Journal of Knowledge Information Technology and Systems

ISSN 1975-7700

http://www.kkits.or.kr

Time Series Expression E.coli Prediction for Gene Regulatory Network Reconstruction

Hee-Jin Yoon*

IT Collage, Jangan University

ABSTRACT

Presenting the relationship of gene and gene is Gene Regulatory Networks(GRN). In the relationship of gene and gene, one of the target gene is affected by regulate genes. Gene regulatory network is divided into two ways by regulate gene activate expression of target gene and repress it. This paper predicted time series expression E.coli data to reconstruction Gene Regulatory Networks. E.coli used in test data is composed of eight genes and has 50 time point. To predict a one target gene, the rest is used for regulator genes. To predict E.coli data, it used Neural Network with Weighted Fuzzy Membership Function(NEWFM). E.coli Data which has time series is used Wavelet for feature extraction. Feature is used for prediction of a Gene Regulatory Networks. By the NEWFM, features were trained and selected minimum feature by Weighted Fuzzy Membership Function Bounded Sum. Data value of selected features is defuzzificated by Takagi-Sugeno value, and by using previous time expression value of regulator genes predicted current time expression value of target gene. Predicted value indicated mean square error(MSE). As comparing, it shows that predicted result is improved by each MSE result; NFRN algorithm is 0.12%, EK algorithm is 0.12%, and NEWFM is 0.003%.

KEYWORDS: Fuzzy neural networks, Microarray data, Gene regulatory networks, Wavelet Transcriptions, Time series data

ARTICLE INFO: Received 12 August 2015, Revised 8 October 2015, Accepted 8 October 2015.

*Corresponding author is with the Department of Internet Communication, Jangan University, whasung kyungi-doo, KOREA.

E-mail address: hjyoon@jangan.ac.kr

1. 서 론

최근 유전자 간의 동적이고 불확실한 관계를 밝히려는 연구가 증대되고 있다. 그리하여 high-throughput이나 마이크로어레이의 기술의 발달로 유전자와 유전자의 상호관계를 예측 할 수 있게 되었으나, 생화학적 연구 방법은 외부적 환경과 낮게 발현되는 유전자에 대해 유전자 관계를 밝혀내지 못하는 단점을 가지고 있다. 이런 단점을 보완하기 위해 기계학습이나, 확률 통계적 기법을 이용하여 유전자 상호관계를 나타내고 있다[1]. 유전자와 유전자의 상관관계를 나타내는 네트워크를 유전자 조절 네트워크(Gene Regulatory Network, GRN)라 한다. 유전자 조절 네트워크에서 유전자 간의 영향을 미치는 유전자를 조절유전자(regulator gene)라 하며, 대상이 되는 유전자를 타겟유전자(target gene)라 한다[10]. 유전자 조절 네트워크를 연구하는 방법으로는 부울 네트워크(Boolean Network)[2, 3] 확률적 기법을 이용한 베이지안 네트워크(Bayesian Network)[11], 베이지안 네트워크에 피드백을 추가하여 네트워크를 구성한 동적 베이지안 네트워크(Dynamic Bayesian Network)[4], 신경망을 이용한 Neural Fuzzy Recurrent Network(NFRN)[5] 등이 있다. 부울 네트워크는 계산이 간단하나 정의의 한계가 있다. 또한, 확률적인 기법을 이용한 베이지안 네트워크는 확률적인 관계로 유전자의 관계를 추론한다는 큰 장점을 가지고 있으나, 피드백이 이루어지지 않으므로 시계열 데이터를 처리 할 수 없다는 단점을 가지고 있다. 베이지안 네트워크의 단점을 보완한 동적 베이지안 네트워크는 피드백으로 시계열을 처리 할 수 있으나, 유전자의 상관관계에서 활성자와 억제자를 구분 하지 못하는 단점을 가지고 있다[12]. 또한 NFRN은 if-then 형식의 추출 기법을 이용하여 퍼지규칙을 단순화 시켜, 데이터를 학습시키는 장점이 있으나, 활성자와 억제자의 구분을 할 수 없는 단점이 있다. 유전자 조절 네트워크는 하나의 타겟 유전자에 대해 나머지 유전자 데이터가 조절유전자가 되어 유전자 간의 상관관계를 예측하고 나타낼 수 있다.

본 논문은 가중퍼지소속함수기반 신경망(Neural Network with Weighted Fuzzy Membership Function, NEWFM)[6]을 이용하여 시계열 E.coli 데이터의 각 유전자를 예측하였다. 가중퍼지소속함수기반 신경망(NEWFM)은 가중퍼지소속함수의 경계합을 이용하여 클래스분류에 대해 우수한 최소의 특징을 선택 할 수 있게 하여, 타겟유전자에 대한 예측을 가능하게 한다. 또한 베이지안과 신경망인 NFRN의 단점을 보완 하였다. NEWFM을 이용하여 SOS E.coli 데이터 간의 예측 실험을 한 결과 평균 제곱오차의 평균 값 0.0039의 Elman 알고리즘 0.0103, NFRN 알고리즘 0.22로 다른 알고리즘보다 향상된 결과를 보여주고 있다.

2. 가중퍼지소속함수기반 신경망

2.1 가중퍼지소속함수기반 신경망의 구조

가중퍼지소속함수기반 신경망은 <Figure 1>에 나타나 있는 것과 같이 입력노드에서 특징을 입력받아 하이퍼박스에서 클래스 2개로 분류(B1과 B2)되어 학습되어지며 각각 특징의 하이퍼박스들은 대, 중, 소의 가중퍼지소속함수를 갖는다. 가중퍼지소속함수 경계합에 의해 특징선택이 되어지고 최소로 선택된 특징들은 Takagi-Sugeno에 의해 역퍼지화 되어진다[7]. 역퍼지화 된 값은 예측을 위해 사용된다.

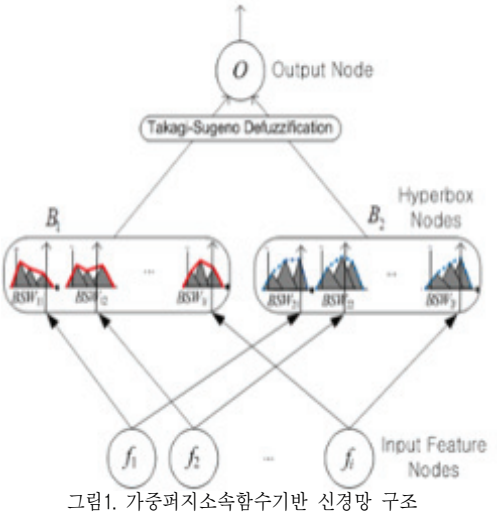

그림1. 가중퍼지소속함수기반 신경망 구조
Figure 1. Structure of Neural Network Weighted with Weighted Fuzzy Membership Function

3. E.coli 데이터 유전자 간 예측

<Figure 2>은 E.coli 데이터 유전자 간 예측의 처리과정을 보여주고 있다.

그림2. E.coli 예측 과정
Figure 2. Processing of E.coli prediction

1 step: 전처리과정으로 정규화 및 클래스 분류하는 과정이다.

2 step: 시계열 특성에 맞는 특징 추출 단계이다.

3 step: 가중퍼지소속함수기반 신경망의 가중퍼지소속함수 경계합을 이용하여 특징 선택 단계이다.

4 step: 유전자-유전자 예측 단계로 Takagi-Sugeno 기법[13]의 역퍼지화를 이용하여 예측한다.

3.1 실험데이터

유전자 간 예측을 위해서 사용한 데이터는 Alon's 홈페이지 (http://www.weizmann.ac.il/ UriAlon/ Papers/SOSData/)에서 제공하는 시계열 E.coli 데이터를 사용하였다. E.coli 데이터는 8개의 유전자:uvrD, lexA, umuD, recA, uvrA, uvrY, ruvA, polB로 이루어졌으며, 6분간격의 50개의 타임포인트를 갖는다. 아래 <Figure 3>은 실험에 사용되어진 E.coli데이터를 보여주고 있다.

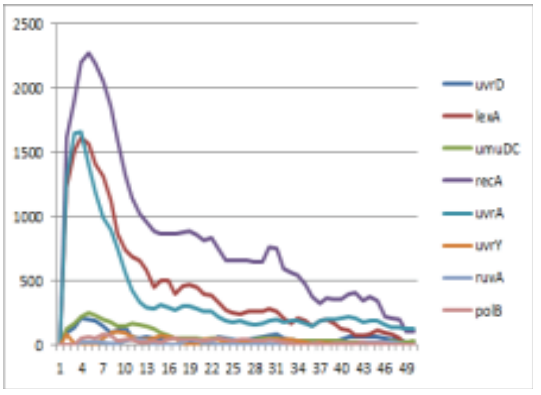

그림 3. E.coli의 8개 유전자 실제데이터
Figure 3. E.coli of 8 gene real data

데이터의 잡음을 제거하기 위해 전처리 과정으로 시그모이드 함수를 이용하여 정규화 하였다. 클래스 분류는 샘플의 평균값을 기준으로 하여 분류하였다. 입력데이터의 형태
I={gene1_feature1,gene2_feature2
.... gene n_feature n, class분류}의 형태를 갖는다.

3.2 특징추출

유전자의 시계열 데이터를 이산화하여 측정할 수 있게 하기 위하여 Harr Wavelet변환을 사용하였다[8]. 입력데이터를 A(approximation)과 D(detail)

로 분해하고, 다시 A(i)와 D(i)로 분해해 가는 과정이지만, E.coli예측 실험에서는 샘플의 수가 많지 않으므로 A1만 사용하였다. A1은 수식1과 같다.

$$A1 = (t_n + t_{n-1})/\sqrt{2} \quad (1)$$

t_n은 현재시간의 데이터를 의미하며, t_{n-1}은 이전 시간의 데이터를 의미한다. 시계열 데이터는 이전 time point의 데이터가 현재 time point의 데이터를 예측 할 수 있게 한다. 이 실험에서도 Figure 4.와 같이 time point를 적용하여 타겟유전자를 예측하였다[9].

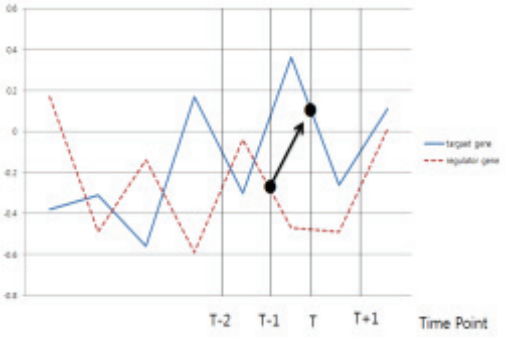

그림 4. time point의 예측

Figure 4. Prediction of time point

3.3 예측

E.coli의 8개 유전자를 예측하기 위하여 가중퍼지소속함수기반 신경망으로 특징선택 한 후 Takagi-Sugeno 기법으로 역퍼지화 한 값을 이용하였다. 각 샘플(time point)의 값은 Takagi-Sugeno 평균과의 비율을 계산하여 예측 값을 구하였다. <Figure 5>와 <Figure 6>은 실제 데이터와 예측한 값을 그래프를 나타내고 있다.

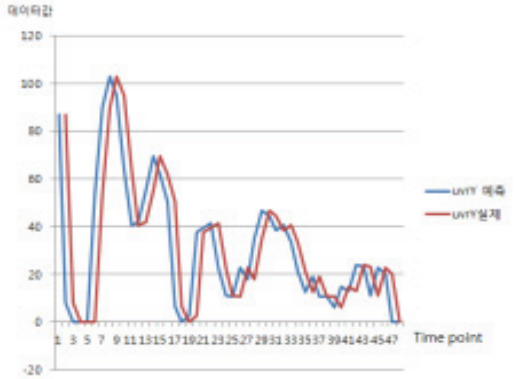

그림 5. uvrD의 예측: 빨간색은 실제데이터이고 파란색은 예측 값

Figure 5. Predict of uvrD: red line is real data and blue line is predicted value

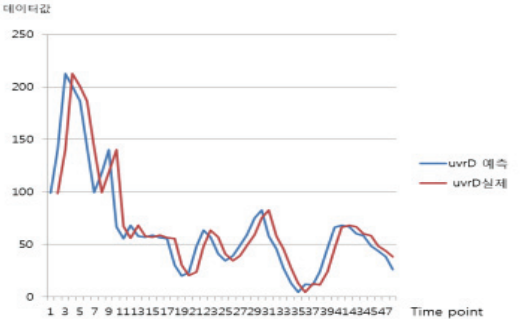

그림 6. uvrY의 예측: 빨간색은 실제데이터이고 파란색은 예측 값

Figure 6. Predic of uvrY: red line is real data and blue line is predicted value

4. 실험결과

가중퍼지소속함수기반 신경망을 이용하여 시계열 E.coli 데이터에 웨이블릿 특징을 적용하여 특징을 선택 한 후 예측 알고리즘을 이용하여 유전자 예측을 하였다. 예측한 결과의 값은 실제의 데이터 값과 예측 값의 차이를 계산한 평균제곱오차(Mean Square Error, MSE)로 나타냈다. MSE의 계산 식은 수식2와 같다[14].

$$E_i = \frac{1}{N}\sum_{i=1}^{N}(y_i^{predicted} - y_i^{measured})^2 \quad (2)$$

본 논문에서는 NFRN[5]과 EK[15] 두 개의 알고리즘과 비교한 결과 MSE의 평균이 NFRN는 0.12, EK알고리즘은 0.22였으나, 가중퍼지소속함수기반 신경망을 이용한 결과는 0.003으로 향상됨을 보여주고 있다. <Table 1>에서는 NFRN알고리즘과 EK 알고리즘, 가중퍼지소속함수기반 신경망을 이용한 알고리즘의 각 8개의 유전자의 MSE 값을 비교하여 보여주고 있다.

표 1. SOS데이터의 표준평균오차비교
Table 1. Compare of MEAN ERROR on the SOS data

model	uvrD	lexA	umuD	recA	uvrA	uvrY	ruvA	polB	avg (%)
NERN	0.17	0.08	0.09	0.10	0.09	0.16	0.20	0.08	0.12
EK	0.20	0.10	0.21	0.12	0.14	0.45	0.22	0.31	0.22
NEWFM	**0.003**	**0.007**	**0.008**	**0.0009**	**0.00008**	**0.12**	**0.0008**	**0.00019**	**0.003**

5. 결 론

E.coli의 8개 유전자를 예측하기 위하여 가중퍼지소속함수기반 신경망으로 특징선택 한 후 Takagi-Sugeno 기법으로 역퍼지화 한 값을 이용하였다. 각 샘플(time point)의 값은 Takagi-Sugeno 평균의 비율을 계산하여 현재의 time point를 이전의 time point가 예측하였다. 결과 값은 평균제곱오차를 계산한 결과 NERN알고리즘은 8개의 유전자에 대해 평균제곱오차 값이 0.12%, EK알고리즘 0.22%, 가중퍼지소속함수기반 신경망을 이용하여 특징선택을 사용한 방법은 0.003%로 향상됨을 보여주고 있다.

References

[1] G. D. Hong, *Cognitive load while learning to use a computer program. Applied cognitive psychology*, Vol. 2, No. 9, pp. 151-170, 1999.

[2] T.Akutsu, S.Miyano, and S. Kuhara, *Identification of genetic networks from a small number of gene expression patterns under the Boolean network model*, Pacific Symposium on Biocomputing pp. 17-28, 1999.

[3] M. Chaves, E. D. Sontag, and R, Albert, *Methods of robustness analysis for Boolean models of gene control networks,* IEEE Proceeding, Vol. 153, No. 4, pp. 154-167, 2006.

[4] S. Y. Kim, S.Imoto, and S. Miyano, *Inferring gene networks from time series microarray data using dynamic Bayesian network,* Briefings in Bioinformatics, Vol. 4, No. 3, pp. 228-235, 2003.

[5] I. A. Maraziotis, A. Dragonir, and A. Bezerianos, *Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks,* IET Syst. Biol., pp. 41-50.

[6] Joon Shik Lim, *Extracting minimized feature input and fuzzy rules using a fuzzy neural network and non-oferlay area distrubution measurement method,* Vol. 15, No. 5, pp. 599-604. 2005.

[7] Sang-Hong Lee, and Joon S. Lim, *Forcasting KOSPI based on a neural network with weighted fuzzy membership function,* Expert System with Application, Vol. 38, Issue 4, pp. 4259-4263, 2011.

[8] Abdulhamit Subasi, *EEG signal classification using wavelet feature extraction and a mixture of expert model:*, Expert Systems with Application, Vol. 32, Issue 4, pp. 1084-1093, 2007.

[9] Hee Jin Yoon, *Algorithm for prediction of gene-gene interaction and reconstruction network using fuzzy neural network,* gachon university, 2015.

[10] Husmeire, D., *Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks,* Bioinformatics, 19:2271-2282, 2003.

[11] Friedman, N., Linial, M., Nachman, I., and Pe'er D., *Using bayesian networks to analysis expression data*, J.Comp. Biol. 7, pp. 601-620, 2000.

[12] Lian En Chai, and Mohd Saberi Mohamad, *Inferring E.cloi SOS response pathway from gene expression data using IST-DBN with time lag estimation*, Biomedical Infrastructure SCI 477, pp. 5-14, 2013.

[13] Takagi T, and Sugeno M, *Fuzzy identification of systems and its applications to modeling and control*, IEEE Trans Syst Man Cybem 15:116-132. 1985.

[14] Roozbeh Manshaei, and Pooya Sobhe Bidari, *Hybrid-controlled neurofuzzy networks analysis resulting in genetic regulatory networks reconstruction*, ISRN Bioinformatics, Vol. 2012.

[15] I. A. Marziotis, A. Dragomir, and A. Bezerianos, *Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks*, IET Syst Biol, Vol. 1. No. 1, pp. 41-50. 2007.

유전자 조절 네트워크 재구성을 위한 시계열 발현 E.coli 예측

윤희진

장안대학교 인터넷정보통신과

요 약

유전자와 유전자의 관계를 나타내는 것을 유전자 조절 네트워크라고 한다. 유전자와 유전자의 관계에서 하나의 타겟유전자는 조절유전자들에 의해 영향을 받는다. 유전자 조절 네트워크는 조절유전자가 타겟유전자에 대해 발현을 활성화하는 것과 억제하는 것으로 구분되어진다. 본 논문에서는 유전자 조절 네트워크를 재구성하기 위해 시계열 발현 E.coli 데이터를 예측하였다. 실험 데이터로 사용된 E.coli는 8개의 유전자로 이루어졌으며, 50 time point를 갖고 있다. 하나의 유전자를 예측하기 위하여 나머지 7개의 유전자를 조절 유전자로 사용하였다. 예측을 위해 가중퍼지 소속함수기반 신경망을 이용하였다. 시계열 특성을 갖는 E.coli 데이터는 Wavelet을 유전자 조절 네트워크를 위해 예측하는데 특징으로 사용하였다. 가중퍼지소속함수기반 신경망에 의해 특징들은 훈련되어지고 가중퍼지소속함수 경계합에 의해 최소 특징을 선택하였다. 선택된 특징들은 데이터 값을 Takagi-Sugeno 값으로 역퍼지화하여, 조절유전자들의 이전시간의 발현 양으로 타겟유전자의 현재시간의 발현 양을 예측하였다. 예측한 결과 값은 평균제곱오차(MSE)로 나타냈다. MSE를 비교한 결과 NFRN알고리즘은 0.12% EK알고리즘 0.22% 가중퍼지소속함수기반 신경망을 이용한 결과 0.003%로 예측 결과가 향상됨을 보여주고 있다.

감사의 글

본 논문은 장안대학교의 2015학년도 학술연구비를 지원 받음.

Hee Jin Yoon received the master's degree in the Department of Computer Engineering from Dongkuk University in 2001. She received doctor's degree in the Department of Computer Engineering from Gachon University in 2015. She has been a Professor in the Department of Internet Communication at Jangan University since 2013. Her current research interest include Artificial Intelligence.

E-mail address: hjyoon@jangan.ac.kr