Indian Journal of Science and Technology, *Vol 8(26), DOI: 10.17485/ijst/2015/v8i26/80728, October 2015*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Prediction of Time Series Microarray Data using Neurofuzzy Networks

Hee J. Yoon[1], Bo H. Wang[2], Joon S. Lim[2*]

[1]IT College, Jangan University, Whasung - 445756, Gyeonggi-do, South Korea; hjyoon@jangan.ac.kr
[2]IT College, Gachon University, Seongnam - 461701, Gyeonggi-do, South Korea; bhwang99@hanmail.net, jslim@gachon.ac.kr

Abstract

There have been many studies recently that predict the interactions between genes and reconstruct the gene control network. In this paper, we propose the approach to predict the expression values between the genes of the yeast cell using a neural network based on weighted fuzzy membership function. This neuro fuzzy system makes the exact prediction possible through choosing best rules automatically. Features extracted from original data are used for learning. We extract the five features and they take into account the characteristics of time series by using wavelet transform, Current Position (CP) and time point. The best features to be good for prediction are selected through the Bounded Sum Weight of the weighted fuzzy membership function. The selected features are defuzzified through the Takagi-Sugeno method to calculate the prediction values of original gene expression data. We evaluate mean square error to indicate prediction accuracy of the proposed approach and then compare to the existing algorithm RNN using the neural network. The proposed method outperformed RNN.

Keywords: Feature, GRN, Neurofuzzy, NEWFM, Prediction

1. Introduction

The network that shows the interactions between genes is called the Gene Regulatory Network (GRN)[5]. GRN is an important topic in biology because it can help to find the causal relationships of genes[1]. Genes can be classified into the activator and the repressor of the regulators in the relationships. The activator is the regulator helping the expression of target genes. The regulator can also serve as a repressor if it restrains the expression of the target gene. As one of the means for measuring the accuracy of GRN including activator and repressor, it plays an important role that the expression value of any gene is predicted using several other genes.

In this paper, we propose the approach to predict the gene expression value by novel neuro fuzzy system with the weighted fuzzy membership functions. There are several researches for predicting using neuro fuzzy system as in hybrid method[13] and recurrent method[12]. They produce many rule combinations in order to improve prediction accuracy. If genes are increased, the number of rule combination will be increased exponentially. However, as the proposed approach use the weighted fuzzy membership function which the three membership functions are integrated in one, there is no rule combination and the rules are increased linearly with the number of genes.

Prediction experiment is done with yeast cell data. Yeast cell data consist of many time point values. In time series microarray data, one gene influences others over time. So the effect of data over time needs to be reflected in experiment. For this, we extract the features including changes of data at time series. Five features are extracted using wavelet transform, CP and time point. These features include one time and two times prior data, as well as current time data and are used for learning in neuro fuzzy system with the weighted fuzzy membership function (NEWFM)[11]. As current time data and prior data are used in experiment at the same time, the trend of data over time can be reflected in learning. Some of extracted features are selected through the bounded sum

*Author for correspondence

weight of the weighted fuzzy membership function for the NEWFM after each learning to improve prediction accuracy. Minimum features mean that the learning performance is improved and the prediction of the target gene is done the best.

After learning using the finally selected features, Takage-Sugeno values as results are defuzzified through the Takagi-Sugeno method and is used to predict the expression value of target gene. We calculate MSE between original expression gene data and the predicted values and then compare them with MSE for the existing algorithm RNN. CDC 15 data set is used for comparison. The average MSE of RNN is 0.3326 while when NEWFM is used, it is 0.2845. We have 14.4% improvement in the proposed approach compared to the existing method.

In the remainder of this paper, we propose the prediction method chapter 2. In chapter 3, we demonstrate improvement of our approach in terms of MSE with experimental results and then make conclusion in last.

2. Material and Method

In this chapter, we explain the using data, normalization and feature extraction methods, neuro fuzzy system and prediction method. Normalization of original data is done for removing noise and transforming to suitable data for neuro fuzzy system. We don't use the original data but the features extracted from original normalized data in neuro fuzzy system. As the trend over time is reflected in features, more accurate prediction is possible. When predicting is doing, Takagi-Sugeno values are used[10].

2.1 Experiment Data

For experiment data, we use the yeast cell time series microarray data by Spellman et al[7]. The 12 genes that are explained in the Yeast Proteome data base were used. Yeast cell time series microarray dataset has four datasets that are cdc15, cdc28, alpha and Elu, with each dataset having time points of 24, 17, 18 and 14, respectively. For the training data used for the prediction, cdc15 dataset was used, while cdc28 and alpha were used as the test data set.

2.2 Normalization

Experimental data is normalized to a scale between 0.5 and 1 in order to remove the noise from original data and facilitate the fuzzy process[8]. We apply the modified sigmoid function as normalization method[8]. Standard sigmoid function is given by

$$\text{sigmoid function} = \frac{1}{1+e^{-t}} \tag{1}$$

In (1), we substitute t with formula (2).

$$\frac{o_{i,j} - \min(o_i)}{\max(o_i) - \min(o_i)} \tag{2}$$

where o_i means every time points values of gene i and $o_{i,j}$ is jth time point value of gene i. Normalized data $g_{i,j}$ is produced through formula (1) and (2) using $o_{i,j}$.

2.3 Feature Extraction

In this paper, the features are used for the prediction method. The features are extracted from normalized data of original expression values. Time series data has characteristics in which a time point value affects another time point of other gene or same gene over time. We extracted the features which such characteristics are reflected in. The five features are extracted and Table 1 shows them. t, t + 1, t + 2 are time point in Table 1.

Feature T is the normalized value of gene data itself in current time. In this experiment, a target gene is set with data of current time (t) and all remaining genes as regulator are also set with current time data. This feature is used to reflect the effect on the target gene of input gene of current time. Figure 1 depicts the effect on the current value of target gene of current input gene[9].

$T^*(T + 1)$ is the feature which multiply the current time point value and the value after one time point. The product of adjacent two time point values is intended to reflect prior time point effect in experiment.

$$a1 = \frac{g_t + g_{t+1}}{\sqrt{2}} \tag{3}$$

$$d1 = \frac{g_t - g_{t+1}}{\sqrt{2}} \tag{4}$$

Table 1. Genes and features

	g_1					.....	g_{12}				
	T	T*(T+1)	a1	d1	cp	.....	T	T*(T+1)	a1	d1	cp
t						.....					
t+1						.....					
t+2						.....					
.	.	.	.	.	.	.....	.	.	.	.	.

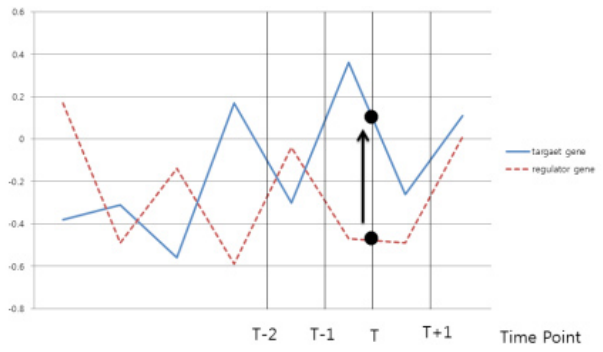

Figure 1. Current time feature.

The features *a1 and d1 are* given by formula (3) and (4), respectively. Formula (3) and (4) are derived from Harr wavelet transform[9]. Wavelet transform makes it possible to analyze time domain in data processing[9].

The feature *cp* is given by formula (5). In *cp*, the adjacent three time points are used for production of feature. After subtracting the average of values in prior two time points from current time value, the result is divided by the average of values in prior two time points. This feature aims to follow the trend of data during three time points.

$$cp = \frac{g_{t+2} - average(g_t + g_{t+1})}{average(g_t + g_{t+1})} \quad (5)$$

It is difficult to realize in which time a gene affects other gene. As it can be seen from the five characteristics, the trend of data over time is reflected in them. This trend enables us to predict accurate target gene data and further, to develop reverse engineering.

2.4 Neural Network based on Weighted Fuzzy Membership Function

Neuro fuzzy system used for learning in this paper is composed of weighted fuzzy membership functions. The bounded sum weights are calculated through weighted fuzzy membership functions. The bounded sum weight of the weighted fuzzy membership function is the synthesis of the bounded sum weights of three fuzzy membership functions of large, medium and small to create one fuzzy function. NEWFM is neuro fuzzy system using weighted fuzzy membership function. In the learning process, the bounded sum weight of each feature is calculated and used for selecting good features to predict[11]. The value of the bounded sum weights for the selected features is defuzzified using the Takagi-Sugeno method and this value is used as the prediction value[4].

2.5 Prediction Method

As in section 2.3, each gene has five features so total number of all gene's feature are 60. 60 features are divided by 6 groups randomly such that each group has 10 features. Table 2 shows configuration of one group.

In Table 2, class is decided by each time point expression value of target gene and obtained through formula (6)[9].

$$class(g_{i,t}) = \begin{cases} 1, if\ g_{i,t} < average(g_i) \\ 2, otherwise \end{cases} \quad (6)$$

Where $g_{i,t}$ is t time point value of target gene i and $average(g_i)$ is the average of all time points values of target gene i. For each target gene, 6 times learnings are executed using each group. After finishing each learning, we remove bad feature in one group or collect good features in different groups using bounded sum weight and then conduct learning again. After last learning of each target gene, prediction values are obtained by Takagi-Sugeno method[4]. This proceeds to two steps. First step, we calculate the average of all time point values of Takagi-Sugeno and then subtract each time point value of Takagi-Sugeno from the average. And the ratio of the difference to range of every time point Takagi-Sugeno values of any gene is calculated. Second step, the ratio is applied to the average of original expression data[9]. Equation (7) and (8) are about two step of prediction. $TS(g_{i,t})$ is Takagi-Sugeno value of t time point of gene i and $TS(g_i)$ means every time point values. In equation (8), α means amplitude control number.

$$ratio\,(TS(g_{i,t})) = \frac{TS(g_{i,t}) - average\,(TS(g_i))}{Max(TS(g_i)) - Min(TS(g_i))} \quad (7)$$

$$Prediction\ value = average\,(o_{i,t}) + ratio(TS(g_{i,t})) \times \alpha \quad (8)$$

3. Experimental Result

As seen in the Table 3, the methods by RNN and NEWFM using neural networks show a MSE (Mean Square Error)

Table 2. An example of input data group

	g_i.T	g_k.T*(T+1)	g_k.T	g_j.a1	g_s.cp	g_m.a1	class
t	*value11*	*value21*	*value31*	*value41*	*value51*	*value61*	1
t+1	*value12*	*value22*	*value32*	*value42*	*value52*	*value62*	2
t+2	*value13*	*value23*	*value33*	*value43*	*value53*	*value63*	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

Table 3. MSE of NEWFM and RNN

Gene	*NEWFM*	*RNN*[6]
SIC01	0.5176	0.4542
CLB05	0.1705	0.1721
CDC20	0.2450	0.5523
CLN03	0.0889	0.2493
SWI06	0.1643	0.2833
CLN01	0.3735	0.1874
CLN02	0.5779	0.5642
CLB06	0.3822	0.4005
CDC28	0.0407	0.1300
MBP01	0.5747	0.2742
CDC06	0.2281	0.3655
SWI04	0.0505	0.3604
Average	**0.2845**	**0.3326**

value. Equation for MSE is presented in (9) where t is the number of every time point and g^p is the predicted gene expression value and g^m is the measured value.

$$\text{mse} = \frac{1}{t}\sum_{t=1}^{t}(g^p - g^m)^2 \quad (9)$$

It was verified that the error rate of NEWFM is decreased comparing with RNN MSE. The proposed approach selects the best features in order to get the least mse through the bounded sum weight. So we can get the 14% more decreased average mse than the RNN's. Small MSE may be a measurement of the correct configuration GRN.

4. Conclusions

In this paper, we showed that the method of extracting features by using the traits of a time series when predicting genes using a weighted fuzzy membership function based on a neural network for yeast cell time series microarray data. This prediction can be used for reconstructing GRN. We first extracted five features from each gene and then learned repeatedly with good genes after feature selection. For prediction, Takagi-Sugeno values are obtained after finishing learning. Prediction is done with those values. The proposed approach is compared with RNN result. The result outperforms that of RNN.

5. Acknowledgement

This work (Grants No. S2230566) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2014.

6. References

1. Cho K, Choo S, Jung S, Kim J, Choi H, Kim J. Reverse engineering of gene regulatory networks. IET Syst Biol. 2007; 1(3):149–63.
2. Sugimoto N, Iba H. Inference of gene regulatory networks by means of dynamic differential Bayesian networks and nonparametric regression. Genome Inform. 2004; 15(2):121–30.
3. Kim S, Imoto S, Miyano S. Dynamic Bayesian network and nonparametric regression for nonlinear modelling of gene networks from time series gene expression data. Biosystems. 2004; 75(1-4):57–65.
4. Lim JS, Ryu TW, Kim HJ, Gupa S. Feature selection for specific antibody deficiency syndrome by neural network with weighted fuzzy membership. LNCS Springer Verlag. 2005; 3614:811–20.
5. Chen T, Filkov V, Skiena SS. Identifying gene regulatory networks from experimental data. Parallel Comput. 2001; 27(1-2):94–103.
6. Maraziotis IA, Dragonir A, Bezerianos A. Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks. System Biology IET. 2007; 1(1):41–50.
7. Spellman PT, Sherlock G, Zhang MQ, Iver VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Bio Cell. 1998; 9(12):3273–97.
8. Jun H, Claudio M. The influence of the sigmoid function parameters on the speed of backpropagation learning. From Natural to Artificial Neural Computation. 1995; 930:195–201.
9. Yoon HJ, Lim JS. Algorithm of prediction of gene-gene interaction and reconstruction of gene regulatory network using fuzzy neural network and microarray [Doctoral thesis]; 2015.
10. Takagi T, Sugeno M. Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybem. 1985; 15:116–32.
11. Lee S-H, Lim JS. Forecasting KOSPI based on a neural network with weighted fuzzy membership functions. Expert System with Applications. 2011; 38(4):4259–63.
12. Maraziotis IA, Dragomir A, Bezerianos A. Gene networks reconstruction and time-series prediction from microarray data using recurrent neural fuzzy networks. IET Syst Biol. 2007; 1(1):41–50.

13. Manshaei R, Bidari PS. Hybrid-controlled neurofuzzy networks analysis resulting in genetic regulatory networks reconstruction. ISRN Bioinformatics. 2012; 1–16.
14. Soinov LA, Krestyaninova MA, Brazma A. Towards reconstruction of gene networks from expression data by supervised learning. Genome Biology. 2003; 4(1).
15. Cheng C, Fu Y, Shen L, Gerstein M. Identification of yeast cell cycle regulated genes based on genomic features. BMC Syst Biol. 2013; 7(70):1–13.
16. Available from: http://www.genome.jp/kegg/
17. Du P, Gong J, Wurtele ES, Dickerson JA. Modeling gene expression networks using fuzzy logic. IEEE Transactions on Systems, Man, and Cybernetics B. 2005; 35(6):1351–9.
18. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520–5.