Hierarchical Clustering of Ensemble Prediction Using LOOCV Predictable Horizon for Chaotic Time Series

Shuichi Kurogi
Department of Control Engineering
Kyushu Institute of Technology
Kitakyushu, Japan
kuro@cntl.kyutech.ac.jp

Naoto Shimoda
Department of Control Engineering
Kyushu Institute of Technology
Kitakyushu, Japan
shimoda@kurolab.cntl.kyutech.ac.jp

Kazuya Matsuo
Department of Control Engineering
Kyushu Institute of Technology
Kitakyushu, Japan
matsuo@cntl.kyutech.ac.jp

***Abstract*—Recently, we have presented a method of ensemble prediction of chaotic time series. The method employs strong learners capable of making predictions with small error, where usual ensemble mean does not work well owing to the long term unpredictability of chaotic time series. Thus, we have developed a method to select a representative prediction from a set of plausible predictions by means of using LOOCV (leave-one-out cross-validation) measure to estimate predictable horizon. Although we have shown the effectiveness of the method, it sometimes fails to select the representative prediction with long predictable horizon. In order to cope with this problem, this paper presents a method to select multiple candidates of representative prediction by means of employing hierarchical K-means clustering with $K = 2$. From numerical experiments, we show the effectiveness of the method and an analysis of the property of LOOCV predictable horizon.**

***Index Terms*—hierarchical clustering of predictions, ensemble prediction of chaotic time series, leave-one-out predictable horizon, long-term unpredictability,**

I. Introduction

So far, there have been a number of studies on time series prediction [1], [2], where our methods awarded 3rd and 2nd places in the competitions held at IJCNN'04 [3] and ESTSP'07 [4], respectively, as well as a number of other methods have utilized model selection methods evaluating the mean square prediction error (MSE) for holdout and/or cross-validation datasets. Furthermore, our method in [5] utilizes moments of predictive deviation as ensemble diversity measures for model selection, and achieves better performance from the point of view of MSE than the conventional holdout method. The method in [6] uses direct multi-step 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 of time after the predictable horizon (see [6] for the analysis and [1] for properties of chaotic time series).

Recently, we have presented a method of ensemble prediction of chaotic time series [7]–[9]. Here, from [10], [11], we can see that 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 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 [10] use ensemble mean to obtain representative forecast, our method in [7]–[9] selects representative individual prediction from a set of plausible predictions because our method employs strong learners capable of making predictions with small error and there are individual predictions showing better performance than ensemble mean.

Our method in [7]–[9] employs LOOCV (leave-one-out cross-validation) measure to estimate predictable horizon to select a representative prediction from plausible predictions generated by strong learning machines. Comparing with our previous methods embedding model selection techniques using MSE [5], [6], the method has an advantage that it selects a representative prediction for each start time of prediction. Furthermore, it has provided long predictable horizons on average, while it sometimes fails in selecting representative predictions with long predictable horizons from plausible predictions.

In order to cope with this problem, this paper presents a method to select multiple candidates of representative prediction which is expected to have long predictable horizon. Here, it is mentioned in [11], [12] that a useful tool to provide alternative scenarios, or representative deterministic forecasts, is clustering which automatically groups the ensemble members. However, the implementation of clustering is not so straightforward in chaotic time series prediction with long prediction horizon The present method employes hierarchical K-means clustering with $K = 2$ accompanied with stepwise

increase of prediction horizon, which enable us to execute multiclass classification of chaotic time series predictions with long prediction horizon. From a survey of clustering methods for time series data [13], we can say that the present method is a row-data-based method of clustering but has a novelty of stepwise increase of prediction horizon. By means of analyzing this method, we show the property and the validity of LOOCV predictable horizon. We show the method in **II**, experimental results and analysis in **III**, and the conclusion in **IV**.

II. Ensemble Prediction of Chaotic Time Series

A. IOS Prediction of Chaotic Time Series

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

$$y_t = r(\boldsymbol{x}_t) + e(\boldsymbol{x}_t), \tag{1}$$

where $r(\boldsymbol{x}_t)$ is a nonlinear target function of a vector $\boldsymbol{x}_t = (y_{t-1}, y_{t-2}, \cdots, y_{t-k})^T$ generated by the delay embedding with dimension k from a chaotic differential dynamical system (see [1] for chaotic timeseries). Here, y_t is obtained not analytically but numerically, and then y_t involves an error $e(\boldsymbol{x}_t)$ owing to an executable finite calculation precision. In general, a time series generated with higher precision has small prediction error for longer duration of time from the prediction start time. Thus, let a time series generated with a very high precision be ground truth time series $y_t^{[\mathrm{gt}]}$ (see **III** for details), while we execute the prediction shown below with standard 64 bit precision.

Let $y_{t:h} = y_t y_{t+1} \cdots y_{t+h-1}$ denote a time series with the initial time t and the horizon h. For a given and training time series $y_{t_\mathrm{g}:h_\mathrm{g}} (= y^{[\mathrm{train}]}_{t_\mathrm{g}:h_\mathrm{g}})$, we are supposed to predict succeeding time series $y_{t_\mathrm{p}:h_\mathrm{p}}$ for $t_\mathrm{p} \geq t_\mathrm{g} + h_\mathrm{g}$. Then, we make the training dataset $D^{[\mathrm{train}]} = \{(\boldsymbol{x}_t, y_t) \mid t \in I^{[\mathrm{train}]}\}$ for $I^{[\mathrm{train}]} = \{t \mid t_\mathrm{g} \leq t < t_\mathrm{g} + h_\mathrm{g}\}$ to train a learning machine. After the learning, the machine executes IOS (iterated one-step ahead) prediction by $\hat{y}_t = f(\boldsymbol{x}_t)$ for $t = t_\mathrm{p}, t_\mathrm{p+1}, \cdots$, recursively, where $f(\boldsymbol{x}_t)$ denotes prediction function of $\boldsymbol{x}_t = (x_{t1}, x_{t2}, \cdots, x_{tk})$ whose elements are given by $x_{tj} = y_{t-j}$ for $t - j < t_\mathrm{p}$ and $x_{tj} = \hat{y}_{t-j}$ for $t - j \geq t_\mathrm{p}$. Here, we suppose that y_t for $t < t_\mathrm{p}$ 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 $f(\boldsymbol{x}_t)$ with small error for a duration of time from the initial time of prediction by means of using strong learning machienes.

B. Generation of IOS Predictions by CAN2s

As a strong learning machiene, we use CAN2 (competitive associative net 2), or an artificial neural net for learning efficient piecewise linear approximation of nonlinear function by means of the following schemes (See [14] for details): A single CAN2 has N units. The jth unit has a weight vector $\boldsymbol{w}_j \triangleq (w_{j1}, \cdots, w_{jk})^T \in \mathbb{R}^{k \times 1}$ and an associative matrix (or a row vector) $\boldsymbol{M}_j \triangleq (M_{j0}, M_{j1}, \cdots, M_{jk}) \in \mathbb{R}^{1 \times (k+1)}$ for $j \in I^N \triangleq \{1, 2, \cdots, N\}$. The CAN2 after learning the training dataset $D^n = \{(\boldsymbol{x}_i, y_i) | y_i = r(\boldsymbol{x}_i) + e_i, i \in I^n\}$ approximates the target function $r(\boldsymbol{x}_i)$ by $\widehat{y}_i = \widetilde{y}_{c(i)} = \boldsymbol{M}_{c(i)} \widetilde{\boldsymbol{x}}_i$, where $\widetilde{\boldsymbol{x}}_i \triangleq (1, \boldsymbol{x}_i^T)^T \in \mathbb{R}^{(k+1) \times 1}$ denotes the (extended) input vector to the CAN2, and $\widetilde{y}_{c(i)} = \boldsymbol{M}_{c(i)} \widetilde{\boldsymbol{x}}_i$ is the output value of the $c(i)$th unit of the CAN2. The index $c(i)$ indicates the unit who has the weight vector $\boldsymbol{w}_{c(i)}$ closest to the input vector $\boldsymbol{x}_i$, or $c(i) \triangleq \underset{j \in I^N}{\operatorname{argmin}} \|\boldsymbol{x}_i - \boldsymbol{w}_j\|$. The above function approximation partitions the input space $V \in \mathbb{R}^k$ into the Voronoi (or Dirichlet) regions $V_j \triangleq \{\boldsymbol{x} \mid j = \underset{i \in I^N}{\operatorname{argmin}} \|\boldsymbol{x} - \boldsymbol{w}_i\|\}$ for $j \in I^N$, and performs piecewise linear prediction for the function $r(\boldsymbol{x})$. We use the learning algorithm shown in [14] whose high ability of learning has been shown in Evaluating Predictive Uncertainty Challenge [15].

We make a number of IOS predictions $\hat{y}_{t_p:h_p}$ by means of using CAN2s with different parameter values. Namely, we use multiple $\hat{y}_{t_p:h_p} = y^{[\theta_N]}_{t_p:h_p}$ generated by different learning machines, where $\theta_N \in \Theta$ indicates a learning machine indexed by N, and Θ the set of learning machines.

C. Selecting Plausible Predictions Via Similarity of Attractors

Among $\hat{y}_{t_p:h_p} = y^{[\theta_N]}_{t_p:h_p}$ for $\theta_N \in \Theta$, there are implausible predictions which do not contribute to improve the accuracy of ensemble prediction. To remove such implausible predictions, we select the following set of plausible predictions (see [7] for illustration of the method):

$$\begin{aligned} Y^{[\Theta_S]}_{t_p:h_p} &\triangleq \left\{ y^{[\theta_N]}_{t_p,h_p} \;\middle|\; \theta_N \in \Theta_S \right\} \\ &\triangleq \left\{ y^{[\theta_N]}_{t_p,h_p} \;\middle|\; S\left(y^{[\theta_N]}_{t_p,h_p}, y^{[\mathrm{train}]}_{t_p:h_p}\right) / S_{\max} \geq S_{\mathrm{th}}, \theta_N \in \Theta \right\} \end{aligned} \tag{2}$$

where $S_{\max} = \max\left\{ S\left(y^{[\theta_N]}_{t_p,h_p}, y^{[\mathrm{train}]}_{t_p:h_p}\right) \;\middle|\; \theta_N \in \Theta \right\}$, S_{th} is a threshold, and

$$S\left(y^{[\theta_N]}_{t_p,h_p}, y^{[\mathrm{train}]}_{t_p:h_p}\right) \triangleq \frac{\sum_i \sum_j a^{[\theta_N]}_{ij} a^{[\mathrm{train}]}_{ij}}{\sqrt{\sum_i \sum_j \left(a^{[\theta_N]}_{ij}\right)^2} \sqrt{\sum_i \sum_j \left(a^{[\mathrm{train}]}_{ij}\right)^2}} \tag{3}$$

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

$$a_{ij} = \sum_{s=t}^{t+h-1} \mathbf{1}\left\{ \left\lfloor \frac{y_s - v_0}{\Delta_a} \right\rfloor = i \wedge \left\lfloor \frac{y_{s+1} - v_0}{\Delta_a} \right\rfloor = j \right\}, \tag{4}$$

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, $\mathbf{1}\{z\}$ is an indicator function equals to 1 if z is true, and 0 otherwize, and $\lfloor \cdot \rfloor$ indicates the floor function.

D. LOOCV Predictable Horizon and Hierarchical Clustering for Multiple Representative Predictions

Although plausible predictions have almost the same attractors, they show different prediction error. As a measure of

prediction error, let us define predictable horizon between two predictions $y_{t_p:h_p}^{[\theta_N]}$ and $y_{t_p:h_p}^{[\theta_{N'}]}$ in $Y_{t_p:h_p}^{[\Theta_S]}$ as

$$h\left(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[\theta_{N'}]}\right) = \max\left\{h \mid \forall s \le h \le h_p; |y_{t_p+s}^{[\theta_N]} - y_{t_p+s}^{[\theta_{N'}]}| \le e_y\right\}, \quad (5)$$

where e_y indicates a threshold. Furthermore, we introduce LOO (leave-one-out) predictable horizon given by

$$\tilde{h}_{t_p}^{[\theta_N,\Theta_S]} = \left\langle h\left(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[\theta_{N'}]}\right)\right\rangle_{y_{t_p:h_p}^{[\theta_{N'}]} \in Y_{t_p:h_p}^{[\Theta_S]} \setminus \{y_{t_p:h_p}^{[\theta_N]}\}}. \quad (6)$$

Here, $\langle\cdot\rangle$ denotes the mean and the subscript indicates the range of the mean. We select representative prediction $y_{t_p:h_p}^{[\theta_{\sigma_1}]}$ with the longest LOO predictable horizon given by $\tilde{h}_{t_p}^{[\theta_{\sigma_1},\Theta_S]} = \max\left\{\tilde{h}_{t_p}^{[\theta_N,\Theta_S]} \mid \theta_N \in \Theta_S\right\}$ which we call LOOCV predictable horizon as introduced in [7]. Here, θ_{σ_1} indicates the learning machine which has generated the prediction with the maximum $\tilde{h}_{t_p}^{[\theta_{\sigma_N},\Theta_S]}$.

In order to select multiple candidates of representative prediction with long predictable horizon involved in $Y_{t_p:h_p}^{[\Theta_S]}$, we execute hierarchical binary clustering of $Y_{t_p:h_p}^{[\Theta_S]}$ into $Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]}$ for $L = 0, 1, \cdots, L_{\max}$ and $c = 0, 1, 2, \cdots, 2^L - 1$ by means of the algorithm shown in Fig. 1. Here, we do not execute further clustering of $\Theta_S^{[L,c]}$ with $|\Theta_S^{[L,c]}| \le 3$ because LOO predictable horizons for predictions less than 3 are not effective. Furthermore, we define

$$h_0^{[L,c]} \triangleq \min\left\{h\left(y_{t_p:h_p}^{[\theta_N]}, \left\langle y_{t_p:h_p}^{[\theta]}\right\rangle_{\theta\in\Theta_S^{[L,c]}}\right) \;\middle|\; \theta_N \in \Theta_S^{[L,c]})\right\} \quad (7)$$

which denotes the shortest predictable horizon between $y_{t_p:h_p}^{[\theta_N]}$ and the mean prediction $\left\langle y_{t_p:h_p}^{[\theta]}\right\rangle_{\theta\in\Theta_S^{[L,c]}}$ for $\theta_N \in \Theta_S^{[L,c]}$ (see Fig. 2). After the clustering, we obtain representative predictions $y_{t_p}^{\left[\theta_{\sigma_1}^{[L,c]},\Theta_S^{[L,c]}\right]}$ with the LOOCV predictable horizon $\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L,c]},\Theta_S^{[L,c]}\right]}$ for each cluster of predictions $Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]}$.

E. Analysis of LOOCV Predictable Horizons in Hierarchical Clusters

We have shown that the performance of the selection of representative prediction using LOOCV predictable horizon is better than an intuitive selection using the maximum similarity of attractors (see [7]–[9]). In [8], [9], we have shown that longer predictable horizons can be achieved by means of employing bagging method to improve the prediction performance of learning machines. However, the validity has not been clarified, yet, and it is examined here as follows: let $y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]}\right]}$, $y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]}\right]}$ and $y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]}\right]}$ are representative predictions with LOOCV predictable horizons $\tilde{h}_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]},\Theta_S^{[L,c]}\right]}$, $\tilde{h}_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L+1,c_0]}\right]}$ and $\tilde{h}_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L+1,c_1]}\right]}$, respectively. Here, c_0 and c_1 represent the clusters divided by K-means clustering with $K = 2$ from the cluster c. Let us suppose $h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]}\right]}, y_{t_p:h_p}^{[\mathrm{gt}]}\right) > h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]}\right]}, y_{t_p:h_p}^{[\mathrm{gt}]}\right)$ as shown in Fig. 2. Furthermore, we suppose $h_0^{[L,c]} \simeq h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]}\right]}, Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]}\right) \simeq h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]}\right]}, Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]}\right)$. Then, we have LOO predictable horizons for $\theta_{\sigma_1}^{[L+1,c_0]}$ and $\theta_{\sigma_1}^{[L+1,c_1]}$ as

$$\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L,c]}\right]} = \frac{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| \tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L+1,c_0]}\right]} + \left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right| h_0^{[L,c]}}{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| + \left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right|} \quad (8)$$

$$\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L,c]}\right]} = \frac{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| h_0^{[L,c]} + \left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right| \tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L+1,c_1]}\right]}}{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| + \left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right|} \quad (9)$$

Then, when the condition

$$d\left(\Theta_S^{[L+1,c_0]}, \Theta_S^{[L+1,c_1]}\right) \triangleq \frac{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| \left(\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L+1,c_0]}\right]} - h_0^{[L,c]}\right)}{\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right| \left(\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L+1,c_0]}\right]} - h_0^{[L,c]}\right)} > 1 \quad (10)$$

holds, we have $\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L,c]}\right]} \ge \tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L,c]}\right]}$ and then

$$y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]}\right]} = y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]}\right]} \quad (11)$$

because $\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L,c]}\right]}$ has the maximum value among $\tilde{h}_{t_p}^{\left[\theta_N,\Theta_S^{[L,c]}\right]}$ for all $\theta_N \in \Theta_S^{[L,c]}$. Here, $\left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}\right| > \left|Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}\right|$ and $\left(\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_0]},\Theta_S^{[L+1,c_0]}\right]} - h_0^{[L,c]}\right) > \left(\tilde{h}_{t_p}^{\left[\theta_{\sigma_1}^{[L+1,c_1]},\Theta_S^{[L+1,c_0]}\right]} - h_0^{[L,c]}\right)$ are expected to be satisfied when we use a set of sufficiently large number of sufficiently strong learning machines, where the actual predictable horizon $h\left(y_{t_p:h_p}^{[\theta_N]}, y_{t_p:h_p}^{[\mathrm{gt}]}\right)$ of a learning machine and the number of learning machines with larger predictable horizons, respectively, become larger with the increase of the strength (or prediction precision) of learning machines. An equivalent relationship between (10) and (11) can be written as follows; i.e. we have

$$y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L-1,\lfloor c/2\rfloor]}\right]} = y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]}\right]} \quad (12)$$

when

$$D^{[L,c]} > D^{[L,\bar{c}]}. \quad (13)$$

for $h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]}\right]}, y_{t_p:h_p}^{[\mathrm{gt}]}\right) > h\left(y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,\bar{c}]}\right]}, y_{t_p:h_p}^{[\mathrm{gt}]}\right)$. Here, $\bar{c} = 2\lfloor c/2\rfloor + 1$ if $c = 2\lfloor c/2\rfloor$ and $\bar{c} = 2\lfloor c/2\rfloor$ if $c = 2\lfloor c/2\rfloor + 1$,

Algorithm Hierarchical_Clustering $\left(Y_{t_p:h_p}^{[\Theta_S]}, L_{\max}\right)$

step 1: Set $L := 0$, $c := 0$, and $\Theta_S^{[L,c]} := \Theta_S$.

step 2: Execute binary clustering of $\Theta_S^{[L,c]}$ into $\Theta_S^{[L+1,2c]}$ and $\Theta_S^{[L+1,2c+1]}$ for each $c = 0$ to $2^L - 1$ as follows: if $|\Theta_S^{[L,c]}| \leq 3$, then set $\Theta_S^{[L+1,2c]} := \Theta_S^{[L,c]}$ and $\Theta_S^{[L+1,2c+1]} := \phi$. Otherwise, obtain two clusters of predictions $Y_{t_p:h_0^{[L,c]}}^{\left[\Theta_S^{[L+1,2c]}\right]}$ and $Y_{t_p:h_0^{[L,c]}}^{\left[\Theta_S^{[L+1,2c+1]}\right]}$ by means of applying K-means clustering with $K = 2$ to $Y_{t_p:h_0^{[L,c]}}^{\left[\Theta_S^{[L,c]}\right]}$. Here, the horizon $h_0^{[L,c]}$ is given by (7) in text and shown in Fig. 2.

step 3: Set $L := L + 1$. Go to **step 2** if $L \leq L_{\max}$.

step 4: Return $\left\{ Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]} \middle| L=1,\cdots,L_{\max}; c=0,1,2,\cdots,2^L-1 \right\}$.

Fig. 1. Hierarchical clustering of predictions in $Y_{t_p:h_p}^{[\Theta_S]}$.

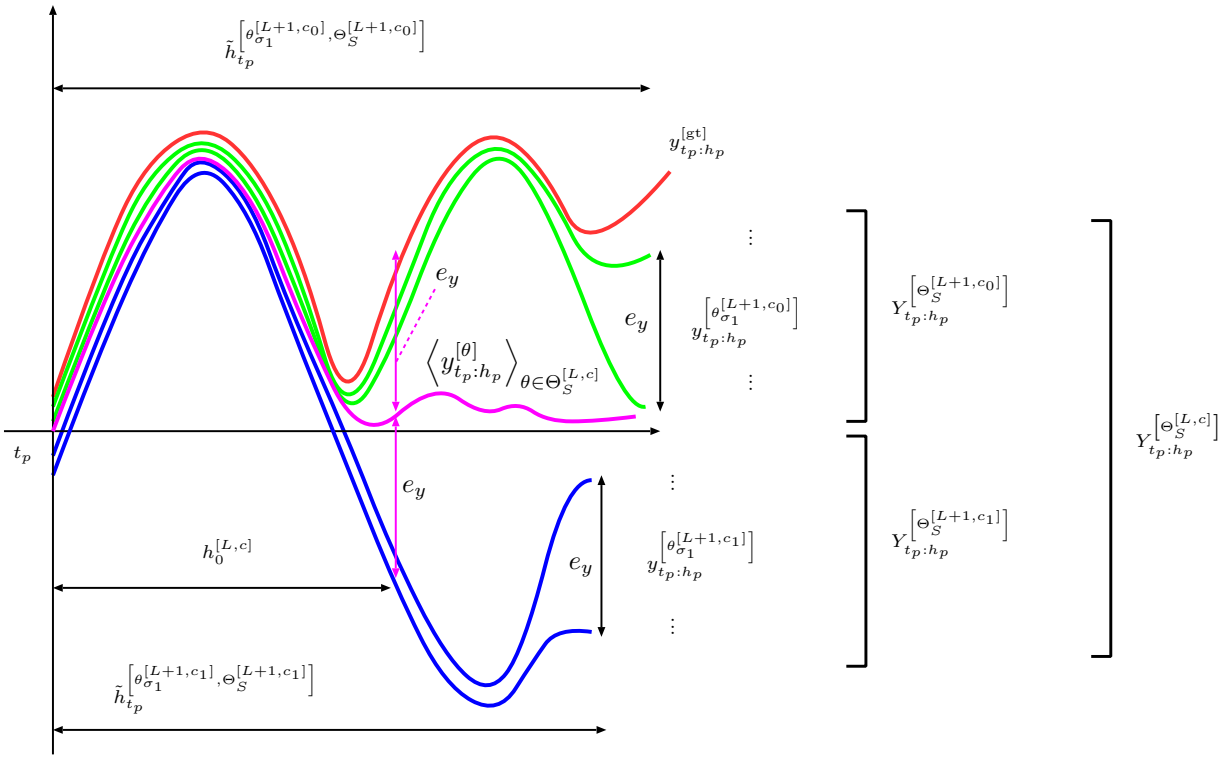


Fig. 2. Schematic illustration of binary clustering of predictions $Y_{t_p:h_p}^{\left[\Theta_S^{[L,c]}\right]}$ into $Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_0]}\right]}$ and $Y_{t_p:h_p}^{\left[\Theta_S^{[L+1,c_1]}\right]}$ via using the horizon $h_0^{[L,c]}$.

and

$$D^{[L,c]} = \frac{\prod_{l=1}^{L} d^{\left[l,\lfloor c/2^{L-l}\rfloor\right]}}{\sum_{c=0}^{2^L-1}\prod_{l=1}^{L} d^{[l,\lfloor c/2^{L-l}\rfloor]}} \tag{14}$$

$$d^{[l,c]} = \begin{cases} \frac{1}{1+d\left(\Theta_S^{[l,2\lfloor c/2\rfloor+1]},\Theta_S^{[l,2\lfloor c/2\rfloor]}\right)} & \text{if } c = 2\lfloor c/2\rfloor \\ \frac{1}{1+d\left(\Theta_S^{[l,2\lfloor c/2\rfloor]},\Theta_S^{[l,2\lfloor c/2\rfloor+1]}\right)} & \text{if } c = 2\lfloor c/2\rfloor + 1). \end{cases} \tag{15}$$

Here, $D^{[L,c]}$ holding $\sum_{c=0}^{2^L-1} D^{[L,c]} = 1$ for each L is a normalized version of $d(\cdot,\cdot)$ in (10) which reflects a degree of sufficient number of sufficiently long predictable horizons in the cluster of predictions, $Y_{t_p:h_p}^{\left[\Theta_S^{[L-1,\lfloor c/2\rfloor]}\right]}$.

The relationship between (12) and (13) indicates that the representative prediction with the LOOCV predictable horizon in a higher level of clusters will be also selected as the representative prediction with the LOOCV predictable horizon in the lower level of clusters when there are sufficiently large number of sufficiently strong learning machines. In other words, the representative prediction $y_{t_p:h_p}^{\left[\theta_{\sigma_1}^{[L,c]},\Theta_S^{[L,c]}\right]}$ with $[L,c] = [0,0]$ for the original set of plausible predictions $Y_{t_p:h_p}^{\left[\Theta_S^{[0,0]}\right]} = Y_{t_p:h_p}^{[\Theta_S]}$ will be the representative prediction in higher level clusters which are expected to have sufficiently

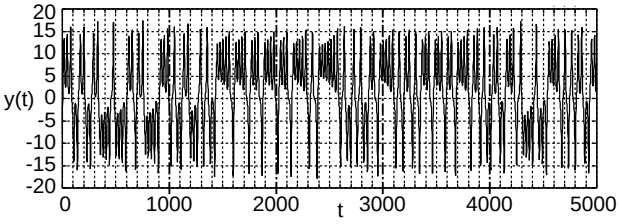


Fig. 3. Lorenz time series $y(t)$ for $t = 0, 1, 2, \cdots, 4999$, or ground truth time series $y^{[\mathrm{gt}]}_{0:5000}$.

long predictable horizon $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t}\right)$ when we use a set of sufficient number of sufficiently strong learning machines. When we do not have a set of sufficiently large number but several number (more than or equal to 3) of sufficiently strong learning machines, a higher level cluster involving them can select a representative prediction with long predictable horizon $h\left(y^{\left[\theta_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t}\right)$ by means of LOOCV predictable horizon, which is reflected by a small $D^{[L,c]}$ and a long predictable horizon $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t}\right)$ as shown in **III**.

III. Numerical Experiments and Analysis

A. Experimental Settings

We use the Lorenz time series, as shown in Fig. 3 and [6], obtained from the original differential dynamical system given by

$$\frac{dx_c}{dt_c} = -\sigma x_c + \sigma y_c, \tag{16}$$

$$\frac{dy_c}{dt_c} = -x_c z_c + r x_c - y_c, \tag{17}$$

$$\frac{dz_c}{dt_c} = x_c y_c - b z_c, \tag{18}$$

for $\sigma = 10$, $b = 8/3$, $r = 28$. Here, we use t_c for continuous time and $t\ (= 0, 1, 2, \cdots)$ for discrete time related by $t_c = tT$ with sampling time T. We have generated the time series $y(t) = x_c(tT)$ for $t = 1, 2, \cdots, 5000$ from the initial state $(x_c(0), y_c(0), z_c(0)) = (-8, 8, 27)$ with $T = 25$ms via Runge-Kutta method with 128 bit precision of GMP (GNU multi-precision library). We use $y_{0:h_{\mathrm{tr}}} = y_{0:2000}$ for training learning machines, and execute IOS prediction of $y_{t_{\mathrm{p}}:h_{\mathrm{p}}}$ with the initial input vector $\boldsymbol{x}_{t_{\mathrm{p}}} = (y(t_{\mathrm{p}} - 1), \cdots, y(t_{\mathrm{p}} - k))$ for prediction start time $t_{\mathrm{p}} = 2000 + 100i\ (i = 0, 1, 2, \cdots, 19)$ and prediction horizon $h_p = 500$. For learning machines, we have employed CAN2s (competitive associative nets) for learning piecewise linear approximation of nonlinear functions (see [14] for details).

B. Results and Analysis

For generating original predictions, we have used CAN2s with the number of units (or the number of piecewise linear regions) being $N = 5i\ (i = 1, 2, \cdots, 60)$. In Fig. 4. we show examples of original and selected predictions $\hat{y}^{[\theta_N]}_{t_p:h_p}$ for $t_p = 2300, 3100$ and 4700, which correspond to the cases with the achieved actual predictable horizon $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right)$ being smaller than 100 for $[L, c] = [0, 0]$ as shown in Fig. 5.

From Fig. 4 (b), we can see that plausible predictions selected by thresholding the similarity of attractors with $S_{\mathrm{th}} = 0.5$ involve predictions with predictable horizons shorter than 100. From (c), we can see that the representative prediction $y^{\left[\theta^{[0,0]}_{\sigma_1}\right]}_{t_p:h_p}$ (dark green) has predictable horizon $h\left(y^{\left[\theta^{[0,0]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right)$ smaller than 100. But, for $t_p = 2300$, 3100 in (e) and for $t_p = 4700$ in (d), we can see the predictions with predictable horizon larger than 100, which corresponds to the predictable horizons $h\left(y^{\left[\theta^{[0,0]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right) = 227$, 167, 180 in the cell for $(t_p, L) = (2300, 2)$, $(3100, 2)$ and $(4700, 1)$, respectively, in Fig. 5(b).

From Fig. 5(a), we can see that predictable horizon larger than 100 is achieved by the maximum of predictable horizons of representative predictions in hierarchical clusters not for $L = 0$ but for $L = 0, 1, 2$. In the two cells for $(t_p, L) = (4700, 1)$, we have $D^{[L,c]}(= 0.38) < D^{[L,\bar{c}]}(= 0.62)$ for $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right) = 180 > h\left(y^{\left[\theta^{[L,\bar{c}]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right) = 95$, which does not satisfy the condition (13) and indicates that we do not have sufficient number of sufficiently long predictable horizons in the cluster of predictions $Y^{\left[\Theta^{[L-1,\lfloor c/2 \rfloor]}_{S}\right]}_{t_p:h_p} = Y^{\left[\Theta^{[0,0]}_{S}\right]}_{t_p:h_p}$. Furthermore, it is difficult to select the maximum predictable horizon without knowing $y^{[\mathrm{gt}]}_{t_p:h_p}$. One of the solution is using much stronger learning machines as shown in [8], [9], and other solutions may be obtained by using further information of ground truth time series, which is for our future research studies.

IV. Conclusion

We have presented a method using hierarchical clustering and LOOCV predictable horizon for ensemble prediction of chaotic time series. The method selects multiple candidates of representative prediction with long predictable horizons, and enables us to analyze the property of LOOCV predictable horizons. By means of executing numerical experiments using CAN2s, we have shown that the method is able to select the candidates of representative predictions with long predictable horizons. In our future research studies, we would like to introduce probabilistic prediction using the candidates of representative predictions selected by the LOOCV predictable horizon.

References

[1] K. Aihara, "Theories and applications of chaotic time series analysis," Sangyo Tosho, Tokyo, 2000

[2] A. Lendasse, E. Oja, "Time series prediction competition: the cats benchmark," Proc. of IJCNN2004, pp.1615–1620, 2004

[3] S. Kurogi, T. Ueno, M. Sawa, "Time series prediction of the CATS benchmark using Fourier bandpass filters and competitive associative nets," Neurocomputing, Vol. 70, No.13–15, pp.2354–2362, 2007

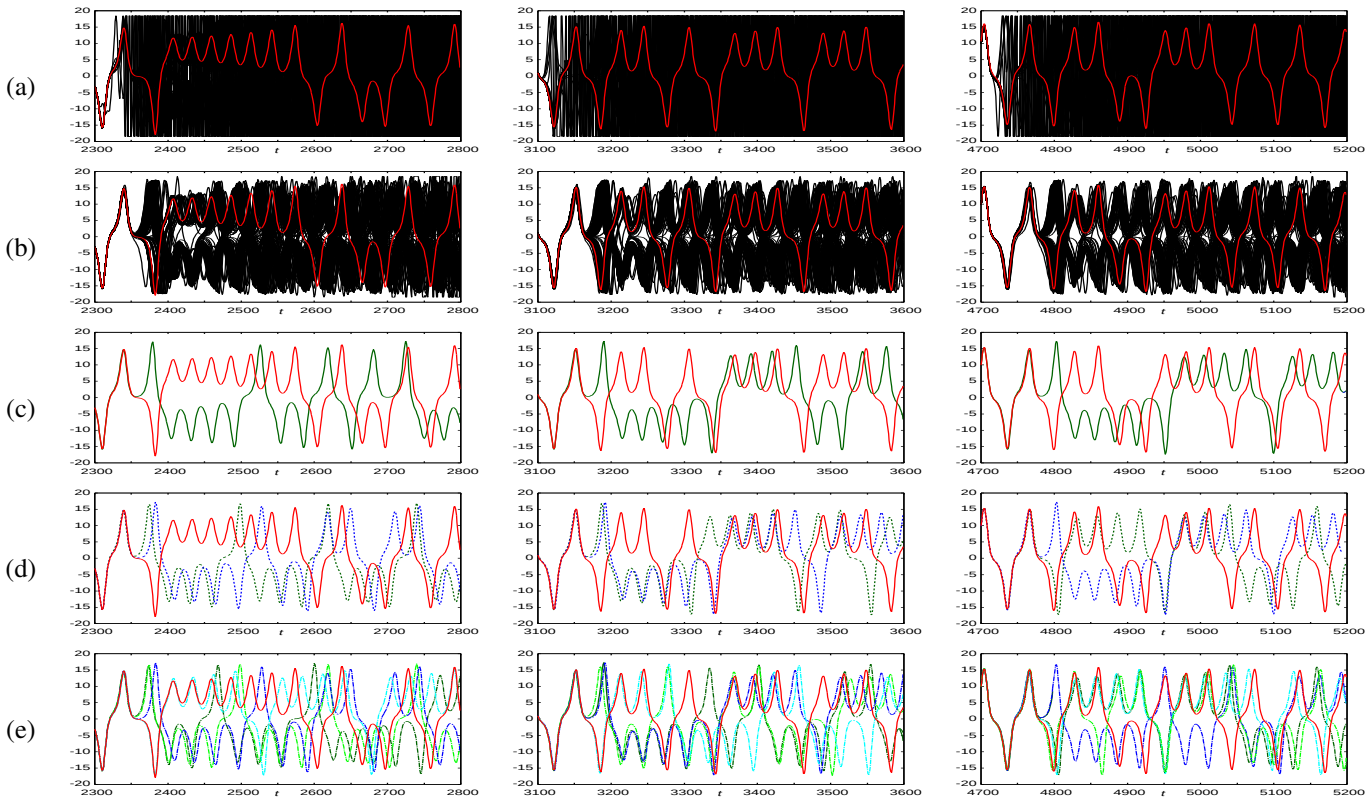


Fig. 4. Experimental results of $y^{[\mathrm{gt}]}_{t_p:h_p}$ (red) and superimposed predictions $\hat{y}^{[\theta_N]}_{t_p:h_p}$ obtained for $t_p = 2300$ (left), $t_p = 3100$ (middle), $t_p = 4700$ (right) and $h_p = 500$. (a) shows all original predictions $\hat{y}^{[\theta_N]}_{t_p:h_p}$, (b) plausible predictions selected with the similarity threshold $S_{\mathrm{th}} = 0.5$, and (c), (d) and (e) the representative predictions for the hierarchical level $L = 0, 1, 2$, respectively. The threshold of the error for predictable horizon is set $e_y = 15$.

[4] S. Kurogi, S. Tanaka, R. Koyama, "Combining the predictions of a time series and the first-order difference using bagging of competitive associative nets," Proc. of the European Symposium on Time Series Prediction (ESTSP) 2007, pp.123–131, 2007

[5] S. Kurogi, K. Ono, T. Nishida, "Experimental analysis of moments of predictive deviations as ensemble diversity measures for model selection in time series prediction," Proc. of ICONIP2013, Part III, LNCS 8228, pp.557–565, Springer, Heidelberg, 2013

[6] S. Kurogi, R. Shigematsu, and K. Ono, "Properties of direct multi-step ahead prediction of chaos time series and out-of-bag estimate for model selection," Proc ICONIP2014, Part II, LNCS 8835, pp.421–428. Springer, Heidelberg, 2014

[7] S. Kurogi, M. Toidani, R. Shigematsu, K. Matsuo, "Prediction of chaotic time series using similarity of attractors and LOOCV predictable horizons for obtaining plausible Predictions," Proc. of ICONIP2015, LNCS 9491, pp.72–81, 2015

[8] M. Toidani, K. Matsuo,S. Kurogi, "Performance Improvement via bagging in ensemble prediction of chaotic time series Using similarity of attractors and LOOCV predictable horizon," ICONIP2016, Part IV, LNCS 9950, pp.590–598, 2016

[9] S. Kurogi, M. Toidani, R. Shigematsu, K. Matsuo, "Performance improvement via bagging in probabilistic prediction of chaotic time series using similarity of attractors and LOOCV predictable horizon," Neural Computing and Applications Journal, DOI 10.1007/s00521-017-3149-7, July 15, 2017

[10] J. Slingo, T. Palmer, "Uncertainty in weather and climate prediction," Phil. Trans. R. Soc. A Vol. 369, pp.4751–4767, 2011

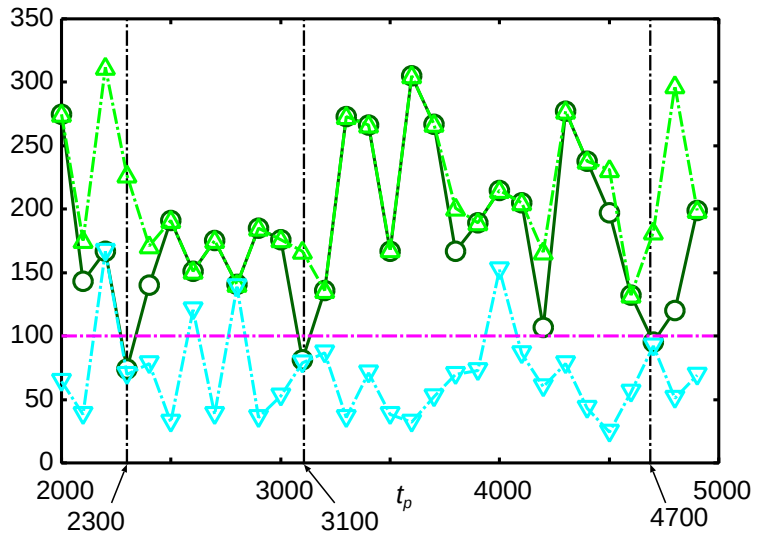
[11] WMO, "Guidelines on Ensemble Prediction Systems and Forecasting," WMO-No.1091, World Meteorological Organization, Geneva, Switzerland, 2012

[12] M. Xue, S. Roy, S. Zobell, Y. Wan, C. Taylor, C. Wanke, "A stochastic spatiotemporal weather-impact simulator: Representative scenario selection," AIAA Aviation Technology, Integration, and Operations Conference, Virginia, 2011

[13] T.W. Liao, "Clustering of Time Series Data—A Survey," Pattern Recognition, Vol. 38, No. 11, pp.1857–1874, 2005

[14] S. Kurogi, M. Sawa, S. Tanaka, "Competitive associative nets and cross-validation for estimating predictive uncertainty on regression problems," Lecture Notes on Artificial Intelligence (LNAI), Vol. 3944, pp.78–94, 2006

[15] J. Quiñonero-Candela, C.E. Rasmussen, F.H. Sinz, Q. Bousquet, B. Schölkopf, "Evaluating Predictive Uncertainty Challenge," In J.Quiñonero-Candela et al. (Eds.): MLCW 2005, LNAI 3944, pp.1–27, Springer, Heidelberg, 2006



(a)

L	$t_p = 2300$				$t_p = 3100$				$t_p = 4700$			
0	74				81				95			
1	77 0.67		71 0.33		82 0.43		80 0.57		95 0.62		**180** 0.38	
2	77 0.32	**227** 0.18	71 0.27	70 0.23	82 0.32	**167** 0.18	79 0.15	80 0.35	92 0.12	96 0.38	**181** 0.18	**182** 0.32

(b)

Fig. 5. Experimental result of (a) predictable horizons $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right)$ for $(L,c) = (0,0)$ (dark green), the maximum (green) and the minimum (cyan) among the hierarchical level $L = 0, 1, 2$ and all clusters $c = 0, 1, \cdots, 2^L - 1$ for each $t_p = 2000, 2100, \cdots, 4900$. In (b), an integer number shows the predictable horizon $h\left(y^{\left[\theta^{[L,c]}_{\sigma_1}\right]}_{t_p:h_p}, y^{[\mathrm{gt}]}_{t_p:h_p}\right)$ for $t_p = 2300, 3100, 4700$, $L = 0, 1, 2$ and $c = 0, 1, \cdots, 2^L - 1$, and a dotted-decimal number indicates the degree $D^{[L,c]}$ of sufficient number of sufficiently long LOOCV horizons for $L = 1$ and 2.