Pareto RBF network ensemble using multi-objective evolutionary computation

Nobuhiko Kondo, Toshiharu Hatanaka and Katsuji Uosaki

Department of Information and Physical Sciences,
Graduate School of Information Science and Technology,
Osaka University, Suita, 565-0871, Japan
Tel: +81-6-6879-7834; Fax: +81-6-6879-7836;
Email:{nobuhiko,hatanaka,uosaki}@ist.osaka-u.ac.jp

Abstract: In this paper, evolutionary multi-objective selection method of RBF networks structure is considered. The candidates of RBF network structure are encoded into the chromosomes in GAs. Then, they evolve toward Pareto-optimal front defined by several objective functions concerning with model accuracy and model complexity. An ensemble network constructed by such Pareto-optimal models is also considered in this paper. Some numerical simulation results indicate that the ensemble network is much robust for the case of existence of outliers or lack of data, than one selected in the sense of information criteria.

Keywords: RBF network, Evolutionary Multi-Objective Optimization, Genetic Algorithm, Nonlinear System Modeling, Ensemble Learning

1. Introduction

Mathematical models of actual systems are fundamentally important in a lot of engineering problems such as development of new system, existing system analysis, control system design, fault detection and diagnosis, signal processing, time series prediction and so on. In many cases, actual existing systems have some kind of nonlinear properties, so the system models have to represent such properties appropriately. From this viewpoint nonlinear system modeling has been studying and plenty of modeling algorithms have been developed in this two decades. Most of these approaches give a good model under some criterion based on the prior knowledge. However, there are usually several demands to a system model, the system model optimized under the specific criterion is not the optimal model in means of the other criterion. For example, it required that the model should be easy to handle and well explainable for the modeling data set contaminated by observation noise, but these properties are mutually exclusive. On the other hand, since multi-objective optimization is receiving much attentions in the field of system optimization, the evolutionary computation is much being studied as an efficient technique to providing the Pareto-optimum solutions [1] [3]. From this point of view, an application of multi-objective evolutionary computing to nonlinear system identification is proposed [5] [13]. These approaches deal with polynomial dynamic system model and give the optimal model set concerning model accuracy and complexity.

Artificial neural networks have wide variety of applications due to their powerful nonlinear mapping ability, so nonlinear system modeling using these has been receiving much attentions in the last two decades motivated by its applicability. The primary importance in applying neural network to nonlinear system modeling is to select its structure suitably. Then, some approaches to determine the neural network structure have been proposed. However, a general method of the structure determination has not established, because the optimum structure depends on a class of objective system, application area, learning algorithm and so on. So the network structure is generally determined by trial and error or a heuristic method. Moreover, there generally exists a tradeoff between the model accuracy and the model complexity in the system modeling [11], so it makes the structure determination problems more difficult. From this viewpoint, we have considered multi-objective optimization based modeling using evolutionary algorithms.

In this paper, we deal with the static nonlinear system modeling using RBF (Radial Basis Function) network, which is a kind of artificial neural network. RBF network has in their hidden layer a number of basis function which respond locally in input space. The network output is the linear sum of the basis function values. If the parameters of RBF networks, i.e. the number of basis functions and the widths and centers of each basis function, are determined, output layer weights can be calculated with the training data. This parameter setting affects the quality of function approximation. Therefore we consider the structure determination problem of RBF networks as a multi-objective optimization problem that concerns with the model accuracy, the model complexity and the output layers' weights. Then a method of obtaining the candidates of model as a Pareto-optimal set based on evolutionary algorithms is proposed.

The designers will be able to select one model from the Pareto-optimal set obtained by the proposed method according to their use or some criteria. On the other hand, by introducing the concept of the ensemble learning, one model can be obtained by constructing the ensemble network of the Pareto set. It is expected that the ensemble network is much robust than one of the Pareto models. Then by numerical simulations, we compare the ensemble network with the models which are selected from the Pareto set by AIC(Akaike information criterion) and BIC(Bayesian information criterion).

In the section 2, the concept of multi-objective optimization

by GA is introduced. In the section 3, an outline of RBF networks is described and the proposed method is introduced. The section 4 explains the ensemble of Pareto-optimal RBF networks. Some numerical study results are shown in the section 5 and concluding remarks are given at the last section.

2. Multi-objective Optimization by GA

2.1. Genetic Algorithm

GA (Genetic algorithm) is an algorithm of search or optimization. It was invented based on genetics and evolution. Initially, the initial population of individuals which have a binary digit string as the "chromosome" is generated at random. Each bit of chromosome is called "gene" [1]. The "fitness", which is a measure of adaptation to environment, is calculated for each individual. Then, "selection" operation leaving individuals to next generation is performed based on fitness value, and then "crossover" and "mutation" are performed on the selected individuals to generate new population by transforming chromosomes into offspring's ones. This procedure is continued until the end condition is satisfied. This algorithm is conforming to the mechanism of evolution, in which the genetic information changes for every generation and the individuals which adapt to environment better survive preferentially.

GA has several advantage. First, since GA is a stochastic multi-point search, it can search parallel. Second, GA requires only fitness value. Third, GA is expected to escape from local optimum by genetic operator.

2.2. Multi-objective Optimization

In the multi-objective optimization problems, there generally exists tradeoff among the objective functions. And so two concept, "domination" and the "Pareto-optimum", are considered. First, $\boldsymbol{x}_1$ is said to "dominate" $\boldsymbol{x}_2$ if

$$\forall i = 1, 2, \ldots, n \quad f_i(\boldsymbol{x}_1) \leq f_i(\boldsymbol{x}_2)$$

and

$$\exists j = 1, 2, \ldots, n \quad f_j(\boldsymbol{x}_1) < f_j(\boldsymbol{x}_2)$$

And $\boldsymbol{x}_0$ which is not dominated by any other $\boldsymbol{x}$ is called the "Pareto-optimal solution" [2] [3]. Pareto-optimal solution is considered to be the best solution comprehensively. And generally many Pareto-optimal solutions exist simultaneously. Considering tradeoff among the objective functions, on multi-objective optimization problems it is important to obtain a Pareto-optimal solution set.

2.3. Multi-objective GA based on *rank*

A parameter *rank* is introduced in order to apply the concepts of domination and Pareto-optimum to GA. Though there are some ranking methods, this study adopts Fonseca's ranking method [3]. According to Fonseca's ranking method, a *rank* of an individual $\boldsymbol{x}_i$ on a generation t is:

$$rank(\boldsymbol{x}_i, t) = 1 + p_i^{(t)}$$

where p_i is the total number of individuals which dominate $\boldsymbol{x}_i$. By calculating this *rank* for each individual and selecting based on it, a population can evolve toward a Pareto-optimal solution set. Since GA is a multi-point search algorithm, GA is expected to find a Pareto-optimal set in a single simulation run.

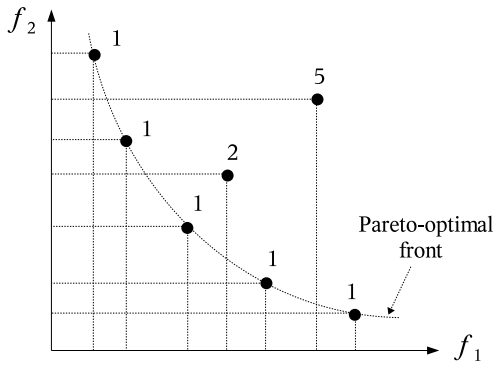

Fig. 1. Fonseca's ranking method

3. Construction of Pareto RBF Networks

3.1. RBF Network

RBF (Radial Basis Function) network is constructed of three layers as shown in Fig.2 and has basis functions which respond locally in input space.

Basis function $\phi_j(\boldsymbol{x})$ in this study is defined by Gaussian function,

$$\phi_j(\boldsymbol{x}) = \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{c}_j)^T(\boldsymbol{x} - \boldsymbol{c}_j)}{2\sigma_j^2}\right) \tag{1}$$

Here, $\boldsymbol{x}$ is input variable, $\boldsymbol{c}_j$ is center vector, and σ_j^2 is a parameter which decides function width. Using this $\phi_j(\boldsymbol{x})$, RBF network is constructed as:

$$u(\boldsymbol{x}) = w_0 + \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) \tag{2}$$

Here, m is the number of hidden units, i.e., the basis functions, and w_j are the output layer weights. RBF network will be determined if the parameters $m, \boldsymbol{c}_j, \sigma_j$, and w_j are estimated based on the data observed from the system. In this study, these parameters are estimated by two GAs. The parameters σ_j are assumed to be constant value for simplicity.

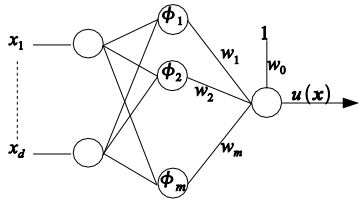

Fig. 2. RBF network

3.2. Genetic Representation

In this study, we apply MOGA (Multi-Objective Genetic Algorithm) to determine the both of the number of basis functions and the centers of them. The candidate of the center of basis function is assumed to be the position of the training data points. The chromosomes of MOGA population indicate that the data points are employed as centers of basis functions i.e. "1" represents that a basis function is located at the corresponding training data point, as shown in Fig.3. By this setting, the length of the chromosome becomes equal to the number of training data, the number of "1" gene in the chromosome indicates the number of basis functions and the locus of the "1" shows the center position of the basis functions. In the proposed method, the weight parameters are estimated by real-coded GA.

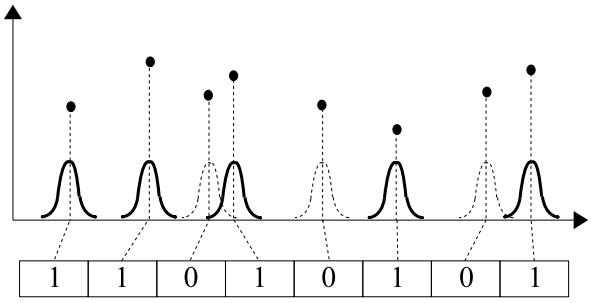

Fig. 3. Chromosome representation

3.3. Evaluation

It is generally demanded that the mathematical models not only can explain the relationship between input and output enough but also is simple in order to have the generalization ability. Then in this study three evaluation criteria are set for the evaluation of MOGA which determines the network architecture. The first fitness is the number of basis function. This fitness indicates the complexity of the model. The second fitness is logMSE. This fitness indicates the extent of a fit of the model to the training data. MSE(Mean Squared Error) is defined as :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\{y_i - \hat{y_i}\}^2 \tag{3}$$

Here, y_i is the observed output, $\hat{y_i}$ is the model output. The third fitness is the sum of the absolute value of weights. These three evaluation criteria are to be minimized. MSE is used for evaluation in real-coded GA which estimates the weight parameters.

3.4. Construction of Pareto RBF Networks

The Pareto-optimal RBF network construction algorithm is consist of MOGA including real-coded GA. As MOGA we adopt NSGA-II which is one of MOGAs and is known to have the capability to maintain diversity[2]. After estimating all the parameters of the network, *rank* is assigned for each individual by the concept of multi-objective optimization problem, in which three fitnesses are to be minimized. Then Pareto-optimal individuals will be obtained in accordance with NSGA-II algorithm. In NSGA-II, we apply genetic operations which are the uniform crossover and the bit reversal mutation.

In real-coded GA, about genetic operation, UNDX (Unimodal Normal Distribution Crossover) [12] is applied in the proposed method. UNDX generates two offsprings by normal random numbers which is determined by three parents, as shown in Figure 4. Basically offsprings are generated by normal distribution around segment connecting two parents. The third parent is used to determine the standard deviation of normal distribution.

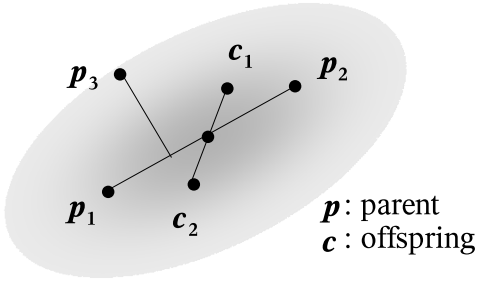

Fig. 4. UNDX

MGG[14] is adopted as the generation alternation model of real-coded GA used in the proposed method. MGG is said to have an ability to preserve the diversity of population. MGG procedure is as follows.

1. Plurality of real number vector is generated at random as the initial population.
2. Two parents are selected at random from population.
3. $2n_c$ offsprings are generated by applying UNDX to two parents n_c times. Here the third parent which determines the standard deviation of normal distribution is selected from population.
4. Fitness values of each offspring are calculated, then two individuals are selected from the set which is composed of two parents and all offsprings, then two parents are replaced by the selected two individual. The individuals selected here are elite and the individual selected by roulette selection in which the elite was pruned.
5. Continue 2 4 until the end condition is met.

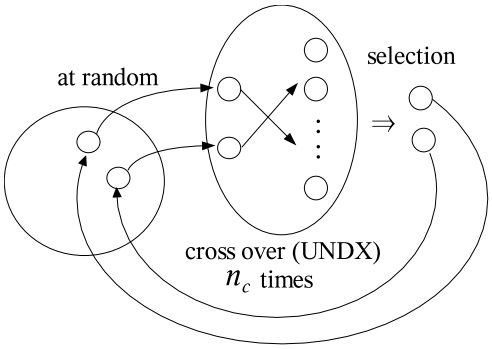

Fig. 5. MGG

The procedure of the proposed method is shown in Fig.6.

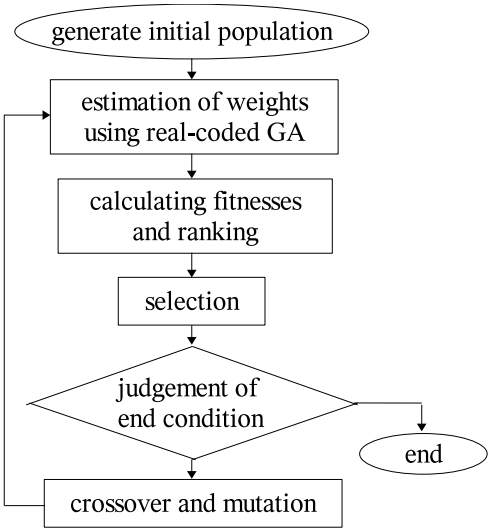

Fig. 6. Flowchart of the proposed method

4. Pareto RBF Network Ensemble

Various models based on three criterion can be obtained by proposed method, so the designers will be able to select one model flexibly. On the other hand, there are the demand to obtain one model with good generalization ability. For instance, model selection by information criteria has been studied.

Recently the ensemble learning is receiving much attentions in the field of machine learning. In the ensemble learning, a monolithic model is constructed by combining several models. While some learning methods to make models constructing ensemble have been proposed, in this study ensemble is constructed of Pareto-optimal models obtained by the proposed method.

Suppose that the number of Pareto models is m and the output of j-th network is $y_j(\boldsymbol{x})$, then the output of ensemble network $y^{EN}(\boldsymbol{x})$ is :

$$y^{EN}(\boldsymbol{x}) = \sum_{j=1}^{m} w_j y_j(\boldsymbol{x}) \tag{4}$$

Here, w_j is the weight on the output of j-th network. In this study, w_j is assumed to be $1/m$ about every j, for simplicity. In this study, difference of performance between the ensemble network and networks selected from Pareto-optimal set based on information criteria is considered from numerical simulation results. Following $AIC(Akaike\ Information\ Criterion)$ and $BIC(Bayesian\ Information\ Criterion)$ are used as information criteria.

$$AIC = n \log MSE + 2(m+1) \tag{5}$$

$$BIC = n \log MSE + (m+1) \log n \tag{6}$$

5. Numerical Simulation

5.1. Function approximation problem

In the numerical simulation, the nonlinear function approximation problem is considered.

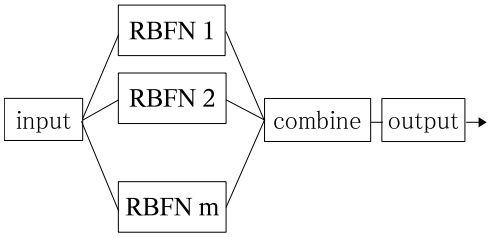

Fig. 7. RBF network ensemble

Let the true function be:

$$v(x) = 2x + 3\sin(5\pi x) + \sin(10\pi x). \tag{7}$$

Training data set is sampled by

$$\begin{array}{lcll} y_i & = & v(\boldsymbol{x}_i) + \varepsilon_i, & i = 1, 2, \ldots, n \\ \varepsilon_i & & N(0, \sigma^2), & \end{array} \tag{8}$$

where $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ are input-output data and n is the number of training data. x_i are sampled from uniform distribution over $[0, 1]$. The observation data are disturbed by normal white noise ε_i with mean 0, variance σ^2.

In order to investigate difference of performance between the ensemble network and networks selected from Pareto-optimal set based on information criteria, following numerical simulations have been implemented.

At first Pareto-optimal RBF networks are constructed using training data. About the parameters of NSGA-II, population size is 50, crossover rate is 0.7, mutation rate is 0.1 and generation size is 10. The RBF width parametersσ_j^2, $j = 1, 2, \cdots$ are set to 0.01. MGG is iterated 10,000 times with population size 30 and n_c is 30. Next, MSE for test data not used in training is calculated to investigate the generalization ability.

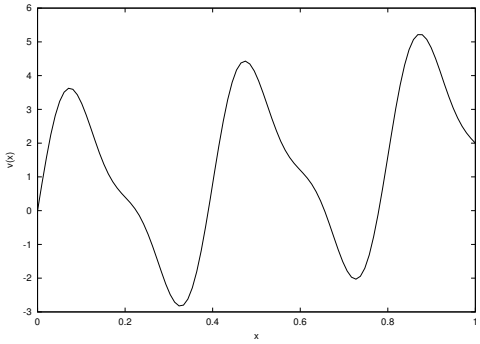

Fig. 8. True function of simulation 1

5.2. Simulation 1

Pareto-optimal RBF networks were constructed using training data with the observation noise $\varepsilon_i \quad N(0, 0.04)$. Then MSE for 50 test data not used in training was calculated for both the ensemble network and networks selected from Pareto-optimal set based on AIC and BIC. The number of training data was changed as 30, 40, 50 and 60. For each number of training data, data set was changed 5 times respectively. Results are shown in Table 1.

Table 1. Simulation 1 : The column of "data" indicates the number of training data, the column of "RBFN" indicates the number of obtained Pareto RBF networks. The column of "ensemble" indicates MSE of the ensemble network. The columns of "AIC" and "BIC" indicate MSE of the network which are selected by AIC and BIC, respectively.

data	RBFN	ensemble	AIC	BIC
30	50	0.191255	0.099932	0.088485
	48	0.171739	0.147592	0.147592
	50	0.435584	0.255772	0.255772
	36	0.630114	0.978955	0.652990
	50	0.174871	0.075501	0.075501
40	50	0.169304	0.102441	0.102441
	50	0.174922	0.060889	0.060889
	43	0.176029	0.049408	0.069194
	44	0.396571	0.045771	0.045771
	50	0.193280	0.047974	0.047974
50	43	0.162108	0.062830	0.062830
	46	0.217515	0.064232	0.114111
	42	0.186255	0.063561	0.074141
	41	0.202725	0.089901	0.089901
	50	0.090469	0.037580	0.037580
60	34	0.222999	0.096943	0.096943
	50	0.126545	0.056253	0.056253
	46	0.196389	0.101005	0.101005
	36	0.338770	0.108320	0.108320
	50	0.115663	0.077814	0.077814

5.3. Simulation 2

Next, the variance of the observation noise was assumed to become big with a certain probability. 90% of noise variance was 0.16 and 10% of it was 16. Then MSE for 50 test data not used in training was calculated for both the ensemble network and networks selected from Pareto-optimal set based on AIC and BIC. The number of training data was changed as 30, 40, 50 and 60. For each number of training data, data set was changed 5 times respectively. Results are shown in Table 2.

5.4. Simulation 3

Next, 80% of noise variance was 0.16 and 20% of it was 16. Then MSE for 50 test data not used in training was calculated for both the ensemble network and networks selected from Pareto-optimal set based on AIC and BIC. The number of training data was changed as 30, 40, 50 and 60. For each number of training data, data set was changed 5 times respectively. Results are shown in Table 3.

5.5. Simulation 4

Next, 70% of noise variance was 0.16 and 30% of it was 16. Then MSE for 50 test data not used in training was calculated for both the ensemble network and networks selected from Pareto-optimal set based on AIC and BIC. The number of training data was changed as 30, 40, 50 and 60. For each number of training data, data set was changed 5 times respectively. Results are shown in Table 4.

Table 2. Simulation 2

data	RBFN	ensemble	AIC	BIC
30	38	1.051167	0.957824	0.957824
	47	0.728419	0.348573	0.348573
	32	0.927173	2.525247	1.654921
	29	1.752871	1.738257	1.933985
	32	2.513687	2.191173	2.191173
40	17	0.778018	0.784632	1.036531
	46	0.507676	0.364317	0.610114
	16	0.978053	1.238106	1.238106
	22	2.217106	2.252523	2.427382
	19	0.758372	0.982137	0.982137
50	28	0.680729	0.821663	0.821663
	37	0.820927	0.602269	0.602269
	34	0.569077	0.771010	0.728347
	38	0.785553	0.874953	0.874953
	17	0.431460	0.830321	0.830321
60	16	0.957909	0.659941	0.659941
	15	0.814199	0.785704	0.785704
	24	0.894919	0.903971	0.903971
	34	0.320639	0.243167	0.243167
	8	0.734543	0.804033	0.804033

5.6. Discussion of the results

In simulation 1, the case of stationary noise with small variance was considered. There networks selected from Pareto-optimal set based on information criteria have smaller test data MSE than the ensemble network. However, at the fourth row of Table 1, MSEs for test data are significantly large. In this simulation run the ensemble network has the smallest MSE for test data. Training data set lacks within the range from 0.7 to 0.85 in this simulation run. The ensemble network is expected to have good generalization ability in such cases.

In simulation 2, 3 and 4, the case where there are outliers in training data is considered. There the ensemble network often have smaller test data MSE than networks selected from Pareto-optimal set based on information criteria. The ensemble network may ease the effect of bad networks such as networks overfitting the outliers.

Summing up the results, the ensemble network has higher generalization ability than networks selected based on AIC or BIC for the case of existence of outliers or lack of data. This results indicate that the ensemble network is robust, though more simulation and discussion are needed.

6. Conclusions

In this study, we have proposed a method of obtaining a Pareto-optimal RBF network set based on evolutionary algorithms. Then, an ensemble network constructed by such Pareto-optimal models is also considered. Numerical simulation results indicate that the ensemble network is much robust than networks selected based on information criteria. Reduction of the computational costs, improvement of the ensemble method, comparison to the conventional approaches, application to dynamic system identification and

Table 3. Simulation 3

data	RBFN	ensemble	AIC	BIC
30	35	1.022114	1.768824	1.768824
	32	3.090617	3.723074	3.723074
	30	1.372639	1.898516	1.898516
	36	4.202276	5.199522	5.199522
	28	16.542650	94.693893	94.693893
40	25	1.031609	1.284169	1.284169
	38	0.924128	1.045010	1.045010
	27	3.924914	2.424119	1.425429
	49	1.006428	1.434848	1.434848
	37	0.866881	0.956438	0.956438
50	13	1.228426	1.445913	1.445913
	21	0.838093	1.347125	1.347125
	32	1.411863	1.186526	1.186526
	33	1.168921	1.339410	1.733162
	11	0.709311	0.808769	0.808769
60	20	0.814219	1.081037	1.081037
	32	0.412720	0.445307	0.445307
	12	0.787953	0.778415	0.778415
	26	1.250772	0.974598	1.398607
	29	0.800697	0.943744	0.943744

Table 4. Simulation 2-4

data	RBFN	ensemble	AIC	BIC
30	21	3.612076	4.014570	4.014570
	36	7.883370	8.813035	8.813035
	30	2.545823	5.239426	5.239426
	32	0.718388	0.729001	0.729001
	47	1.321716	4.653079	4.653079
40	35	5.670076	3.722443	3.722443
	19	3.039101	3.793278	1.407294
	12	1.278787	1.710274	1.710274
	30	3.257849	2.846025	2.846025
	26	0.617000	0.725918	0.658233
50	13	0.953733	0.990628	0.990628
	9	2.517913	3.027015	3.027015
	22	1.781070	2.862607	2.862607
	13	1.577241	1.458167	0.942308
	26	3.334088	4.163925	4.163925
60	31	1.097832	1.201612	1.424725
	7	2.583948	2.422122	2.422122
	25	1.172458	1.095895	2.169411
	27	0.799774	0.723667	0.723667
	10	0.930999	0.810712	0.810712

so on are the future works.

References

[1] T. Bäck , *Evolutionary Computation*, Oxford Press, 1996.

[2] Kalyanmoy Deb, "Multi-objective Optimization using Evolutionary Algorithms," Chichester, John Wiley & Sons, 2001.

[3] C. M. Fonseca and P. J. Fleming, "Genetic algorithms for multiobjective optimization : Formulation, discussion and generalization," *Proceedings of the Fifth International Conference on Genetic Algorithms*, pp.416-423, 1993.

[4] T. Hatanaka, K. Uosaki and T. Kuroda, "Structure Selection of RBF Neural Network Using Information Criteria," *Proceedings of Fifth International Conference on Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies*, 2001.

[5] T. Hatanaka, K. Uosaki and Y. Hossaka, "Application of Evolutionary Multi-Objective Optimization Nonlinear System Identification," *Proceedings of the SICE Kansai Branch Symposium 2002*, pp.54-56, 2002. (in Japanese)

[6] Md. Monirul Islam, X. Yao and K. Murase, "A constructive algorithm for training cooperative neural network ensembles," *IEEE Transactions on Neural Networks*, 14(4):820-834, 2003.

[7] Yaochu Jin, Tatsuya Okabe and Bernhard Sendhoff, "Neural network regularization and ensembling using multi-objective evolutionary algorithms," *Proceedings of Congress on Evolutionary Computation*, pp.1-8, 2004

[8] N. Kondo, T. Hatanaka and K. Uosaki, "Pareto RBF Networks Based on Multiobjective Evolutionary Computation," *SICE Annual Conference in Sapporo*, pp.2177-2182, 2004.

[9] N. Kondo, T. Hatanaka and K. Uosaki, "System identification using evolutionary computation and Pareto ensemble," *SICE Symposium on Systems and Information 2004*, pp.101-106, 2004.

[10] L. Ljung, *System Identi cation: Theory for the User*, (2nd Edition), Prentice-Hall, 1999.

[11] Oliver Nelles, *Nonlinear System Identi cation*, Springer, 2001.

[12] I. Ono and S. Kobayashi, "A Real-coded Genetic Algorithm for Function Optimization Using Unimodal Normal Distribution Crossover," *Proc. 7th Int. Conf. on Genetic Algorithms*, pp.246-253, 1997.

[13] Katya Rodriguez-Vazquez, C.M. Fonseca and P.J. Flemming, "Multiobjective Genetic Programming : A Non-Linear System Identification Application," *Genetic Programming 97 Conference*, pp. 207-212, 1997.

[14] H. Sato, I. Ono and S. Kobayashi, "A New Generation Alternation Model of Genetic Algorithms and Its Assessment," *Journal of Japanese Society for Arti cial Intelligence*, Vol.12, No.5, pp.734-744, 1997.

[15] N. Ukai, T. Hatanaka, K. Uosaki and H. Kurakake, "Process Modeling Using Multi-Objective Genetic Programming," *SICE 1st Annual Conference on Control Systems*, pp.685-688, 2001. (in Japanese)

[16] K. Uosaki and T. Hatanaka, "Nonlinear System Identification based on Genetic Programming," *Journal of The Society of Instrument and Control Engineers*, Vol.39, No.5, pp.352-357, 2000. (in Japanese)

[17] Zi-Jiang Yang, "Identification of Nonlinear Systems," *Journal of The Society of Instrument and Control Engineers*, Vol.37, No.4, pp.249-255, 1998. (in Japanese)