


Performance improvement via bagging in probabilistic prediction of chaotic time series using similarity of attractors and LOOCV predictable horizon

Shuichi Kurogi¹  · Mitsuki Toidani¹ · Ryosuke Shigematsu¹ · Kazuya Matsuo¹

Received: 5 June 2016 / Accepted: 6 July 2017

© The Author(s) 2017. This article is an open access publication

Abstract Recently, we have presented a method of probabilistic prediction of chaotic time series. The method employs learning machines involving strong learners capable of making predictions with desirably long predictable horizons, where, however, usual ensemble mean for making representative prediction is not effective when there are predictions with shorter predictable horizons. Thus, the method selects a representative prediction from the predictions generated by a number of learning machines involving strong learners as follows: first, it obtains plausible predictions holding large similarity of attractors with the training time series and then selects the representative prediction with the largest predictable horizon estimated via LOOCV (leave-one-out cross-validation). The method is also capable of providing average and/or safe estimation of predictable horizon of the representative prediction. We have used CAN2s (competitive associative nets) for learning piecewise linear approximation of nonlinear function as strong learners in our previous study, and this paper employs bagging (bootstrap aggregating) to improve the performance, which enables us to analyze the validity and the effectiveness of the method.

Keywords Probabilistic prediction of chaotic time series · Long-term unpredictability · Attractors of chaotic time series · Leave-one-out cross-validation · Estimation of predictable horizon

1 Introduction

So far, a number of methods for time series prediction have been studied (cf. [1, 2]), and our methods have awarded 3rd and 2nd places in the competitions of time series prediction held at IJCNN'04 [3] and ESTSP'07 [4], respectively. Our methods have used model selection methods evaluating MSE (mean square prediction error) for holdout and/or cross-validation datasets. Recently, we have developed several model selection methods for chaotic time series prediction [5, 6]. The method in [5] utilizes moments of predictive deviation as ensemble diversity measures for model selection in time series prediction and achieves better performance from the point of view of MSE than the conventional holdout method. The method in [6] uses direct multistep ahead (DMS) prediction to apply the out-of-bag (OOB) estimate of MSE. Although both methods have selected the models to generate good predictions on average, they cannot always have provided good predictions, especially when the horizon to be predicted is large. This is owing mainly to the fact that the MSE of a set of predictions is largely affected by a small number of predictions with short predictable horizons even if most of the predictions have long predictable horizons. This is because the prediction error of chaotic time series increases exponentially with the increase in time after the predictable horizon (see [6] for the analysis and [1] for properties of chaotic time series).

Instead of using model selection methods employing the estimation of the MSE, we have developed a method of probabilistic prediction of chaotic time series [7]. Here, from [8], the probabilistic prediction has come to dominate the science of weather and climate forecasting, mainly because the theory of chaos at the heart of meteorology shows that for a simple set of nonlinear equations (or Lorenz's equations shown below) with initial conditions changed by

✉ Shuichi Kurogi
kuro@cntl.kyutech.ac.jp;
<http://kurolab.cntl.kyutech.ac.jp>

¹ Department of Control Engineering, Kyushu Institute of Technology, Tobata, Kitakyushu, Fukuoka 804-8550, Japan

minute perturbations, there is no longer a single deterministic solution and hence all forecasts must be treated as probabilistic. Although most of the methods shown in [8] use ensemble mean for representative forecast, our method in [7] (see below for details) uses an individual prediction selected from a set of plausible predictions for the representative because our method employs learning machines involving strong learners capable of making predictions with small error for a desirably long duration and we can see that ensemble mean does not work when the set of predictions for the ensemble involves a prediction with short predictable horizon. This is owing mainly to the exponential increase in prediction error of chaotic time series after the predictable horizon (see Sect. 3.2 for details)

Thus, instead of using ensemble mean, our method in [7] firstly selects plausible predictions by means of evaluating the similarity of attractors between training and predicted time series and then obtains the representative prediction by means of LOOCV (leave-one-out cross-validation) to select the prediction with longer predictable horizon. Comparing with our previous methods using the MSE for model selection [5, 6], the method in [7] has an advantage that it is capable of selecting the representative prediction from plausible predictions for each start time of prediction and providing the estimation of predictable horizon. Furthermore, it has achieved long predictable horizons on average. However, there are several cases where the method selects representative prediction with short predictable horizon, although there are plausible predictions with longer predictable horizons.

To overcome this problem, this paper tries to improve the performance of learning machines by using bagging (bootstrap aggregating) method and show the analysis of LOOCV predictable horizon. Here, the bagging is known to use ensemble mean to have an ability to reduce the variance of predictions by single learning machines, and then, we can expect that the performance in time series prediction becomes more stable and higher. Note that, in this paper, the bagging ensemble is employed for iterated one-step-ahead (IOS) prediction of time series, and we deal with probabilistic prediction as an ensemble of longer-term predictions. Furthermore, we use CAN2 (competitive associative net 2) as a learning machine (see [3] for the details of CAN2), where CAN2 has been introduced for learning piecewise linear approximation of nonlinear function and the performance has been shown in evaluating predictive uncertainty challenge [9], where our method has been awarded the first place in regression problems. The CAN2 has been used in our methods [3, 4] for the competitions of time series predictions shown above.

We show the present method of probabilistic prediction of chaotic time series in Sect. 2, experimental results and analysis in Sect. 3, and the conclusion in Sect. 4.

2 Probabilistic prediction of chaotic time series

2.1 IOS prediction of chaotic time series

Let $y_t (\in \mathbb{R})$ denote a chaotic time series for a discrete time $t = 0, 1, 2, \dots$ satisfying

$$y_t = r(\mathbf{x}_t) + e(\mathbf{x}_t), \quad (1)$$

where $r(\mathbf{x}_t)$ is a nonlinear target function of a vector $\mathbf{x}_t = (y_{t-1}, y_{t-2}, \dots, y_{t-k})^T$ generated by k -dimensional delay embedding from a chaotic differential dynamical system (see [1] for the theory of chaotic time series). Here, y_t is obtained not analytically but numerically, and then, y_t involves an error $e(\mathbf{x}_t)$ owing to an executable finite calculation precision. This indicates that there are a number of plausible target functions $r(\mathbf{x}_t)$ with allowable error $e(\mathbf{x}_t)$. Furthermore, in general, a time series generated with higher precision has small prediction error for longer duration of time from the initial time of prediction. Thus, let a time series generated with a high precision (or 128-bit precision; see Sect. 3 for details), be ground truth time series $y_t^{[st]}$, while we examine predictions generated with standard 64-bit precision.

Let $y_{t:h} = y_t y_{t+1} \dots y_{t+h-1}$ denote a time series with the initial time t and the horizon h . For a given training time series $y_{t_g:h_g} (= y_{t_g:h_g}^{[rain]})$, we are supposed to predict succeeding time series $y_{t_p:h_p}$ for $t_p \geq t_g + h_g$. Then, we make the training dataset $D^{[rain]} = \{(\mathbf{x}_t, y_t) \mid t \in I^{[rain]}\}$ for $I^{[rain]} = \{t \mid t_g + k \leq t < t_g + h_g\}$ to train a learning machine. After the learning, the machine executes IOS prediction by

$$\hat{y}_t = f(\mathbf{x}_t) \quad (2)$$

for $t = t_p, t_{p+1}, \dots$, recursively, where $f(\mathbf{x}_t)$ denotes prediction function of $\mathbf{x}_t = (x_{t1}, x_{t2}, \dots, x_{tk})$ whose elements are given by $x_{ij} = y_{t-j}$ for $t-j < t_p$ and $x_{ij} = \hat{y}_{t-j}$ for $t-j \geq t_p$. Here, we suppose that y_t for $t < t_p$ is known as the initial state for making the prediction $\hat{y}_{t_p:h_p}$. As explained above, we execute the prediction with standard 64-bit precision, and we may say that there are a number of plausible prediction functions $f(\mathbf{x}_t)$ with small error for a duration of time from the initial time of prediction by means of using strong learning machines.

2.2 Single CAN2 and the bagging for IOS prediction

We use CAN2 as a learning machine. A single CAN2 has N units. The j th unit has a weight vector $\mathbf{w}_j \triangleq (w_{j1}, \dots, w_{jk})^T \in \mathbb{R}^{k \times 1}$ and an associative matrix (or a row vector) $\mathbf{M}_j \triangleq (M_{j0}, M_{j1}, \dots, M_{jk}) \in \mathbb{R}^{1 \times (k+1)}$ for $j \in I^N \triangleq \{1, 2, \dots, N\}$. The CAN2 after learning the training dataset $D^{[rain]} = \{(\mathbf{x}_t, y_t) \mid t \in I^{[rain]}\}$ approximates the target function $r(\mathbf{x}_t)$ by

$$\hat{y}_t = \tilde{y}_{c(t)} = \mathbf{M}_{c(t)} \tilde{\mathbf{x}}_t, \tag{3}$$

where $\tilde{\mathbf{x}}_t \triangleq (1, \mathbf{x}_t^T)^T \in \mathbb{R}^{(k+1) \times 1}$ denotes the (extended) input vector to the CAN2, and $\tilde{y}_{c(t)} = \mathbf{M}_{c(t)} \tilde{\mathbf{x}}_t$ is the output value of the $c(t)$ th unit of the CAN2. The index $c(t)$ indicates the unit who has the weight vector $\mathbf{w}_{c(t)}$ closest to the input vector \mathbf{x}_t , or $c(t) \triangleq \operatorname{argmin}_{j \in I^N} \|\mathbf{x}_t - \mathbf{w}_j\|$. Note that the above prediction performs piecewise linear approximation of $y = r(\mathbf{x})$ and N indicates the number of piecewise linear regions. We use the learning algorithm shown in [10] whose high performance in regression problems has been shown in evaluating predictive uncertainty challenge [9].

We obtain bagging prediction by means of using a number of single CAN2s as follows (see [11, 12] for details); let $D^{[n\alpha; j]} = \{(\mathbf{x}_t, y_t) \mid t \in I^{[n\alpha; j]}\}$ be the j th bag (multiset, or bootstrap sample set) involving $n\alpha$ elements, where the elements in $D^{[n\alpha; j]}$ are resampled randomly with replacement from the training dataset $D^{[\text{train}]}$ involving $n = |D^{[\text{train}]}|$ elements. Here, $\alpha (> 0)$ indicates the bag size ratio to the given dataset, and $j \in J^{[\text{bag}]} \triangleq \{1, 2, \dots, b\}$. Here, note that $\alpha = 1$ is used in many applications (see [12, 13]), which we use in the experiments shown below after the tuning of α (see [12] for validity and effectiveness of using variable α). Using multiple CAN2s employing N units after leaning $D^{[n\alpha; j]}$, which we denote $\theta_N^{[j]} (\in \Theta_N \triangleq \{\theta_N^{[j]} \mid j \in J^{[\text{bag}]}\})$, the bagging for predicting the target value $r_{t_c} = r(\mathbf{x}_{t_c})$ is done by

$$\hat{y}_t^{[\theta_N^{[\text{bag}]}]} \triangleq \frac{1}{b} \sum_{j \in J^{[\text{bag}]}} \hat{y}_t^{[j]} \equiv \left\langle \hat{y}_t^{[j]} \right\rangle_{j \in J^{[\text{bag}]}} \tag{4}$$

where $\hat{y}_t^{[j]} \triangleq \hat{y}_t^{[j]}(\mathbf{x}_{t_c})$ denotes the prediction by the j th machine $\theta_N^{[j]}$. The angle brackets $\langle \cdot \rangle$ indicate the mean, and the subscript $j \in J^{[\text{bag}]}$ indicates the range of the mean. For simple expression, we sometimes use $\langle \cdot \rangle_j$ instead of $\langle \cdot \rangle_{j \in J^{[\text{bag}]}}$ in the following.

2.3 Probabilistic prediction and estimation of predictable horizon

2.3.1 Similarity of attractors to select plausible predictions

First, we make a number of IOS predictions $\hat{y}_{t_p:h_p}^{[\theta_N]} = y_{t_p:h_p}^{[\theta_N]}$ by means of learning machines or CAN2s, $\theta_N \in \Theta$, with different number N of units, where Θ indicates the set of all learning machines. We employ single and bagging CAN2s, which we denote $\theta_N^{[\text{single}]}$ and $\theta_N^{[\text{bag}]}$, respectively, if necessary. We suppose that there are a number of plausible prediction functions $f(\cdot) = f^{[\theta_N]}(\cdot)$, and we have to remove

implausible ones. To have this done, we select the following set of plausible predictions:

$$Y_{t_p:h_p}^{[S_{th}]} = \left\{ y_{t_p:h_p}^{[\theta_N]} \mid S(y_{t_p:h_p}^{[\theta_N]}, y_{t_g:h_g}^{[\text{train}]}) \geq S_{th}, \theta_N \in \Theta \right\} \tag{5}$$

where

$$S(y_{t_p:h_p}^{[\theta_N]}, y_{t_g:h_g}^{[\text{train}]}) \triangleq \frac{\sum_i \sum_j a_{ij}^{[\theta_N]} a_{ij}^{[\text{train}]}}{\sqrt{\sum_i \sum_j (a_{ij}^{[\theta_N]})^2} \sqrt{\sum_i \sum_j (a_{ij}^{[\text{train}]})^2}} \tag{6}$$

denotes the similarity of two-dimensional attractor (trajectory) distributions $a_{ij}^{[\theta_N]}$ and $a_{ij}^{[\text{train}]}$ of time series $y_{t_p:h_p}^{[\theta_N]}$ and $y_{t_g:h_g}^{[\text{train}]}$, respectively, and S_{th} is a threshold. Here, the two-dimensional attractor distribution, a_{ij} , of a time series $y_{t:h}$ is given by

$$a_{ij} = \sum_{s=t}^{t+h-1} \mathbf{I} \left\{ \left\lfloor \frac{y_s - v_0}{\Delta_a} \right\rfloor = i \wedge \left\lfloor \frac{y_{s+1} - v_0}{\Delta_a} \right\rfloor = j \right\}, \tag{7}$$

where v_0 is a constant less than the minimum value of y_t for all time series and Δ_a indicates a resolution of the distribution. Furthermore, $\mathbf{I}\{z\}$ is an indicator function equal to 1 if z is true, and 0 if z is false, and $\lfloor \cdot \rfloor$ indicates the floor function.

2.3.2 LOOCV measure to estimate predictable horizons

Let us define predictable horizon between two predictions $y_{t_p:h_p}^{[\theta_N]}$ and $y_{t_p:h_p}^{[\theta_{N'}]}$ in $Y_{t_p:h_p}^{[S_{th}]}$ as

$$h(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[\theta_{N'}]}) = \max \left\{ h \mid \forall s < h \leq h_p; |y_{t_p+s}^{[\theta_N]} - y_{t_p+s}^{[\theta_{N'}]}| \leq e_y \right\}, \tag{8}$$

where e_y indicates the threshold of prediction error to determine the horizon. Then, we employ LOOCV method to estimate predictable horizon of $y_{t_p:h_p}^{[\theta_N]}$ in $Y_{t_p:h_p}^{[S_{th}]}$. Namely, we use

$$\begin{aligned} \tilde{h}_{t_p:h_p}^{[\theta_N]} &= h(y_{t_p:h_p}^{[\theta_N]}, Y_{t_p:h_p}^{[S_{th}]} \setminus \{y_{t_p:h_p}^{[\theta_N]}\}) \\ &= \left\langle h(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[\theta_{N'}]}) \right\rangle_{y_{t_p:h_p}^{[\theta_{N'}]} \in Y_{t_p:h_p}^{[S_{th}]} \setminus \{y_{t_p:h_p}^{[\theta_N]}\}}, \end{aligned} \tag{9}$$

which we call LOOCV measure of predictable horizon or LOOCV predictable horizon. Here, we expect that $h(y_{t_p:h_p}^{[\theta_N]}, Y_{t_p:h_p}^{[S_{th}]} \setminus \{y_{t_p:h_p}^{[\theta_N]}\})$ and $h(y_{t_p:h_p}^{[\theta_N]}, y_t^{[\text{gt}]})$ have positive correlation by means of assuming that $Y_{t_p:h_p}^{[S_{th}]}$ involves a number of predictions neighboring $y_t^{[\text{gt}]}$.

2.3.3 Probabilistic prediction involving longer LOOCV predictable horizons

Let a subset of plausible predictions involving longer LOOCV predictable horizons be

$$Y_{t_p:h_p}^{[H_{th}, S_{th}]} = \left\{ y_{t_p:h_p}^{[\theta_{\sigma(i)}]} \mid \frac{i}{|Y_{t_p:h_p}^{[S_{th}]}|} \leq H_{th} \right\}, \tag{10}$$

where $\sigma(i)$ denotes the order of LOOCV predictable horizons satisfying $\tilde{h}_{t_p:h_p}^{[\theta_{\sigma(i)}]} \geq \tilde{h}_{t_p:h_p}^{[\theta_{\sigma(i+1)}]}$ for $i = 1, 2, \dots, |Y_{t_p:h_p}^{[S_{th}]}|$. The threshold H_{th} ($0 < H_{th} \leq 1$) indicates the ratio of the number of elements in $Y_{t_p:h_p}^{[H_{th}, S_{th}]}$ and $Y_{t_p:h_p}^{[S_{th}]}$, or $|Y_{t_p:h_p}^{[H_{th}, S_{th}]}| = H_{th} |Y_{t_p:h_p}^{[S_{th}]}|$. Now, we derive the probability of the prediction y_t for $t_p \leq t < t_p + h_p$ as

$$p(v_i \leq y_t < v_{i+1}) = \left\langle \mathbf{I} \left\{ \left\lfloor \frac{y_t^{[\theta]} - v_0}{\Delta_v} \right\rfloor = i \right\} \right\rangle_{\theta \in \Theta^{[H_{th}, S_{th}]}} \tag{11}$$

where $\Theta^{[H_{th}, S_{th}]}$ is the set of parameters θ of learning machines which have generated $y_{t_p:h_p}^{[\theta]} \in Y_{t_p:h_p}^{[H_{th}, S_{th}]}$, and Δ_v denotes the resolution of y_t , and $v_i = i\Delta_v + v_0$ for $i = 0, 1, 2, \dots$. Note that the probability $p(v_i \leq y_t \leq v_{i+1})$ indicates how much the plausible predictions in $Y_{t_p:h_p}^{[H_{th}, S_{th}]}$ take the values in between v_i and v_{i+1} .

2.3.4 Representative prediction and estimation of predictable horizon

Now, we provide $y_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ as a representative prediction, and an estimation of the predictable horizon $h_{t_p:h_p}^{[\theta_{\sigma(1)}]} = h(y_{t_p:h_p}^{[\theta_{\sigma(1)}]}, y_{t_p:h_p}^{[gt]})$ as

$$\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]} = \min \left\{ h(y_{t_p:h_p}^{[\theta_{\sigma(1)}]}, y_{t_p:h_p}^{[\theta]}) \mid \forall y_{t_p:h_p}^{[\theta]} \in Y_{t_p:h_p}^{[H_{th}, S_{th}]} \setminus y_{t_p:h_p}^{[\theta_{\sigma(1)}]} \right\}, \tag{12}$$

where we have to tune H_{th} from the point of view of accuracy and safeness. Here, the safe estimation of $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]}$

indicates that $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ is smaller than or equal to the actual predictable horizon $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$, and we can see that $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ become safer with the increase in H_{th} .

3 Numerical experiments and analysis

3.1 Experimental settings

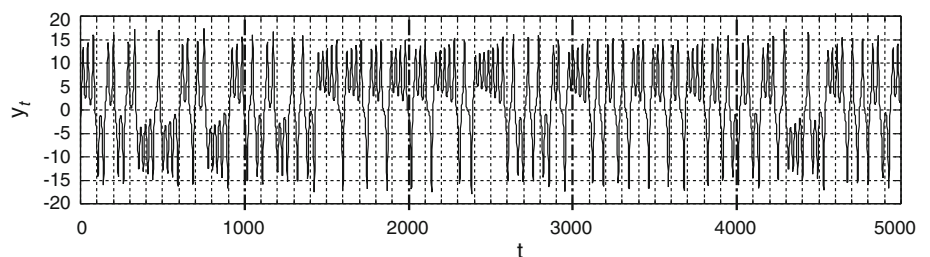
We use the Lorenz time series, as shown in Fig. 1 and [6], obtained from the original differential dynamical system given by

$$\begin{aligned} \frac{dx_c}{dt_c} &= \sigma(y_c - x_c), & \frac{dy_c}{dt_c} &= -x_c z_c + r x_c - y_c, \\ \frac{dz_c}{dt_c} &= x_c y_c - b z_c, \end{aligned} \tag{13}$$

for $\sigma = 10$, $r = 28$ and $b = 8/3$. Here, we use t_c for continuous time and $t (= 0, 1, 2, \dots)$ for discrete time related by $t_c = tT$ with the sampling time or the embedding delay $T = 25$ ms. We have generated the time series $y_t^{[gt]} = x_c(tT)$ for $t = 1, 2, \dots, 5000$ from the initial state $(x_c(0), y_c(0), z_c(0)) = (-8, 8, 27)$ via the fourth-order Runge–Kutta method with step size $\Delta t = 10^{-4}$ and $r = 128$ -bit precision of GMP (GNU multiprecision library).

Using $y_{t_g:h_g}^{[train]} = y_{0:2000}^{[gt]}$, we make the training dataset $D^{[train]} = \{(x_t^{[gt]}, y_t^{[gt]}) \mid t \in I^{[train]}\}$ for $I^{[train]} = \{10 (= k), 11, \dots, 1999\}$ and $x_t^{[gt]} = (y_{t-1}^{[gt]}, \dots, y_{t-k}^{[gt]})^T$. For learning machines θ_N , we have employed single CAN2s $\theta_N^{[single]}$ and bagging CAN2s $\theta_N^{[bag]}$ with the number of units $N = 5 + 20i$ ($i = 0, 1, 2, \dots, 14$). After the training, we execute IOS prediction $\hat{y}_t = f^{[\theta_N]}(x_t)$ for $t = t_p, t_p + 1, \dots$ with the initial input vector $x_{t_p} = (y_{t_p-1}^{[gt]}, \dots, y_{t_p-k}^{[gt]})$ for prediction start time $t_p \in T_p = \{2000 + 100i \mid i = 0, 1, 2, \dots, 29\}$ and prediction horizon $h_p = 500$. We show experimental results for the embedding dimension being $k = 10$ and the threshold in (8) being $e_y = 10$ (see [7] for the result with $k = 8$, which is not significantly but slightly different).

Fig. 1 Lorenz time series y_t for $t = 0, 1, 2, \dots, 4999$, or ground truth time series $y_{0:5000}^{[gt]}$



In order to estimate the accuracy of $y_t^{[st]}$, we have obtained an average predictable horizon $\langle h(y_{t:500}^{[st]}, y_{t:500}^{[\Delta t=10^{-5}, r=128]}) \rangle_{t \in T_p} = 230$ steps ($=5.75$ s/25 ms) for the time series $y_{t:500}^{[\Delta t=10^{-5}, r=128]}$ generated with $\Delta t = 10^{-5}$ and $r = 128$ -bit precision via the Runge–Kutta method. This indicates that $y_t^{[st]}$ with $\Delta t = 10^{-4}$ and $r = 128$ is considered to be accurate during 230 steps on average because we have observed that predictable horizon of two time series generated by the Runge–Kutta method with step sizes $\Delta t = 10^{-n}$ and 10^{-n-1} for $n = 3, 4, 5, 6, 7$ increases monotonically with the decrease in step size or the increase in n .

Here, note that we have executed several experiments with using the parameter $\theta = (N, k)$ for $k = 6, 8, 10, 12$ and so on, and we do not have found out any critically different results, although we would like to execute and show the results of comparative study in our future research.

3.2 Results and analysis

First, we show an example of all predictions $y_{t_p}^{[\theta_N]}$ for $t_p = 2300$ in Fig. 2a. Note that $t_p = 2300$ is the start time of representative prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ with predictable horizon $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ being smaller than 100 by single CAN2 (actually $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]} = 72$) and improved by bagging CAN2 as $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]} = 183$ (see Fig. 3a).

In Fig. 2b, we can see that single CAN2s have larger number of predictions with the similarity S smaller than $S_{th} = 0.8$ than bagging CAN2s at $t = 2799$, and their predictions are not selected as plausible predictions. A detailed analysis of the similarity is shown below.

The representative prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ (green) shown in (c) is chosen by means of selecting the largest LOOCV predictable horizon $\tilde{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ shown in (d). From (d), we can see that the single CAN2 (left) has actual predictable horizon $h_{t_p:h_p}^{[\theta_N]}$ larger than 200 and LOOCV predictable horizon $\tilde{h}_{t_p:h_p}^{[\theta_N]}$ smaller than 100, actually $(h_{t_p:h_p}^{[\theta_N]}, \tilde{h}_{t_p:h_p}^{[\theta_N]}) = (209, 72.1)$. Since the present method selects the prediction with the largest $\tilde{h}_{t_p:h_p}^{[\theta_N]}$, the prediction with $h_{t_p:h_p}^{[\theta_N]} = 209$ could not have selected. On the other hand, we can see that bagging CAN2 (right in (d)) successfully selects the prediction with $h_{t_p:h_p}^{[\theta_N]}$ larger than 100, actually $(h_{t_p:h_p}^{[\theta_N]}, \tilde{h}_{t_p:h_p}^{[\theta_N]}) = (183, 191)$. Precisely, bagging CAN2s have successfully provided

large $\tilde{h}_{t_p:h_p}^{[\theta_N]} = 191$ because there are a number of predictions with long predictable horizons around $h_{t_p:h_p}^{[\theta_N]} = 200$ as shown as the group of points neighboring $h_{t_p:h_p}^{[\theta_N]} = 200$ in (d) on the right-hand side. Incidentally, from (c), we can see that ensemble mean does not seem appropriate for producing representative prediction in long-term prediction of chaotic time series.

In Fig. 3, we show the results of actual and estimated predictable horizons. Note that we have obtained $\langle h(y_{t:500}^{[st]}, y_{t:500}^{[\Delta t=5 \times 10^{-4}, r=64]}) \rangle_{t \in T_p} = 172$ steps ($=4.3$ s/25ms) and $\langle h(y_{t:500}^{[st]}, y_{t:500}^{[\Delta t=10^{-3}, r=64]}) \rangle_{t \in T_p} = 142$ steps ($=3.55$ s/25 ms) and the former is almost the same as the mean of predictable horizons achieved by single and bagging CAN2 being 170 and 175 steps, respectively. This indicates that single and bagging CAN2s after learning the training data generated via the Runge–Kutta method with the step size $\Delta t = 10^{-4}$ have almost the same prediction performance as the Runge–Kutta method with $\Delta t = 5 \times 10^{-4}$. Although we do not have no general measure to evaluate time series prediction so far, the above method using the step size of Runge–Kutta method and the mean predictable horizon seems reasonable. In Fig. 3a, we can see that the performance of the stability of prediction by single CAN2 is improved by bagging CAN2 from the point of view that the former has four actual predictable horizons $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]}$ smaller than 100 among all predictions for $t_p \in T_p$ and bagging CAN2 has achieved all $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}$ larger than 100. From (b), we can see that the estimated predictable horizon $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ with $H_{th} = 0.5$ is almost the same as actual predictable horizon $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$, while $H_{th} = 0.9$ has achieved safe estimation, or $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}]} \leq h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$,

In order to analyze the property of the method, we show the attractor distribution of training and representative time series in Fig. 4. We can see that the similarity of attractors $S(y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]}, y_{t_g:h_g}^{[train]}) = 0.859$ obtained by single CAN2 is smaller than $S(y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}, y_{t_g:h_g}^{[train]}) = 0.939$ obtained by bagging CAN2. From the result on the left in Fig. 2b, we can see that there is a prediction with the similarity larger than 0.859 for single CAN2. Actually, the maximum similarity of single CAN2s is 0.931. The prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ with the maximum similarity of attractors in plausible predictions has a possibility to be used for selecting a representative

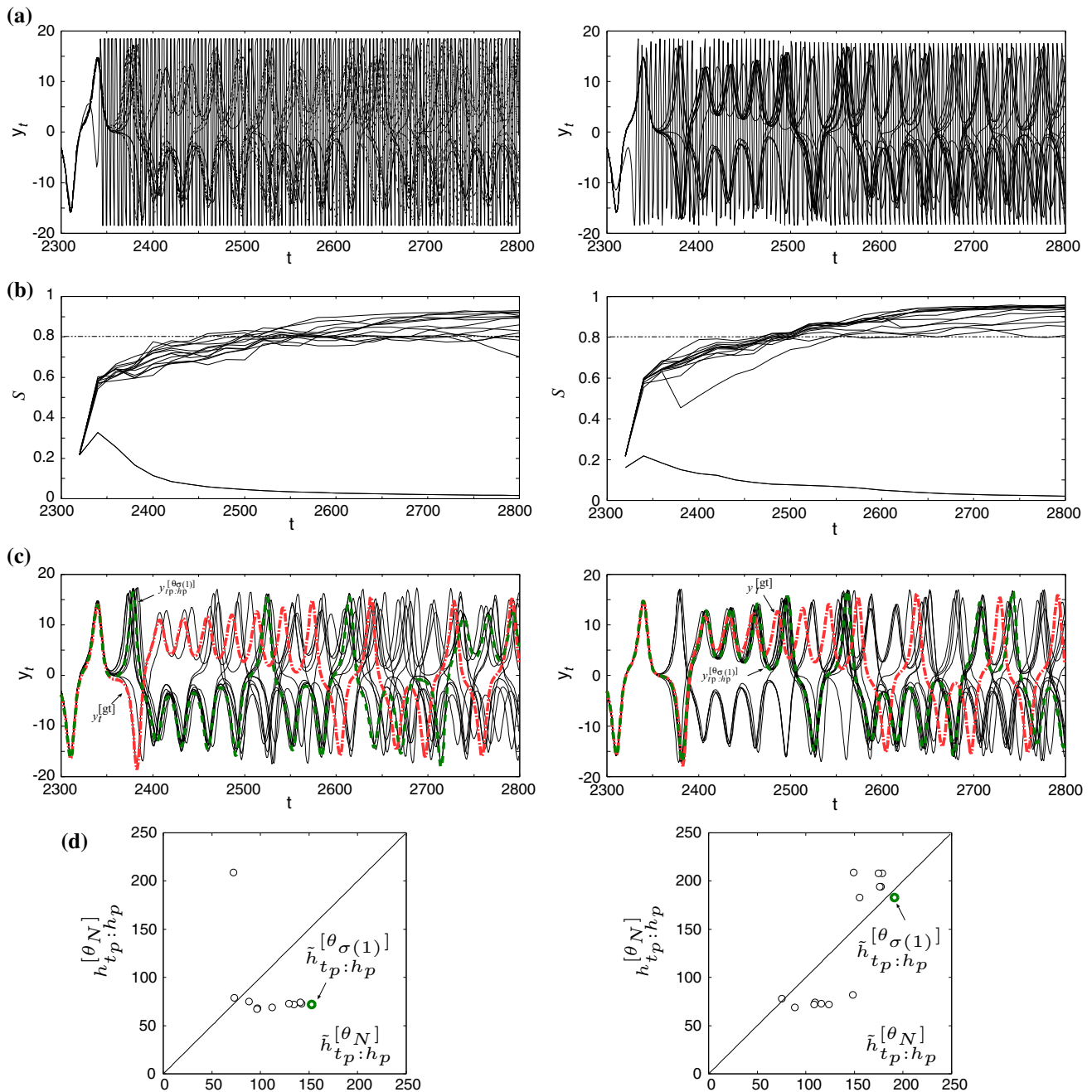


Fig. 2 Experimental results obtained by single CAN2s (left) and bagging CAN2s (right) for the prediction start time $t_p = 2300$ and the horizon $h_p = 500$. The top row, **a**, shows superimposed original predictions $y_{t_p:h_p}^{[\theta_N]}$. **b** Shows time evolution of similarity S of attractors, and the predictions with $S \geq S_{th} = 0.8$ at $t = t_p + h_p - 1 = 2799$ are selected as plausible predictions. **c** Shows selected plausible

predictions $y_{t_p:h_p}^{[\theta_N]}$ as well as ground truth time series $y_t^{[gt]}$ (red) and representative prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ (green). **d** Shows the relationship between actual predictable horizons $h_{t_p:h_p}^{[\theta_N]}$ and LOOCV predictable horizons $\tilde{h}_{t_p:h_p}^{[\theta_N]}$ of plausible predictions (colour figure online)

prediction, where $\theta_{\sigma(1)}$ indicates the learning machine with the maximum similarity. The comparison between $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ and $h_{t_p:h_p}^{[\theta_N]}$ is shown in Fig. 5a, where $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ seems competitive with $h_{t_p:h_p}^{[\theta_N]}$ for single CAN2, but worse for bagging

CAN2. To analyze much more, we have examined the correlation $r(S_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_N]})$ between the similarity $S_{t_p:h_p}^{[\theta_N]} = S(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[train]})$ and the predictable horizon $h_{t_p:h_p}^{[\theta_N]} = h(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[train]})$, as well as the correlation

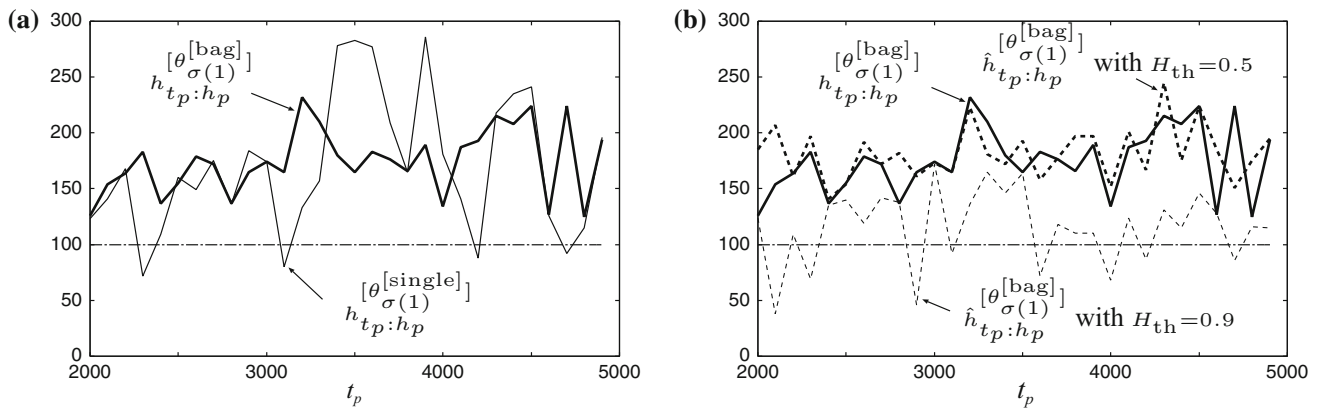


Fig. 3 Experimental result of **a** actual predictable horizons $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]}$ and $h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}$, and **b** estimated predictable horizon $\hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}$ with $H_{th} = 0.9$ and 0.5 for $t_p = 2300$. The mean of the predictable horizons is

$$\left\langle h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]} \right\rangle_{t_p \in T_p} = 170, \quad \left\langle h_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]} \right\rangle_{t_p \in T_p} = 175, \quad \left\langle \hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]} \right\rangle_{t_p \in T_p, H_{th}=0.9} = 115 \text{ and } \left\langle \hat{h}_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]} \right\rangle_{t_p \in T_p, H_{th}=0.5} = 182, \text{ respectively}$$

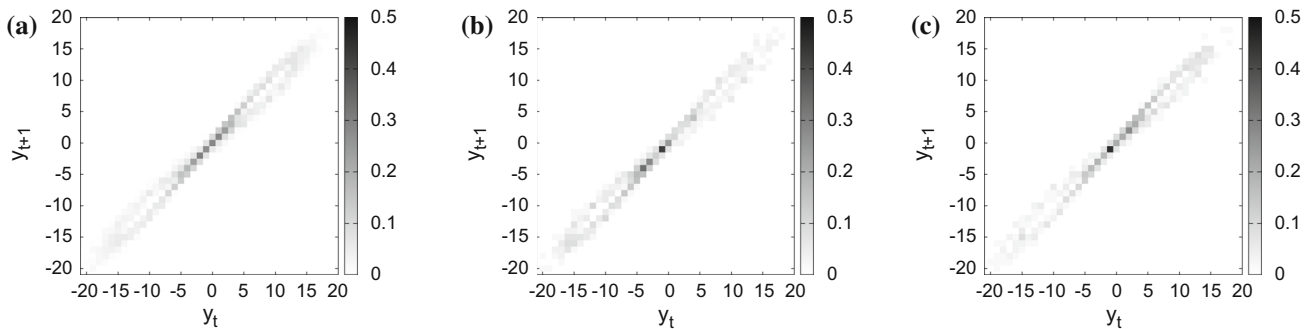


Fig. 4 Experimental result of attractor distribution: **a** $a_{ij}^{[rain]}$ of training time series $y_{t_g:h_g}^{[rain]}$, **b** $a_{ij}^{[\theta_{\sigma(1)}^{[single]}]}$ of the representative prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]}$ obtained by single CAN2 with $\sigma(1) = N = 145$, and **c** $a_{ij}^{[\theta_{\sigma(1)}^{[bag]}]}$

of the representative prediction $y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}$ obtained by bagging CAN2 with $\sigma(1) = N = 225$, at $t = 2799$. The resolution of the distributions is $\Delta_a = (v_{max} - v_0)/40 = (18.5 - (-18.5))/40 = 0.925$. The similarity $S(y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[single]}]}, y_{t_g:h_g}^{[rain]}) = 0.859$ and $S(y_{t_p:h_p}^{[\theta_{\sigma(1)}^{[bag]}]}, y_{t_g:h_g}^{[rain]}) = 0.939$

$r(\tilde{h}_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_{\sigma_S(1)}]})$ as shown in Fig. 5b. From this result, there are a number of cases with positive low or negative value of correlations. In particular, the correlation of similarity, $r(S_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_N]})$, has few cases with the values larger than 0.5 for both single and bagging CAN2. This suggests that the selection of representative prediction by using the similarity measure is not so reliable. On the other hand, bagging CAN2 has larger number of cases with the correlations larger than 0.5 as we can see the thick line of $r(\tilde{h}_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_{\sigma_S(1)}]})$ on the right-hand side in Fig. 5b. Furthermore, we can see that there are several cases of t_p with negative correlations $r(\tilde{h}_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_{\sigma_S(1)}]})$ in (b), and the

corresponding predictable horizons $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ in (a) are shorter than the neighboring (w.r.t. t_p) horizons. This correspondence seems reasonable because negative correlation does not contribute to the selection of the prediction with large predictable horizon. Thus, we have to remove the cases of negative correlations. So far, we have two approaches: one is to improve the performance of learning machine much more as we have done with the bagging method in this paper, and the other is to refine the selection method by means of modifying LOOCV predictable horizon or developing new methods. Actually, we have predictions with much longer predictable horizons not shown in this paper, but we cannot select such predictions without knowing the ground truth time series, so far.

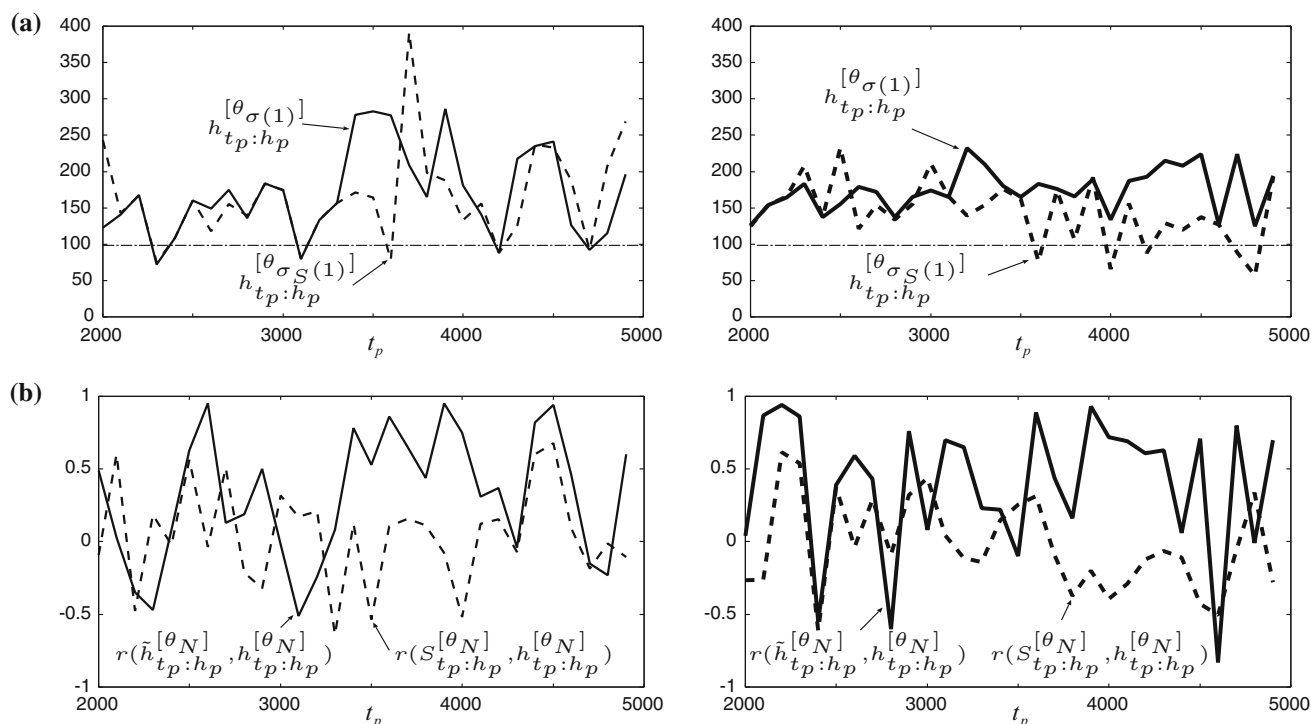


Fig. 5 Experimental result of **a** predictable horizons, $h_{t_p:h_p}^{[\theta_{\sigma(1)}]}$ and $h_{t_p:h_p}^{[\theta_{\sigma_S(1)}]}$, and **b** the correlations $r(\tilde{h}_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_N]})$ and $r(S_{t_p:h_p}^{[\theta_N]}, h_{t_p:h_p}^{[\theta_N]})$ for single CAN2s (left) and bagging CAN2s (right)

4 Conclusion

We have presented a performance improvement in the method for probabilistic prediction of chaotic time series by means of using bagging learning machines. The method obtains a set of plausible predictions by means of using similarity of attractors between training and predicted time series. And then, it provides representative prediction which has the longest LOOCV predictable horizon. By means of executing numerical experiments using single and bagging CAN2s, we have shown that bagging CAN2 improves the performance of single CAN2 and analyzed the relationship between LOOCV and actual predictable horizons. In our future research studies, we would like to overcome the problem of negative correlation between the achieved predictable horizon and the LOOCV predictable horizon, or the measure of selecting representative prediction.

Compliance with ethical standards

Conflict of interest The authors declare no conflicts of interest associated with this article.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use,

distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aihara K (2000) Theories and applications of chaotic time series analysis. Sangyo Tosho, Tokyo
- Lendasse A, Oja E (2004) Time series prediction competition: the cats benchmark. Proc IJCNN 2004:1615–1620
- Kurogi S, Ueno T, Sawa M (2007) Time series prediction of the CATS benchmark using Fourier bandpass filters and competitive associative nets. Neurocomputing 70(13–15):2354–2362
- Kurogi S, Tanaka S, Koyama R (2007) Combining the predictions of a time series and the first-order difference using bagging of competitive associative nets. In: Proceedings of the European symposium on time series prediction (ESTSP) 2007, pp 123–131
- Kurogi S, Ono K, Nishida T (2013) Experimental analysis of moments of predictive deviations as ensemble diversity measures for model selection in time series prediction. In: Proceedings of ICONIP, (2013) Part III, LNCS 8228. Springer, Heidelberg
- Kurogi S, Shigematsu R, Ono K (2014) Properties of direct multi-step ahead prediction of chaos time series and out-of-bag estimate for model selection. In: Proceedings of ICONIP2014, Part II, LNCS 8835. Springer, Heidelberg
- Kurogi S, Toidani M, Shigematsu R, Matsuo K (2015) Prediction of chaotic time series using similarity of attractors and LOOCV predictable horizons for obtaining plausible predictions. In: Proceedings of ICONIP 2015, LNCS 9491, pp 72–81

8. Slingo J, Palmer T (2011) Uncertainty in weather and climate prediction. *Phil Trans R Soc A* 369:4751–4767
9. Quiñonero-Candela J, Rasmussen CE, Sinz FH, Bousquet Q, Schölkopf B (2006) Evaluating Predictive Uncertainty Challenge. In: Quiñonero-Candela J et al (eds) *MLCW 2005*, LNAI 3944. Springer, Heidelberg, pp 1–27
10. Kurogi S, Sawa M, Tanaka S (2006) Competitive associative nets and cross-validation for estimating predictive uncertainty on regression problems. *Lecture Notes on Artificial Intelligence (LNAI) 3944*:78–94
11. Breiman L (1996) Bagging predictors. *Mach Learn* 26:123–140
12. Kurogi S (2009) Improving generalization performance via out-of-bag estimate using variable size of bags. *J Jpn Neural Netw Soc* 16(2):81–92
13. Efron B, Tibshirani R (1997) Improvements on cross-validation: the 632+ bootstrap method. *J Am Stat Assoc* 92:548–560