2003 International Symposium on Advanced Intelligent Systems
September 25-28, 2003, Jeju, Korea

Promoter classification using random generator-controlled generalized regression neural network

Kunho Kim, Byungwhan Kim*
Department of Electronic Engineering, Sejong University
98, Kunja-Dong, Kwangjin-Ku, Seoul, 143-747, Korea.
Email:kbwhan@sejong.ac.kr

Kyungnam Kim
Department of Molecular Biology, Sejong University
98, Kunja-Dong, Kwangjin-Ku,Seoul, 143-747, Korea.

Jin Han Hong, Sang Ho Park
DNA Chip Division, Macrogen
116, Shinmun-Ro 1 Ka, Chongro-Ku, Seoul, Korea.

Abstract **- A new classifier is constructed by using a generalized regression neural network (GRNN) in conjunction with a random generator (RG). The RG played a role of generating a number of sets of random spreads given a range for gaussian functions in the pattern layer. The range experimentally varied from 0.4 to 1.4. The DNA sequences consisted 4 types of promoters. The performance of classifier is examined in terms of total classification sensitivity (TCS), and individual classification sensitivity (ICS). For comparisons, another GRNN classifier was constructed and optimized in conventional way. Compared GRNN, the RG-GRNN demonstrated much improved TCS along with better ICS on average.**

1. Introduction

As a biometric, artificial neural network (ANN) has been extensively applied to map and identify specific biological functions in Deoxyribonucleic Acid (DNA) sequences [1-3]. Compared to other algorithms, ANN demonstrated superior functional mapping ability. This is mainly attributed to the ANN capability of high correlation and interpolation. Many different types of neural networks have been applied classifying various DNA sequences. Among neural networks, the generalized regression neural network [4] is increasingly expected due to its simple training algorithm and optimization procedure. Despite its potential usefulness, the GRNN has rarely been applied to predict or classify DNA sequence little reported. The GRNN performance depends on one training factor called "spread" of gaussian function. Conventionally, the effect of the spread on GRNN predictive performance is optimized by experimentally adjusting it. Most critical problem is that all gaussian functions in the pattern layer consist of the same one single spread. By adopting multi-spreads, it is expected that the GRNN predictive ability could be improved.

In this study, a technique to construct a GRNN classifier of multi-valued spreads is presented. This is accomplished by means of a random generator (RG). For convenience, the RG-based GRNN is referred to as RG-GRNN. The performance of GRNN is evaluated in terms of the classification sensitivity. Each measure is investigated for all or individual promoter data set. The classification sensitivity is more detailed with respect to the threshold. The RG-GRNN is also compared to conventional GRNN.

2. DNA Promoter Data

The DNA data evaluated consist of 4 types of promoter, including *Oriza Sativa* (OS), *Arabidopsis Thaliana* (AT), *Escherichia Coli* (EC), and *Zymomonas Mobils* (ZM). The first two promoters, OS and AT, can be classified into an

eukaryotic promoter. The other EC and ZM belong to prokaryotic promoter. Promoter sequences for AT were obtained by comparing full-length cDNAs [5] with a genomic DNA [6]. Since DNA sequences upstream of the cDNAs contain the promoter activity, approximately 1-kb genomic DNA regions upstream of the translation start site (ATG codon) were selected in constructing the database. The OS promoter sequences were collected in the similar way using the rice database [7]. Meanwhile, the whole genome sequences of two bacterial species, the EC [8] and ZM were obtained from NCBI with accession number U00096 and in-house database of Macrogen, respectively. The open reading frames (ORFs) from ZM were derived from the prediction by using a program 'Glimmer V2.0' [9] and analyzed further with a BlastX [10] program with non-redundant protein database of NCBI. For the two sets of genomic data, a number of promoter sequence were collected by searching promoters, and each sequence consisted of 500 bases upstream and 100 bases downstream from the cordon start site. The 600 bases per each ORF were thus used for promoter prediction.

The training data consist of 115 sets of promoter sequences. More specifically, the data is composed of 20 OS, 25 AT, 35 ET, and 35 ZM. The test data for evaluating model appropriateness are composed of 58 sets of promoters, 13 OS, 15 AT, 15 ET, and 15 ZM. Each sequence pattern consisted of 146 base pairs.

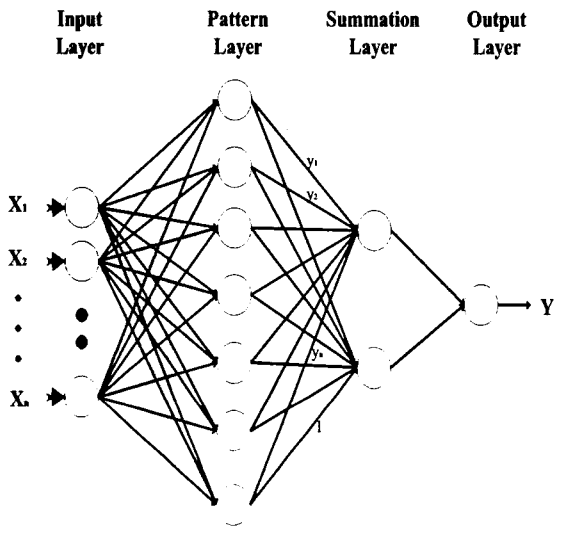

Figure 1: Schematic of generalized regression neural network

3. Generalized Regression Neural Network

A schematic of GRNN is depicted in Fig. 1. As shown in Fig. 1, the GRNN consists of four layers, including the input layer, pattern layer, summation layer, and output layer. Each input unit in the first layer corresponds to individual process parameter. The first layer is fully connected to the second, pattern layer, where each unit represents a training pattern and its output is a measure of the distance of the input from the stored patterns. Each pattern layer unit is connected to the two neurons in the summation layer: S-summation neuron and D-summation neuron. The S-summation neuron computes the sum of the weighted outputs of the pattern layer while the D-summation neuron calculates the unweighted outputs of the pattern neurons.

The connection weight between the ith neuron in the pattern layer and the S-summation neuron is y_i, the target output value corresponding to the ith input pattern. For D-summation neuron, the connection weight is unity. The output layer merely divides the output of each S-summation neuron by that of each D-summation neuron, yielding the predicted value to an unknown input vector x as

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

where n indicates the number of training patterns and the D function in (1) is defined as

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

where p indicates the number of elements of an input vector. The x_j and x_{i_j} represent the jth element of x and x_i, respectively. The ζ is generally referred to as the spread, whose optimal value is conventionally determined by adjusting it within certain experimental range.

4. Results

The performance of classifier is evaluated in terms of the classification sensitivity. The classification sensitivity is defined as the total number of the test sequence patterns correctly classified into their respective classes. It is evaluated as a function of the threshold expressed as

$$\left|d_{ij} - out_{ij}\right| \langle \text{ Threshold} \quad (3)$$

where, d_{ij} and out_{ij} represent the desired and calculated

outputs of the ith output neuron for the jth test pattern. The classification sensitivity is measured for all and individual data sets, each called total classification sensitivity (TCS) and individual classification sensitivity (ICS), respectively. Meanwhile, the threshold varied from 0.6 to 0.9 with an increment of 0.1.

Table I: ICS of GRNN with respect to the threshold

Threshold	OS	AT	EC	ZM	TCS
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

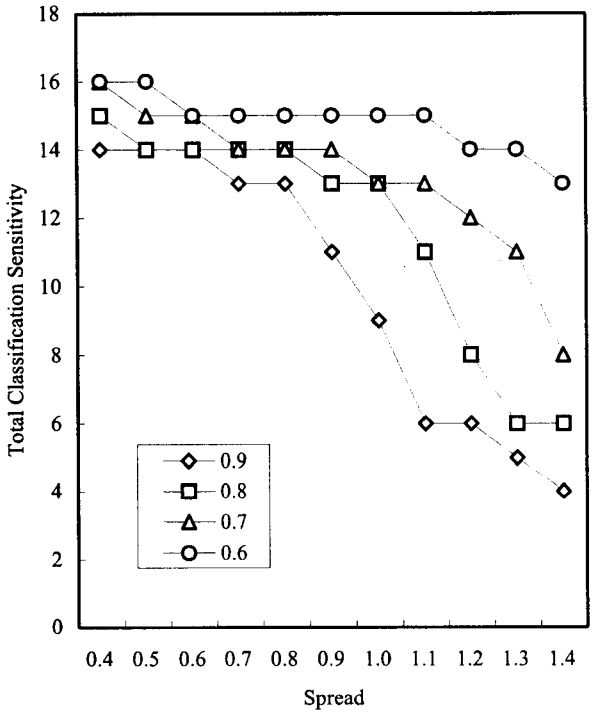

Figure 2: Total classification sensitivity of GRNN as a function of spread.

4.1. Conventional GRNN

First, the performance of conventional GRNN is investigated. The spread varied from 0.4 to 1.4 by 0.1. For each spread, GRNN classifier was constructed. The TCS measured by (3) are displayed in Fig. 2 as a function of the threshold. As depicted in Fig. 2, the TCS decreases with increasing the spread. The highest TCS is commonly obtained at 0.4 for all promoter data. The TCS of the classifier determined at 0.4 is detailed in terms of ICS. This was conducted as a function of the threshold and results are contained in Table I. Compared to other promoters, the GRNN is incapable of classifying the AT.

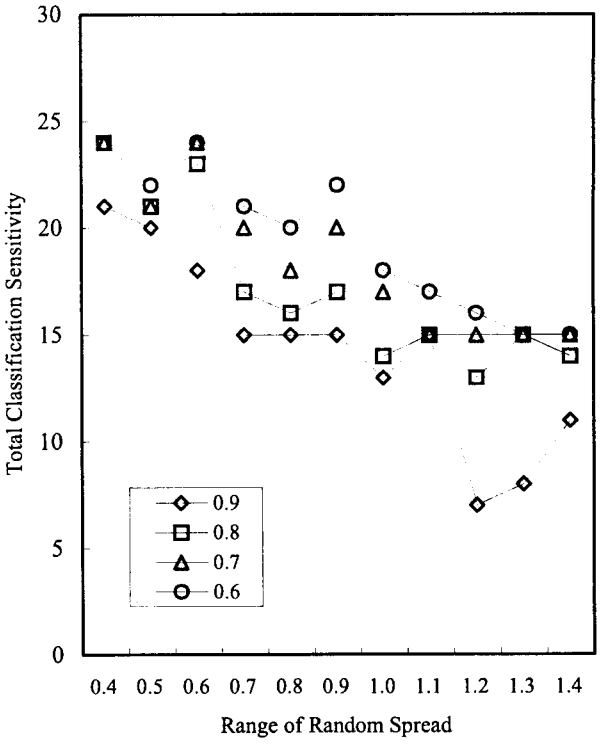

Figure 3: Total classification sensitivity of RG-GRNN as a function of range of random spread.

4.2 RG-GRNN

Using the RG, GRNN classifiers are constructed and compared to conventional GRNN. The RG was used to generate 200 sets of random spreads for the D defined in (2). The experimental range of the random spread is the same as the one employed earlier. Among 200 predictive models generated at a given range and threshold, one model with the highest TCS was selected. The TCSs determined are plotted in Fig. 3 as a function of the range. As depicted in Fig. 3, the TCS generally decreases with increasing the range, but behaves inconsistently with the range. Over the entire ranges, one optimal classifier with the highest TCS is obtained at 0.4 irrespective of the threshold. In Table II, the TCS is detailed in terms of the ICS with respect to the threshold. Compared to Table I, it is noticeable that the RG-GRNN yields a significantly improved TCS over the GRNN. This is illustrated even in the ICS for the AT, which was not possible by the GRNN as noticed earlier. The improvement is also demonstrated for the EC and ZM but the OS. Consequently, the RG-GRNN demonstrated much improved TCS along with better ICS on average. These clearly indicate that the proposed RG-GRNN is an effective way to construct a

classifier of large volume of DNA sequence data.

Table II: ICS of RG-GRNN with respect to the threshold

Threshold	OS	AT	EC	ZM	TCS
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

5. Conclusions

Using the RG, a GRNN classifier was constructed and applied to classify DNA promoter sequences. The RG was used to generate a number of sets of random spreads for the gaussian functions in the pattern layer. The RG-GRNN was compared to conventional GRNN in terms of the classification sensitivity. Comparisons revealed that the RG-GRNN was much better than GRNN in classifying all or individual promoters. Particularly, the improvement was significant in the total classification sensitivity. The proposed classifier is very simple to implement and optimize. By the demonstrated high classification capability, the RG-GRNN is expected widely used for predicting or classifying large volume of other bio-medical data.

Acknowledgements

This work was supported by Korea Health Industry Development Institute from 2000 IMT Fund.

References

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

[5] http://signal.salk.edu/cgi-bin/tdnaexpress.

[6] http://arabidopsis.org.

[7] http://www.ncbi.nlm.nih.gov.

[8] F. R. Blattner, G. III Plunket, C. A. Bloc, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao, "The complete genome sequence of Escherichia coli K-12," *Science*, vol. 277, pp. 1453-1474, 1997.

[9] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Res.*, vol. 27, pp. 4636-4641, 1999.

[10] W. Gish and D. J. States, "Identification of protein coding regions by database similarity search," *Nature Genetic*, vol. 3, pp. 266-272, 1993.