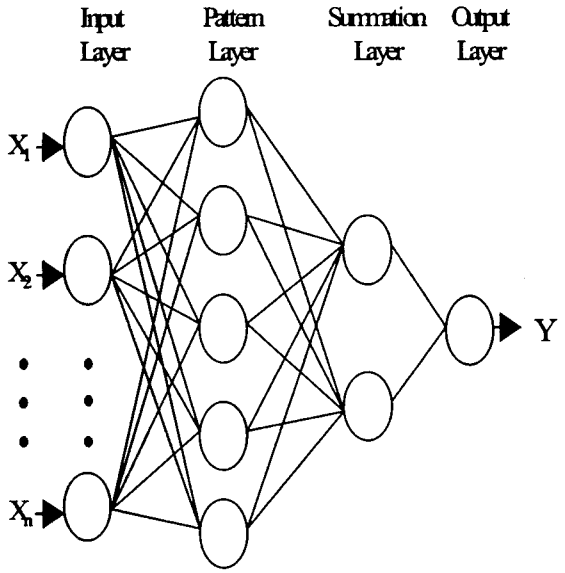


**그림 1** 일반화된 회귀 신경망 구조
**Fig. 1** Architecture of generalized regression neural network

## 3. 일반화된 회귀 신경망

그림 1에서와 같이, GRNN은 총 4개의 층, 즉 입력층, 패턴층, 합층, 그리고 출력층으로 구성된다. 입력층의 뉴런 수는 독립 변수의 수와 일치하며, 패턴층의 뉴런 수는 학습패턴의 수와 일치한다. 합층은 S-합층 뉴런들과 단일 D-합층, 이렇게 두 개의 뉴런으로 구성된다. S-합층 뉴런은 패턴층에 관한 가중된 출력들의 합을 계산하고, D-합층 뉴런은 패턴 뉴런들에 관한 가중되지 않은 출력들의 합을 계산한다. 입력층과 패턴층간의 하중치($W_p$)는 입력패턴 (X)에 의해 결정되며, 이를 표현하면,

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

여기서, "T"는 이항 (transposition)을 의미한다. 패턴층의 하나의 뉴런은 합층의 두 개의 뉴런에 연결되며, 패턴층의 i번째의 뉴런과 합층 첫 뉴런간의 연결 하중치는 $y_i$가 된다. 이 i번째의 뉴런과 합의 다른 하나의 뉴런과의 연결 하중치는 1이 된다. 합층과 출력층간의 하중치 ($W_s$)는 $y_i$와 1에 의해 다음과 같이 결정된다.

$$Ws=[Y\ 1] \tag{2}$$

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

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

여기서 $x_i$는 저장된 i번째의 입력 학습패턴을 지칭하며, n은 전체 학습데이터의 수를 의미한다. (3)에서 함수 D는

$$D(x, x_i) = \sum_{j=1}^{p} \left(\frac{x_j - x_i}{\zeta}\right)^2 \tag{4}$$

여기서 p는 각 입력패턴을 구성하는 전체 독립변수의 수를 지칭한다. $x_j$와 $x_{ij}$는 x와 $x_i$의 j번째의 요소를 의미한다. 그리고 변수 $\zeta$는 spread라 불리며, GRNN의 성능을 결정하는 유일한 학습인자이다. 일반적으로 spread는 실험적으로 일정한 범위내에서 결정하며, 결정된 spread는 그림 1의 패턴층을 구성하는 모든 가우시안 함수에 대해서 동일하다. spread 값을 여러 값에서 결정할 때, GRNN의 성능이 향상될 수 있으며, 이와 관련한 연구는 보고 된 바가 없다.

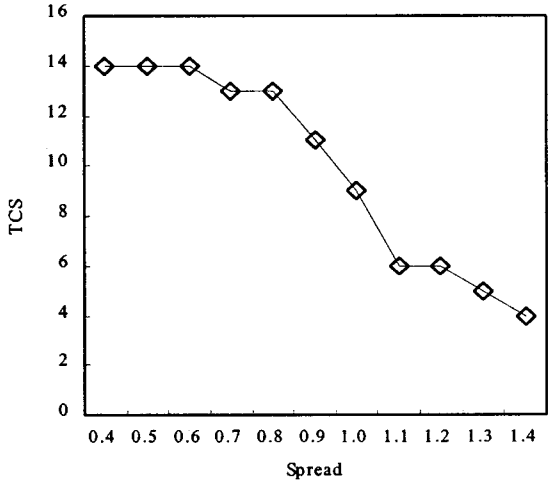


**그림 2** GRNN 분류기의 spread 변화에 따른 TCS
**Fig. 2** TCS of GRNN classifier as a function of spread.

## 4. 분류기의 성능 분석

### 4.1 성능 지표

제안된 분류기의 성능은 예측정확도와 분류민감도 측면에서 평가하였고, 다시 전체와 개별적 프로모터에 대해서 세분화하여 평가하였다. 분류민감도는 정의된 부류(class)로 정확히 분류되는 테스트 입력패턴의 수로 결정된다. 분류민감도는 전체분류민감도 (total classification sensitivity-TCS)와 개별적 프로모터에 대한 분류민감도 (individual classification sensitivity-ICS)로 나누어 평가하였으며, 각 지표는 (5)의 임계치 (threshold)를 기준으로 더욱 세분화하였다.

$$| d_{ij} - out_{ij} | < Threshold \quad (5)$$

여기서, $d_{ij}$와 $out_{ij}$는 j번째의 테스트 입력에 대한 i번째 출력뉴런에 주어지는 실제치와 그 뉴런으로부터의 예측치를 의미한다. 한편, 분류민감도는 여러 spread에서 동일할 수 있으며, 따라서 최적분류기의 결정을 위해서는 각 분류기의 예측성능을 평가하는 것이 요구된다. 예측 성능은 (6)으로 정의되는 root mean squared error (RMSE)를 을 이용하여 계산하였다.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{q}\sum_{j=1}^{r}(d_{ij}-out_{ij})^2}{qr}} \quad (6)$$

여기서 q와 r은 출력층 뉴런의 수와 테스트 패턴의 수를 의미한다. 분류민감도와 마찬가지로, 예측정확도도 전체예측정확도 (total prediction accuracy-TPA)와 개별적 예측정확도 (individual prediction accuracy-IPA)로 나누어 평가하였다.

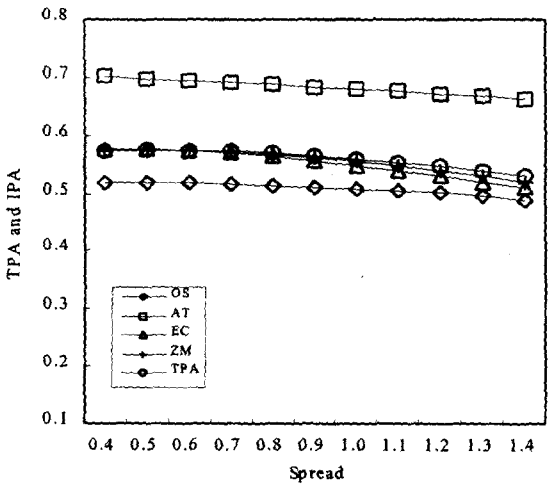


**그림 3** GRNN 분류기의 spread 변화에 따른 TPA와 IPA
**Fig. 3** TPA and IPA of GRNN classifier as a function of spread

### 4.2 종래의 GRNN 분류기

비교를 위해서, 우선 종래의 방식으로 GRNN 분류기를 설계하였다. 가우시안 함수의 spread는 0.4에서 1.4까지 0.1 간격으로 증가시켰다. 각 spread에서 분류기를 구성하였으며, 임계점 0.9를 기준으로 하여 TCS를 계산하였다. 그 결과가 그림 2에 도시되어 있다. 그림 2에서와 같이, TCS는 spread의 증가에 따라 일반적으로 감소하고 있다. 최대의 TCS는 14개이며, 이는 세 spreads (0.4, 0.5, 그리고 0.6) 에서 결정되었다. 그림 3은 분류기의 동일한 spread 범위에서의 TPA와 IPA를 보여주고 있다. 여기서 RMSE는 테스트 패턴 전체에 대해서 계산된 값이다. 그림 2에서와 같이, TPA와 IPA가 spread의 증가에 따라 감소하고 있다. 한편, 최적 분류기의 결정을 위해서 TCS가 가장 우수하였던 3 spreads, 즉 0.4, 0.5, 그리고 0.6에서의 TPA를 계산하였으며, 계산된 TPA는 모두 0.573으로 동일하였다. 편의상, 최적분류기를 spread 0.4에서 결정하였다. 최적분류기의 프로모터 종류별 분류성능을 확인하기 위해, TCS로부터 ICS를 계산하였다, 그 결과, OS, AT, EC, 그리고 ZM에 대한 ICS는 각 6, 0, 4, 그리고 4개이었다. 이로부터, GRNN 분류기는 AT서열은 전혀 분류를 하지 못하고 있음을 알 수 있다.

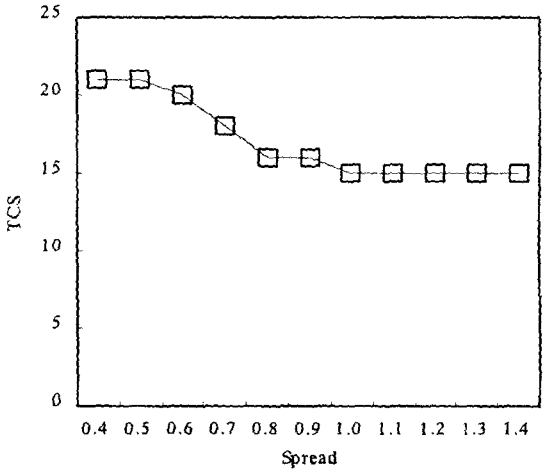


**그림 4** SGA-GRNN 분류기의 spread 변화에 따른 TCS
**Fig. 4** TCS of GA-GRNN classifier as a function of spread

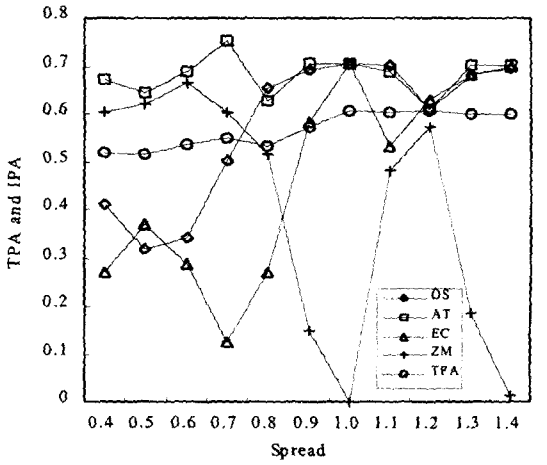


**그림 5** GA-GRNN 분류기의 spread 변화에 따른 TPA and IPA
**Fig. 5** TPA and IPA of GA- GRNN classifier as a function of spread

### 4.3 GA-GRNN 분류기

분류민감도 성능을 향상시키기 위해서 GA를 응용한다.

초기 해는 100개로 설정하였으며, 각 해에 해당하는 string (chromosome)의 길이는 115이며, 이는 학습패턴의 수와 일치한다. 구체적으로, 하나의 string을 구성하는 115개의 slots은 패턴층을 구성하는 동일한 수의 가우시안 함수의 spreds들로 구성된다. 각 string에 대한 적합도 (fitness) 함수는 (7)로 정의하였다.

$$\text{Fitness} = \text{TCS} \tag{7}$$

한편, 교차확률 (crossover probability)와 돌연변이 확률 (mutation probability)은 각기 0.9와 0.1로 설정하였다. GA 연산은 발생 수 (generation number)가 100일 때 종료하는 것으로 설정하였다. 한편, spread range는 GRNN의 경우에서와 같이 동일하며, 차이점은 주어진 spread 범위 내에서 string의 각 slot에는 랜덤 (random)한 값이 할당이 된다는 점이다. 그림 4는 GA-GRNN의 spread 변화에 따른 TCS를 나타내고 있다. 그림 4에서와 같이 TCS는 spread가 증가함에 따라 감소를 하다가 1.0 이상의 spread에 대해서는 15개로 일정하게 유지가 되고 있다. 그림 4에서와 같이 spread 0.4와 0.5에서 TCS가 가장 우수하였으며, 그 수치는 21개로 동일하다. 종래의 GRNN의 성능과 비교했을때 이는 7개가 향상된 결과이다. 그림 5는 GA-GRNN 분류기의 동일한 spread 범위에서의 TPA와 IPA를 보여주고 있다. 그림 5에서와 같이, TPA는 spread의 증가에 따라 증가하는 경향을 보이고 있으며, 이는 그림 3의 GRNN의 TPA와는 반대되는 경향이다.

표 1 최적 분류민감도에 대한 예측성능 비교

**Table 1** Comparison of prediction performance for optimized classification sensitivity

| Promoter type | GRNN | GA-GRNN | Improvement(%) |
|---|---|---|---|
| OS | 0.518 | 0.318 | 38.6 |
| AT | 0.701 | 0.644 | 8.1 |
| EC | 0.577 | 0.370 | 35.8 |
| ZM | 0.574 | 0.620 | -8.0 |

한편, IPA는 spred의 변화에 매우 민감하며, 그 변화도 매우 복잡함을 알 수 있다. 결국, GA-GRNN 분류기는 예측성능에 있어 GRNN 분류기와 매우 다른 양상을 보였다. 그림 4에서 최적의 분류 성능을 보인 spreads, 즉 0.4와 0.5에서 TPA를 계산해 보면 각기 0.517과 0.515이었다. 결국, spread 0.5에서 TCS와 TPA가 최적화된 분류기를 결정할 수 있었다. 최적화된 GRNN 분류기와 비교할 때, 최적화된 GA-GRNN분류기는 대략 10% 정도 향상된 TPA를 보였다. 최적화된 GA-GRNN 분류기의 TCS에 대한 ICS는 OS, AT, EC, 그리고 ZM에 대해서 각 6, 1, 10, 그리고 4였다. GRNN과 비교할 때, OS와 ZM에 대한 ICS는 동일하지만, EC의 ICS는 획기적으로 증가되었다. 특이점은, GRNN 분류기로 전혀 분류가 되지 않았던 AT에 대해서 GA-GRNN은 분류를 할 수 있었다는 사실이다. 결국, ICA측면에서 GRNN에 비해 GA-GRNN은 향상된 성능을 보였다. 한편, 결정된 GRNN와 GA-GRNN의 TPA를 각 프로모터별 IPA로 세분화하여 비교하였으며, 그 결과가 표 1에 나타나 있다. 표 1에서와 같이, GRNN 분류기에 비해 OS, AT, EC의 경우 IPA가 향상이 되었으며, 단지 OS의 경우에 대해서만 저하된 IPA를 보이고 있으나, 그 저하도는 매우 미미하다. 특히 OS와 EC의 경우 IPA의 향상이 두드러졌다. 결과적으로 제안된 GA-GRNN은 모든 4개의 평가지표에서 GRNN보다 향상된 성능을 보였다.

## 5. 결 론

본 GA-GRNN 분류기 설계방식을 제안하였으며, 이를 프로모터 염기서열의 분류에 적용하였다. 분류기의 성능을 예측정확도와 분류민감도 측면에서 평가하였다. 비교를 위해 종래의 방식으로 GRNN 분류기를 구성하였으며, 비교 결과, GA-GRNN 분류기는 TCS와 TPA 측면에서 우수한 성능을 보였다. 특히 TCS측면에서의 성능향상은 주목할 만하다. ICS와 IPA의 경우에도 향상된 성능을 보였으며, 이는 제안된 GA-GRNN 분류기 설계방식이 대 용량, 다 변수DNA 칩 데이터의 해석에 효과적으로 응용될 수 있음을 말한다.

## 참 고 문 헌


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

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

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

[4] D. E. Rummelhart, J. L. MaClelland, Parallel Distributed Processing, MIT Press, Cambridge, 1986.

[5] B. Kim and Gary S. May, "An optimal neural network process model for plasma etching," IEEE Trans. Semicond. Manufact., vol. 7, no. 1, pp. 12-21, 1994.

[6] B. Kim and S. Park, "An optimal neural network plasma model: a case study," Chemom. Intell. Lab. Syst., vol. 56, pp. 39-50, 2001.

[7] Specht D F, "A generalized regression neural networks." IEEE Trans. Neural Networks vol. 2, pp. 568-576, 1991.

[8] B. Kim and S. Park, "Modeling of process plasma using a radial basis function network: a case study," Trans. Contr. Autom. Syst. Eng., vol.2, no. 4, pp. 268-273, 2000.

[9] D. E, Goldberg, Genetic Algorithms in Search,

Optimization & Machine Learning, Addison Wesley, 1989,
[10] http://signal.salk.edu/cgi-bin/tdnaexpress.
[11] http://arabidopsis.org.
[12] http://www.ncbi.nlm.nih.gov.

## 저 자 소 개

**김 성 모 (金 性 模)**

1976년 12월 14일생. 2002년 남서울대학교 전자공학과 졸업. 2002년~현재 세종대학교 일반대학원 전자공학과 석사과정
Tel : 02-3408-3729
Fax : 02-3408-3329
E-mail : ksm1214@hanmir.com

**김 근 호 (金 根 鎬)**

1975년 6월 16일생. 2002년 세종대학교 전자공학과 졸업. 2002년~현재 동 대학원 전자공학과 박사과정
Tel : 02-3408-3729
Fax : 02-3408-3329
E-mail : gno0616@hanmail.net

**김 병 환 (金 秉 桓)**

1962년 10월 30일생. 1985년 고려대 전기공학과 졸업. 1987년 고려대 전기공학과 졸업(석사). 1995년 Georgia Institute of Technology 졸업(공박). 1996-1998년 현대전자 반도체 연구소 책임연구원. 1999-2001년 전남대 전기공학과 전임강사, 2001- 현재 세종대 전자공학과 부교수.
Tel : 02-3408-3729
Fax : 02-3408-3329
E-mail : kbwhan@sejong.ac.kr