Automation, Control and Intelligent Systems
2016; 4(6): 84-88
http://www.sciencepublishinggroup.com/j/acis
doi: 10.11648/j.acis.20160406.11
ISSN: 2328-5583 (Print); ISSN: 2328-5591 (Online)

Improvement of Echo State Network Generalization by Selective Ensemble Learning Based on BPSO

Xiaodong Zhang, Xuefeng Yan

Key Laboratory of Advanced Control and Optimization for Chemical Processes of Ministry of Education, East China University of Science and Technology, Shanghai, P. R. China

Email address:
xfyan@ecust.edu.cn (Xuefeng Yan)

To cite this article:
Xiaodong Zhang, Xuefeng Yan. Improvement of Echo State Network Generalization by Selective Ensemble Learning Based on BPSO. *Automation, Control and Intelligent Systems.* Vol. 4, No. 6, 2016, pp. 84-88. doi: 10.11648/j.acis.20160406.11

Received: November 14, 2016; **Accepted**: November 21, 2016; **Published**: December 1, 2016

Abstract: The Echo State Network (ESN) is a novel and special type of recurrent neural network that has become increasingly popular in machine learning domains such as time series forecasting, data clustering, and nonlinear system identification. This network is characterized by large randomly constructed recurrent neural networks (RNN) called "reservoir", in which the neurons are sparsely connected and the weights remain unchanged during training, leaving the simple training of the output layer. However, the reservoir is criticized for its randomness and instability because of the random initialization of the connectivity and weights. In this article, we introduced the selective ensemble learning based on BPSO to improve the generalization performance of ESN. Two widely studied tasks are used to prove the feasibility and priority of the selective ESN ensemble based on BPSO(SESNE-BPSO) model. And the results indicate that the SESNE-BPSO model performs better than the general ESN ensemble, the single standard ESN and several other improved ESN models.

Keywords: Echo State Network, Reservoir Computing, Artificial Neural Network, Ensemble Learning, Selective Ensemble, Particle Swarm Optimization

1. Introduction

In recent years, reservoir computing (RC) [1, 2] has been extensively studied as a novel kind of training approach in the machine learning community for recurrent neural network (RNN). The RC approach consists of a large randomly constructed RNN called "reservoir", wherein the neurons are sparsely connected and the weights remain unchanged during training. With this approach, only the weights of networks from the reservoir to the readout layer require training through linear regression methods. Therefore, RC approach has numerous advantages such as high modeling accuracy, strong modeling capacity and low computational complexity. The echo state network (ESN) [3, 4], liquid state machines [5] and Evolino [6] are some examples of the RC approach. In this paper, we discuss the most popular form of RC, the ESN.

ESN is characterized by a large reservoir (generally 100--1000 neurons) converting the input data to a high-dimensional dynamic state space, which can be the "echo" of recent input history. ESN has been applied in a wide range of domains, such as nonlinear system identification [7] and time series prediction [8, 9]. However, one of ESN's flaws is its poorly understood reservoir properties. The randomly generated connectivity values and the weight structure of internal neurons in the reservoir may lead to the randomness and instability of ESN in prediction performance. Nevertheless, the random and unstable prediction is not constantly considered a disadvantage of machine learning algorithm. For ensemble learning [10], one of the most popular machine learning algorithm, the randomness and diversity of individual learners in an ensemble contribute in promoting the generalization performance of the learner's ensembles. Therefore, the ensemble learning method is introduced to the ESN model to solve the proposed ESN problem.

Ensemble learning [11-13] is a machine learning algorithm which improves learning performance by training multiple component learners to solve the same task. The final ensemble's output is the average of all individual learners' outputs. The ensemble learning has been widely recognized to provide a better generalization performance compared with a single component learner [14]. The effectiveness of ensemble learning can be explained by the bias and variance

decomposition of the ensemble error [15]. Ensemble learning can reduce both the bias and variance of ensemble error. As is studied in [16], the trade-off of individuals' accuracy and diversity is the key to improve the generalization performance of ensemble. However, whether all the trained individuals networks should be selected into ensemble? Zhou et al. [17] proposed that a selective subset of all individuals can be more effective than ensemble all the individuals. The selective ensemble, which combines the diverse individuals selected from plenty of trained accurate networks, has been proved effective theoretically and practically.

One of the most important procedure for selective ensemble is how to select the diverse individuals from a number of trained accurate networks, which can be regarded as a feature selection problem. Some several classical feature methods such as forward selection, backward elimination can be applied to select the most effective subsets of individuals. However, those methods are almost greedy search algorithms, which suffer from the stagnation in local optima. As well-known, the evolutionary computation techniques are famous for the global search ability. Compared with genetic algorithms (GA) [18], particle swarm optimization (PSO) [19] has many advantages such as fewer parameters and higher convergence speed. Additionally, the optimization of whether the individuals are selected into the ensemble is a discrete optimization problem. Therefore, a discrete binary version of PSO, called binary particle swarm optimization (BPSO) [20], is introduced to solve the binary combinational optimization problem.

In this paper, the selective ensemble based on BPSO algorithm was incorporated introduced to ESN to promote the generalization performance. To my knowledge, this is the first time that the selective ensemble algorithm is applied to ESN.

2. Echo State Network

2.1. Architecture of the ESN

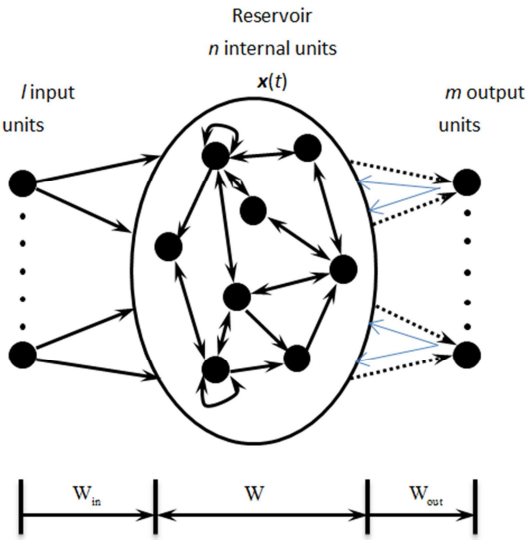

Fig. 1. *Basic structure of a standard ESN. The solid lines denote the randomly created connections fixed prior to training. The dotted lines denote the connections adjusted during training. The grey solid lines denote the feedback connections that are possible but not required.*

The ESN is a kind of RNN whose structure can be divided into three sections: a liner input layer with l input neurons, a large and fixed RNN with n internal neurons, and a linear readout layer with m output neurons. The fixed RNN part where the neurons are sparsely connected and the weights maintain unchanged during training is called "reservoir". Fig. 1 illustrates the basic structure of the ESN.

The states of internal neurons $\boldsymbol{x}(t)$ and output variables $\boldsymbol{y}(t)$ at a specific time point t are expressed as follows [21]:

$$\boldsymbol{x}(t) = f(\boldsymbol{W}_{in} \cdot \boldsymbol{u}(t) + \boldsymbol{W} \cdot \boldsymbol{x}(t-1) + \boldsymbol{W}_{back} \boldsymbol{y}^T(t-1)) \quad (1)$$

$$\boldsymbol{y}(t) = \boldsymbol{x}^T(t) \cdot \boldsymbol{W}_{out} \quad (2)$$

where f is the internal neuron stimulation function(typically a tanh sigmoid function), and $\boldsymbol{u}(t)$, $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$ are the input variable, internal neuron state, and output variable at a specified time step t, respectively. $\boldsymbol{W}_{in}$ is the $n \times l$ matrix, which indicates the input weights to the reservoir; $\boldsymbol{W}$ is the $n \times n$ matrix, which denotes the internal connection weights of the reservoir; $\boldsymbol{W}_{out}$ is the $m \times n$ matrix, which represents the output (readout) weights from the reservoir.; $\boldsymbol{W}_{back}$ is the $n \times m$ matrix, which indicates the feedback weights from the output to the reservoir. The initialization of reservoir state $\boldsymbol{s}(t)$ is a zero vector. The superscripted T represents transpose.

2.2. Training of the ESN

As discussed above, $\boldsymbol{W}_{in}$, $\boldsymbol{W}_{back}$ and $\boldsymbol{W}$ are the fixed matrices generated in advance generated by using the stochastic numerical values obtained from a uniform distribution, which means that only trainable matrix is the output weight matrix $\boldsymbol{W}_{out}$. For ESN to maintain the "Echo State Property", which means that the internal neuron state is a nonlinear transformation of the entire history of the input signal, the spectral radius of the internal connection weights $\boldsymbol{W}$ should be set to less than 1[3]. Thus $\boldsymbol{W}$ is generally scaled by $\alpha / |\lambda_{\max}|$, where $|\lambda_{\max}|$ is the spectral radius of $\boldsymbol{W}$ and α is a scaling parameter between 0 and 1.

The internal neuron state $\boldsymbol{X}$, obtained during the training process, can be expressed as follows:

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}^T(1) \\ \boldsymbol{x}^T(2) \\ \vdots \\ \boldsymbol{x}^T(n) \end{bmatrix} \quad (3)$$

and the output data stream state matrix $\boldsymbol{Y}$ can be expressed as follows:

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}(1) \\ \boldsymbol{y}(2) \\ \vdots \\ \boldsymbol{y}(n) \end{bmatrix} \quad (4)$$

where n represents the number of the training sample. Consequently, the output matrix $\boldsymbol{W}_{out}$ to be adjusted during training should solve a linear regression problem:

$$\boldsymbol{X} \cdot \boldsymbol{W}_{out} = \boldsymbol{Y} \tag{5}$$

the common method uses the least-squares solution:

$$\boldsymbol{W}_{out} = \arg\min_{w} \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{Y} \right\|^2 \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm, and the desired $\boldsymbol{W}_{out}$ is calculated by the following equation:

$$\boldsymbol{W}_{out} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{T}\boldsymbol{Y} \tag{7}$$

This is implemented by the pseudo-inverse algorithm.

3. Selective ESN Ensembles Based on BPSO

3.1. Review of BPSO

Particle swarm optimization (PSO) [22] method was first proposed by Kennedy and Eberhart to solve the numerical optimization problem. As an evolutionary computational technique, PSO introduced a population of particles to simulate the bird flocks to search the best solution to the problem. Each particle represents a candidate solution. Then the discrete binary version of PSO (BPSO)[20] was proposed to solve the combinatorial optimization problem in 1997. In BPSO, each dimension of a particle' position is limited to 0 or 1. The velocity and position of each particle can be updated according to Eq.(9)-(11):

$$v_{id}(t+1) = w \cdot v_{id}(t) + c_1 \cdot rand()(p\_best_{it} - x_{id}) + c_2 \cdot rand()(g\_best_{d} - x_{id}) \tag{8}$$

$$S(v_{id}) = sigmoid(v_{id}) = \frac{1}{1+e^{-v_{id}}} \tag{9}$$

$$\text{if } rand() < S(v_{id}(t+1)) \text{ then } x_{id}(t+1) = 1, \text{ else } x_{id}(t+1) = 0 \tag{10}$$

Where $rand()$ represents a random function on the domain [0,1], p_best_{it} denotes the personal best of the it particle and g_best_{d} denotes global best for the d particle; c_1, c_2 and w are the parameters; $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ donates the position of i^{th} particle. $v_i = (v_{i1}, v_{i2}, \ldots, v_{id})$ represents velocity for particle i. v_{id} is limited in the range of $[-v_{\max}, v_{\max}]$.

3.2. SESNE-BPSO

In this section, the selective ESN ensemble based on BPSO (SESNE-BPSO) is described in detail. For the problem of selective ensemble, each dimension of the particle' position values 0 or 1 to denote whether the originally generated individual ESN is selected or not. The position of particle $\boldsymbol{x} = [x_1, x_2, \ldots x_m]$ denotes the selection status of the ensemble. The dimension of each particle is the size of the originally generated ensemble.

$$SeEn(x_i) = \left\{ ESN_j \,\middle|\, x_{ij} = 1, ESN_j \in En \right\} \tag{11}$$

Where $SeEn$ and En denote the selective ensemble and the originally generated ensemble respectively.

The objective optimization function is the normalized root mean square error(NRMSE).

$$NRMSE = \sqrt{\sum_{t=1}^{N} \frac{(\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t))^2}{N \cdot \sigma^2}} \tag{12}$$

where $\boldsymbol{y}(t)$ is the desired output(target), $\hat{\boldsymbol{y}}(t)$ is the output, σ^2 is the variance of $\boldsymbol{y}(t)$, and N is the total number of $\boldsymbol{y}(t)$.

The procedure for the SESNE-BPSO can be summarized as follows:

(1) All the data are divided into three parts: training, validation and the testing set.

(2) Generate m ESNs with the input weights W_{in} and the internal connection weights W initialized at random values. The other parameters of the standard ESN such as the sparse degree of reservoir, the spectral radius, and the input extension are confirmed through the validation set.

(3) Each generated ESN is trained using the algorithm described in section 2.2 with the training data.

(4) Choose the error function NRMSE of the selected ensemble $SeEn$ represented by x_i according to Eq.(11) as the objective optimization function. Select $SeEn$ from En by minimizing the error function on the validation set with the BPSO.

(5) The best $SeEn$ of the validation performance is found out.

4. Experiment and Result

4.1. Experimental Setup

In this section, the proposed SESNE-BPSO method was evaluated using two extensively studied tasks obtained from previous literature on ESN. The model performance is evaluated by the percentage of the NRMSE. The results of the proposed SESNE-BPSO performance are compared with those of the general ESN ensemble (ESN-En), which ensemble all the originally generated ESNs, and the single standard ESN as well as two other improved ESN models.

The number of originally generated ESNs is 20 for the following two experiments

4.2. Experiment Tasks and Results

A) NARMA system

The 10-th order nonlinear autoregressive moving average (NARMA) system is described in the following equation [7]:

$$o(t+1) = 0.3o(t) + 0.05o(t)\sum_{i=0}^{9} o(t-i) + 1.5u(t-9)u(t) + 0.1 \quad (13)$$

where $o(t)$ denotes the NARMA system output at time t, $u(t)$ represents the system input at time t, and $u(t)$ refers to an independent identically distributed stream of values generated uniformly from [0, 0.5]. The NARMA system identification task has been described in Jaeger [7], the ESN is trained to output $o(t)$ based on $u(t)$. Modeling the NARMA system is generally difficult because the system is strongly nonlinear and requires a substantially long memory to accurately reproduce the output. The current output of the system is decided by both the input data and the previous output data from up to 10 steps ago. The NARMA data-set used in this experiment contains 6,000 items and all the values are divided into 3 parts. The first part is the training data-set with 2,000 values, the second part is the validation data-set with 2,000 values, and the third part is the testing data-set with the remaining 2,000 values. The first 100 values of each part are stored to wash out the initial memory of the dynamic reservoir. The reservoir size (N) of ESNs for this task is set to 100. The experiments are performed 10 times because of the random initialization of ESN. The testing performance NRMSE of the single standard ESN, ESN-En, and SESNE-BPSO are displayed in Table 1. Mean represents the mean value of NRMSE, SD stands for the standard deviation of NRMSE, Max indicates the maximum value of NRMSE, and Min stands for the minimum value of NRMSE.

Table 1. *Testing performance NRMSE of the single standard ESN, ESN-En, and SESNE-BPSO for the NARMA time series task.*

Algorithm	ESN	ESN-En	SESNE-BPSO
Mean	0.169	0.124	0.101
SD	0.0374	0.0273	0.0316
Max	0.189	0.137	0.112
Min	0.152	0.118	0.096

B) Laser Time Series

The laser chaotic time series data [23] used in this prediction task is a real-world sequence obtained from the Santa Fe Competition by sampling the intensity of a far-infrared laser in a chaotic regime. The task is set to forecast the next value $o(t+1)$ (one step ahead forecast) depending on the history values up to time t. Laser time series prediction is generally difficult because of its numerical round-off noise and diverse time scales, especially in the breakdown events of the sequence. The laser data set used in this experiment contained 10,000 values, which are divided into 3 parts. The first part is the training data-set with 6,000 values, the second part is the validation data set with 2,000 values, and the third part is the testing data set with the remaining 2,000 values. The first 1000 values of each part are also stored to wash out the initial memory of the dynamic reservoir. This laser series prediction task needs feedback connections. The reservoir size (N) of the ESNs for this task is also set to 100. The bias input is a constant 0.02 value. The experiments are conducted for 10 times because of the random initialization of the ESN. The testing performance of the single standard ESN, ESN-En and SESNE-BPSO are displayed in Table 2.

Table 2. *Testing performance NRMSE of the single standard ESN, ESN-En and SESNE-BPSO for the laser time series prediction.*

Algorithm	ESN	ESN-En	SESNE-BPSO
Mean	0.0195	0.0167	0.0149
SD	5.854e-3	5.810e-3	5.703e-3
Max	0.0229	0.0173	0.0156
Min	0.0172	0.0145	0.0128

To validate the performance of the proposed SESNE-BPSO model, other improved ESN models, such as L2-Boost ESN [24], Scale-Free Highly Clustered ESN(SFHC-ESN) [25], are performed for the comparison. The result of the comparison is presented in Table 3.

Table 3. *The test performance of SESNE-BPSO and other improved ESN models.*

Algorithm	Task	
	NARMA	Laser
ESN	0.169	0.0195
L2-Boost ESN	0.12	0.0152
SFHC-ESN	-	0.0163
SESNE-BPSO	0.101	0.0149

4.3. Discussion

Based on the data from Table 1 and Table 2, the experimental results indicate that the ESN-En model obviously improved the performance of generalization compared with the standard ESN and the proposed SESNE-BPSO outperformed the ESN-En. Furthermore, SESNE-BPSO performs better than several other improved ESN models based on Table 3. This result illustrates that the selective ensemble learning based on BPSO algorithm promotes the generalization performance of the ESN ensemble.

5. Conclusion

In this paper, a novel ESN ensemble called SESNE-BPSO is proposed. Ensemble learning is introduced to improve the generalization performance of the ESN model. The diversity of the individual ESNs in the ensemble is one of the key factors in reducing the ensemble generalization error. The diverse ESNs are created because of the random initialization of input and internal weights. The selective ensemble learning based on BPSO algorithm is applied as an ensemble learning approach to further increase the performance of ESN ensemble. Two widely used tasks are performed to test the performance of the proposed SESNE-BPSO model. The

results indicate that SESNE-BPSO performs better than the general ESN ensemble, the standard ESN and other improved ESN models. Consequently the findings demonstrate the feasibility and superiority of the selective ensemble learning based on BPSO approach to ESN.

Acknowledgements

The authors gratefully acknowledge the support of the following foundations: 973 project of China (2013CB733605).

References

[1] B. Schrauwen, D. Verstraeten, J. Van Campenhout, An overview of reservoir computing: theory, applications and implementations, Proceedings of the 15th European Symposium on Artificial Neural Networks. p. 471-482 20072007), pp. 471-482.

[2] M. LukošEvičIus, H. Jaeger, Reservoir computing approaches to recurrent neural network training, Computer Science Review, 3 (2009) 127-149.

[3] H. Jaeger, Tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the" echo state network" approach (GMD-Forschungszentrum Informationstechnik, 2002).

[4] H. Jaeger, Reservoir riddles: Suggestions for echo state network research, Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005., (IEEE2005), pp. 1460-1462.

[5] W. Maass, Liquid state machines: motivation, theory, and applications, Computability in context: computation and logic in the real world, (2010) 275-296.

[6] J. Schmidhuber, D. Wierstra, M. Gagliolo, F. Gomez, Training recurrent networks by evolino, Neural computation, 19 (2007) 757-779.

[7] H. Jaeger, Adaptive nonlinear system identification with echo state networks, Advances in neural information processing systems2002), pp. 593-600.

[8] H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, Science, 304 (2004) 78-80.

[9] S.-X. Lun, X.-S. Yao, H.-Y. Qi, H.-F. Hu, A novel model of leaky integrator echo state network for time-series prediction, Neurocomputing, 159 (2015) 58-66.

[10] C. Zhang, Y. Ma, Ensemble machine learning (Springer, 2012).

[11] D. West, S. Dellana, J. Qian, Neural network ensemble strategies for financial decision applications, Computers & operations research, 32 (2005) 2543-2559.

[12] L. K. Hansen, P. Salamon, Neural network ensembles, IEEE transactions on pattern analysis and machine intelligence, 12 (1990) 993-1001.

[13] Z.-H. Zhou, Ensemble learning, Encyclopedia of Biometrics, (2015) 411-416.

[14] L. K. Hansen, P. Salamon, Neural network ensembles, IEEE Transactions on Pattern Analysis & Machine Intelligence, (1990) 993-1001.

[15] G. Valentini, T. G. Dietterich, Bias—Variance Analysis and Ensembles of SVM, International Workshop on Multiple Classifier Systems, (Springer2002), pp. 222-231.

[16] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine learning, 51 (2003) 181-207.

[17] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial intelligence, 137 (2002) 239-263.

[18] L. Davis, Handbook of genetic algorithms, (1991).

[19] J. Kennedy, Particle swarm optimization, Encyclopedia of machine learning, (Springer, 2011), pp. 760-766.

[20] J. Kennedy, R. C. Eberhart, A discrete binary version of the particle swarm algorithm, Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on, (IEEE1997), pp. 4104-4108.

[21] H. Jaeger, The “echo state” approach to analysing and training recurrent neural networks-with an erratum note, Bonn, Germany: German National Research Center for Information Technology GMD Technical Report, 148 (2001) 34.

[22] R. C. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, Proceedings of the sixth international symposium on micro machine and human science, (New York, NY1995), pp. 39-43.

[23] A. Weigend, N. Gershenfeld, Time Series Prediction: Forecasting the Future and Understanding the Past. 1994, Proceedings of a NATO Advanced Research Workshop on Comparative Time Series Analysis, held in Santa Fe, New Mexico).

[24] S. Basterrech, An Empirical Study of the L2-Boost technique with Echo State Networks, arXiv preprint arXiv:1501.00503, (2015).

[25] Z. Deng, Y. Zhang, Collective behavior of a small-world recurrent neural system with scale-free distribution, IEEE Transactions on Neural Networks, 18 (2007) 1364-1375.