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 studied and plenty of modeling algorithms have been developed in this two decades. Most of these approaches introduce the concept of the ensemble learning, one model can be obtained by constructing the ensemble network of several model candidates as a Pareto-optimal set based on 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 algorithm, 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, \( x_1 \) is said to “dominate” \( x_2 \) if

\[
\forall i = 1, 2, \ldots, n \quad f_i(x_1) \leq f_i(x_2)
\]

and

\[
\exists j = 1, 2, \ldots, n \quad f_j(x_1) < f_j(x_2)
\]

And \( x_0 \) which is not dominated by any other \( x \) 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 \( x_t \) on a generation \( t \) is:

\[
\text{rank}(x_t, t) = 1 + p_i^{(t)}
\]

where \( p_i \) is the total number of individuals which dominate \( x_t \). 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.

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 \( \phi_j(x) \) in this study is defined by Gaussian function,

\[
\phi_j(x) = \exp \left( -\frac{(x - c_j)^T(x - c_j)}{2\sigma_j^2} \right)
\]

Here, \( x \) is input variable, \( c_j \) is center vector, and \( \sigma_j^2 \) is a parameter which decides function width. Using this \( \phi_j(x) \), RBF network is constructed as:

\[
u(x) = w_0 + \sum_{j=1}^{m} w_j \phi_j(x)
\]

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 \( m, c_j, \sigma_j \), 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 \( \sigma_j \) are assumed to be constant value for simplicity.
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.

![Chromosome representation](image)

**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 log\(MSE\). This fitness indicates the extent of a fit of the model to the training data. \(<MSE>(Mean\ squared\ error)\) is defined as:

\[
MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  

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.

![UNDX](image)

**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.

![MGG](image)

**Fig. 5. MGG**

The procedure of the proposed method is shown in Fig.6.
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 $y_j(x)$, then the output of ensemble network $y_{EN}(x)$ is:

$$y_{EN}(x) = \sum_{j=1}^{m} w_j y_j(x)$$  \hspace{1cm} (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.

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

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

5. Numerical Simulation

5.1. Function approximation problem

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

Let the true function be:

$$v(x) = 2x + 3 \sin(5\pi x) + \sin(10\pi x).$$  \hspace{1cm} (7)

Training data set is sampled by

$$y_i = v(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where $x_i, y_i$ 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 $\varepsilon_i$, with mean 0, variance $\sigma^2$.

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 parameters $\sigma_j^2, j = 1, 2, \ldots$ 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.

5.2. Simulation 1

Pareto-optimal RBF networks were constructed using training data with the observation noise $\varepsilon_i \sim N(0, 0.04)$. 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.

<table>
<thead>
<tr>
<th>data</th>
<th>RBFN</th>
<th>ensemble</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>50</td>
<td>0.191255</td>
<td>0.099932</td>
<td>0.088485</td>
</tr>
<tr>
<td>48</td>
<td>0.171739</td>
<td>0.147592</td>
<td>0.147592</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.435584</td>
<td>0.255772</td>
<td>0.255772</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.603114</td>
<td>0.978955</td>
<td>0.652990</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.174871</td>
<td>0.075501</td>
<td>0.075501</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>50</td>
<td>0.169304</td>
<td>0.102441</td>
<td>0.102441</td>
</tr>
<tr>
<td>50</td>
<td>0.174922</td>
<td>0.060889</td>
<td>0.060889</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>0.176029</td>
<td>0.049408</td>
<td>0.069194</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>0.396571</td>
<td>0.045771</td>
<td>0.045771</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.193280</td>
<td>0.479794</td>
<td>0.479794</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>43</td>
<td>0.162108</td>
<td>0.062830</td>
<td>0.062830</td>
</tr>
<tr>
<td>46</td>
<td>0.217515</td>
<td>0.064323</td>
<td>0.114111</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>0.186355</td>
<td>0.063561</td>
<td>0.074141</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>0.207225</td>
<td>0.089901</td>
<td>0.089901</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.090469</td>
<td>0.037580</td>
<td>0.037580</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>34</td>
<td>0.222999</td>
<td>0.096943</td>
<td>0.096943</td>
</tr>
<tr>
<td>50</td>
<td>0.126545</td>
<td>0.056253</td>
<td>0.056253</td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>0.196389</td>
<td>0.101005</td>
<td>0.101005</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>0.343770</td>
<td>0.108320</td>
<td>0.108320</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.115663</td>
<td>0.077814</td>
<td>0.077814</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>data</th>
<th>RBFN</th>
<th>ensemble</th>
<th>AIC</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>38</td>
<td>1.051167</td>
<td>0.957824</td>
<td>0.957824</td>
</tr>
<tr>
<td>47</td>
<td>0.728419</td>
<td>0.348573</td>
<td>0.348573</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>0.927173</td>
<td>2.525247</td>
<td>1.654921</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1.752871</td>
<td>1.738257</td>
<td>1.933985</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>2.513687</td>
<td>2.191173</td>
<td>2.191173</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>17</td>
<td>0.778018</td>
<td>0.746432</td>
<td>1.036531</td>
</tr>
<tr>
<td>46</td>
<td>0.507676</td>
<td>0.364317</td>
<td>0.610114</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>0.978053</td>
<td>1.238106</td>
<td>1.238106</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2.217106</td>
<td>2.225233</td>
<td>2.427382</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>0.758372</td>
<td>0.982137</td>
<td>0.982137</td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>28</td>
<td>0.680729</td>
<td>0.821663</td>
<td>0.821663</td>
</tr>
<tr>
<td>37</td>
<td>0.829027</td>
<td>0.602269</td>
<td>0.602269</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>0.569077</td>
<td>0.771010</td>
<td>0.728347</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>0.785553</td>
<td>0.847495</td>
<td>0.847495</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>0.431460</td>
<td>0.830321</td>
<td>0.830321</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>16</td>
<td>0.957990</td>
<td>0.659941</td>
<td>0.659941</td>
</tr>
<tr>
<td>35</td>
<td>0.814199</td>
<td>0.785704</td>
<td>0.785704</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>0.894919</td>
<td>0.903971</td>
<td>0.903971</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.320639</td>
<td>0.243167</td>
<td>0.243167</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>0.734543</td>
<td>0.804033</td>
<td>0.804033</td>
<td></td>
</tr>
</tbody>
</table>

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, $MSE$s 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 AIC or BIC, application to dynamic system identification and
so on are the future works.

References


