Robust Echo State Networks Based on Correntropy Induced Loss Function

Yu Guo\textsuperscript{a,b,c}, Fei Wang\textsuperscript{a,b,c}, Badong Chen\textsuperscript{a,b,c}, Jingmin Xin\textsuperscript{a,b,c}

\textsuperscript{a}Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
\textsuperscript{b}National Engineering Laboratory for Visual Information Processing and Applications, Xi'an Jiaotong University, Xi'an 710049, China
\textsuperscript{c}Shaanxi Provincial Key Laboratory of Digital Technology and Intelligent Systems, Xi'an 710049, China

Abstract

In this paper, a robust echo state network with correntropy induced loss function (CLF) is presented. CLF is robust to outliers through the mechanism of correntropy which is widely applied in information theoretic learning. The proposed method can improve the anti-noise capacity of echo state network and overcome its problem of being sensitive outliers which are prevalent in real-world tasks. The echo state network with CLF inherits the basic architecture of echo state network, but replaces the commonly used mean square error (MSE) criterion with CLF. The stochastic gradient descent method is adopted to optimize the objective function. The proposed method is subsequently verified in nonlinear system identification and chaotic time-series prediction. Experimental results demonstrate that our method is robust to outliers and outperforms the echo state networks with Bayesian regression and Huber loss function.

Keywords: Echo state networks, Correntropy induced loss function, Robust to outliers, Nonlinear systems
1. Introduction

Recurrent neural networks (RNNs) are a class of artificial neural network with internal feedback connections. Owing to their internal memory through feedback, RNNs is successfully applied to modeling highly nonlinear system problems including system identification [1] [2] and time-series prediction [3] [4]. Mathematically, RNNs can approximate arbitrary nonlinear dynamical system with arbitrary precision in theory [5], which can obtain better performance compared with static feedforward neural networks [6] [7].

Echo state networks (ESN), a class of efficient RNNs, have been proposed to reduce the complexity of training process suffered by classical RNNs [8] [9]. The core structure of ESN is a fixed hidden layer with internal feedback, called the “dynamic reservoir”. The reservoir consists of a large number of randomly and sparsely connected neural units. In ESN, only simple output (readout) weights need training, which leads to an easy learning mechanism but also achieves results with high accuracy. A theoretical analysis of ESN was presented in [10]. The necessary and sufficient condition to construct dynamic reservoir was discussed in their work. ESNs are quite different from classical RNNs that stability can be maintained if the condition called “echo state property” is satisfied [8].

ESN has been applied in many tasks including chaotic time-series prediction [11] [12] [13], system identification and control [14] [15] [16] and speech pattern recognition [17] [18]. Many existing works focus on improving the performance of ESN by modifying the equation of dynamic reservoir [19] [11] [20], among which the leaky integrator echo state network (Leaky-ESN) is the most popular.

Traditional ESN and its variants usually optimize the well-known mean square error (MSE) and are rather sensitive to outliers (or impulsive noises). When training set containing large noises or outliers, ESN may obtain poor generalization performance by minimizing MSE criterion. The main reason is that the assumption about data with Gaussian distribution is required for MSE criterion. However, this assumption is not always valid in real-world application-
s. In order to improve the stability and robustness, the Tikhonov regularization known as ridge regression is used in traditional ESN. Regularization parameter $\lambda$ controls the balance between the quality of the approximation and the complexity of the approximation function which needs to be determined using a valid method. Bayesian regression techniques can be used to include regularization parameters in the estimation procedure [21]. Han and Mu [22] utilized the Bayesian framework to estimate the regularization parameter. However, Bayesian regression may also suffer from the drawback of lacking robustness [23], since it is based on the assumption of Gaussian noises. Li et al. [12] improved the robustness of Bayesian framework by replacing Gaussian likelihood with Laplacian. The Huber loss function is another well-known robust loss function [24], which is used to improve the robustness of ESN in [25]. The Huber loss function is quadratic in small residual but grows linearly for large residual, hence it obtains the robust capacity to outliers.

In this paper, we introduce the robust methods from the information theory to suppress the negative effect caused by outliers. In information theoretic learning (ITL) [26] correntropy is a useful local similarity measurement between two vectors [27], which defines an $l_2$ similarity measurement in reproducing kernel Hilbert space (RKHS) and corresponds a nonlinear distance in the original input space [28]. Compared with the global measure MSE, correntropy is local, which has a strong ability to handle non-Gaussian noise with large outliers [29, 30]. Due to the property of rejecting outliers, correntropy has been successfully applied to construct different cost functions in signal processing and machine learning. Yuan and Hu [31] presented a robust feature extraction framework by introducing correntropy and the half-quadratic optimization technique was employed to optimize the corresponding cost function. He et al. [32] proposed an improved principal component analysis method based on maximum correntropy criterion (MCC), which is rotationally invariant and robust to outliers. Liu et al. took advantage of correntropy in training process of linear regression. Compared with the MSE and minimum error entropy (MEE), MCC presented a better performance on the regression with outliers. Singh et al. [33, 34]
extended correntropy to a new cost function named C-loss, which is employed in training neural networks to alleviate the negative effect of outliers. The C-loss criterion is a robust nonlinear measure of similarity, which can further approximate different norms (from $l_0$ to $l_2$) of data. Recently, the MCC has also been successfully applied in adaptive filtering [29, 35].

In this work, we address the issue of robustness of ESN based on correntropy induced loss function that is a useful criterion to handle noise with large outliers, named CLF-based ESN (CESN). Compared with conventional ESNs with MSEs as loss function, CESN is a natural extension by replacing MSE loss function with CLF. We verified the performance of the proposed CESN model on simulations of different nonlinear dynamical signals.

The remainder of this paper is organized as follows. In Section 2, we briefly revisit the basics of standard ESN and Leaky-ESN. In Section 3, we discuss an efficient and robust ESN model based on CLF. First, definition and properties of correntropy are introduced. Then, a new robust model of ESNs used CLF as training loss function is proposed. In Section 4, we evaluated the proposed method on dynamical systems prediction and identification tasks, demonstrating the advantages of CLF-based ESN over conventional ESN. Finally, we summarize this work in Section 5.

2. Mathematical Description of the ESNs

2.1. Echo State Network

ESN as a novel recurrent neural network replaces the standard hidden layers of RNN with the reservoir. The structure of an ESN is depicted in Fig 1. It is composed of an input layer, a reservoir and an output layer. For ESN, the input weights and reservoir weights are initialized randomly, while output weights are determined analytically. Therefore, ESN presents faster learning than traditional RNN model. The state update and output equations of the
ESN can be written as follows:

\[
\begin{align*}
    x(n+1) &= f(W^x x(n) + W^{in} u(n+1) + W^{fb} y(n)) \\
    y(n) &= g(W^{out} x(n))
\end{align*}
\]  

where \(x(n)\) is the reservoir state vector with \(N\) dimension, \(u(n)\) and \(y(n)\) are \(K\) dimensional external input vector and the \(L\) dimensional output vector, \(W^x\) denotes the \(N \times N\) internal connection weight matrix of the reservoir, \(W^{in}\), \(W^{fb}\) and \(W^{out}\) denote the \(N \times K\) input weight matrix, \(N \times L\) feedback weight matrix and \(L \times (K + N)\) output weight matrix, respectively, \(f\) and \(g\) denote the activation function of reservoir units and output units.

For ESN, \(W^{in}\), \(W^{fb}\) and \(W^x\) are randomly determined. Then the nonlinear system can be converted to a linear system as follows \[36, 15\],

\[
Y = XW^{out}
\]  

Figure 1: Structure of the traditional ESN with inputs, outputs and feedback from the outputs back to the reservoir. Solid arrows indicate fixed, random weights and dotted arrows are trainable output weights.
where

\[ X = [x(n), x(n+1), ..., x(k + N - 1)]^T \]  \hspace{1cm} (3)

\[ Y = [y(n), y(n+1), ..., y(k + N - 1)]^T \]  \hspace{1cm} (4)

Thus, the optimal output matrix \( W_{\text{out}} \) can be determined as a linear regression problem, whose solution can be obtained by

\[ W_{\text{out}} = X^\dagger Y = (X^TX)^{-1}X^TY \]  \hspace{1cm} (5)

where \( X^\dagger \) is the Moore-Penrose generalized inverse and \( X^T \) is the transpose of \( X \).

2.2. Regularized ESN

In practice, the matrix \( X \) may become ill-conditioned due to its large dimension, which often leads to a bad estimation of \( W_{\text{out}} \) by pseudoinverse method. For such an ill-posed problem, the traditional Tikhonov regularization method can be utilized to improve the estimation performance. Then, the output matrix \( W_{\text{out}} \) can be determined by minimizing the objective loss function as follows:

\[ \| Y - XW_{\text{out}} \|^2 + \lambda \| W_{\text{out}} \|^2 \]  \hspace{1cm} (6)

where \( \lambda \) is a constant, \( \| \cdot \|_2 \) denotes the L2-norm of a vector.

The solution of (6) can be obtained as

\[ W_{\text{out}} = (\lambda I + X^TX)^{-1}X^TY \]  \hspace{1cm} (7)

where \( I \) is an identity matrix. \( \lambda \) in Equations (6) and (7) denotes the regularization parameter to balance the training accuracy and model complexity.

2.3. Leaky ESN

Different types of neuron units can be integrated into the general framework of echo state network. One of the popular improved ESNs is the leaky integrator echo state network. Leaky integrator units \[ \text{[11]} \] incorporate information of the previous reservoir states when updating current state \( x(n) \). ESN with leaky
neuron units can slow down dynamics in the reservoir. Then, the learning performance for slow dynamical systems can be improved. The reservoir states of Leaky-ESN are updated as following,

\[ x(n+1) = (1 - a\gamma)x(n) + \gamma f(W^x x(n) + W^{in} u(n+1) + W^{fb} y(n)) \]  

where \( a \) and \( \gamma \) are leaking rate parameters. Leaky-ESN model depends on the previous state \( x(n) \) with the factor \( (1 - a\gamma) \). The echo state property \[8\] can be hold if \( |\lambda|_{max}(W^x) = 1, a > 0, \gamma > 0 \) and \( a\gamma \leq 1 \) \[11\].

Jaeger et. al. \[11\] presented that the influence of each parameter can be distributed over the others for the parameters \( \gamma, a \) and \( \rho \). By setting \( a = 1 \), the following update equation of reservoir state \[37\] can be obtained,

\[ x(n+1) = (1 - \gamma)x(n) + \gamma f(W^x x(n) + W^{in} u(n+1) + W^{fb} y(n)) \]  

where the echo state property of model holds if \( 0 < \gamma \leq 1 \). The equation \[9\] corresponds to a low-pass filter on the reservoir states, where the parameter \( \gamma \) can control cutoff frequency of filter.

3. Correntropy Induced Loss Function based ESN

The basic concept of correntropy is introduced in this section. Moreover, a robust ESN model based on MCC is proposed.

3.1. Basis of Correntropy

Correntropy is a concept proposed for ITL, which is a similarity measure derived from the generalized correlation function of random processes \[26\]. Between two random variables \( A = [a_1, \ldots, a_n] \) and \( B = [b_1, \ldots, b_n] \), the correntropy is defined as

\[ V(A, B) = E[(\Phi(A), \Phi(B))] = E[\kappa_\sigma(A, B)] \]  

where \( E[\cdot] \) denotes the expectation operator, \( \kappa_\sigma(\cdot) \) is a kernel function satisfying Mercers theory \[38\] with kernel size \( \sigma \), \( \Phi(\cdot) \) is a nonlinear function induced by \( \kappa_\sigma(\cdot) \) which maps data from the input space to a high-dimensional RKHS and
\[ \langle \cdot, \cdot \rangle \] denotes the inner product. In practice, the joint probability density function is usually unknown and only a finite number of samples \( a_i, b_i \) are available. In this case, the empirical correntropy can be computed as:

\[
\hat{V}(A, B) = \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(a_i, b_i)
\]

(11)

Kernel methods are widely applied in machine learning as powerful and efficient tools. Gaussian kernel is a typical kernel used in correntropy:

\[
\kappa_\sigma(a_i, b_i) = \exp\left(-\frac{\|a_i - b_i\|^2}{2\sigma^2}\right)
\]

(12)

where \( \|\cdot\| \) is the Euclidean norm. In our work, we take Gaussian kernel as the default kernel function of correntropy. In functional analysis view, the inner product is determined by the kernel size parameter \( \sigma \). Since correntropy is always bounded for any distribution, it outperforms MSE when impulsive noises exist in training samples [39, 40].

3.2. Correntropy-induced loss function

C-loss is easily achieved by extending the definition of correntropy [33]. In [34], correntropy induced loss function was maximized to measure the similarity between the model output \( O \) and the label \( L \) in classification problems. The definition of CLF was given as follows:

\[
\text{CLF}(O, L) = \beta [1 - E(\kappa_\sigma(O, L))]
\]

(13)

where \( \beta = [1 - \exp(-\frac{1}{2\sigma^2})]^{-1} \). In practice, the sample estimator of CLF can be computed as

\[
\text{CLF}(O, L) = \beta \left[1 - \frac{1}{N} \sum_{i=1}^{N} \kappa_\sigma(o_i, l_i)\right]
\]

(14)

Let \( e_i = o_i - l_i \). Then one can evaluate the sensitivity of CLF with respect to the error \( e_i \). The derivative with Gaussian kernel can be computed as

\[
\frac{\partial \text{CLF}(e)}{\partial e_i} = -\frac{1}{\sigma^2} \cdot \exp\left(-\frac{e_i^2}{2\sigma^2}\right) e_i
\]

(15)
Figure 2: (a) Loss curves of CLF with respect to the error $e$ for different values of kernel size $\sigma$. (b) Derivative curves of CLF with different values of kernel size $\sigma$. Compared with MSE, large errors have less influence to the objective function when a proper value of $\sigma$ is selected for CLF criterion.
The curves of CLF and its derivative with different kernel sizes are plotted in Fig. 2. The derivative curve of the MSE is also plotted for comparison. Due to the linear property of derivative curves of the MSE, the derivative will be huge if a large error $O - L$ occurs when using MSE loss function. The error caused by outliers or impulsive noises lead serious influences to the weights of ESN model. In contrast, the negative effect of large errors can be alleviated by CLF since its derivatives are bounded when input variable becomes large. Consequently, compared with the MSE-based loss function, CLF may achieve superior performance in real-world tasks, especially when data is contaminated by outliers [41].

3.3. Training ESN based on CLF

To improve the robust performance of conventional ESNs based on MSEs, we extend the traditional model by replacing the MSE by CLF, which leads an efficient and robust loss function to train the ESN as following,

$$J_{CESN} = \min_{W^\text{out}} \sum_{p=1}^{N} \text{CLF}(t_p, y_p)$$  \hspace{1cm} (16)

where $t_p$ and $y_p$ denote the $p$th row of target matrix $T$ and output matrix $Y$ respectively. $Y$ can be calculate using equation [2].

Many approaches can be applied to solve the optimization problem [16] in the literature. In this paper, stochastic gradient descent method is used for its simpleness and efficient computation. The output weights update is derived as follows,

$$w_{ij}^{\text{out}}(n + 1) = w_{ij}^{\text{out}}(n) - \eta \frac{\partial J_{CESN}(n)}{\partial w_{ij}^{\text{out}}(n)}$$

$$= w_{ij}^{\text{out}}(n) - \eta \frac{\partial J_{CESN}(n)}{\partial y_j(n)} g'(z_j(n)) x_i(n)$$  \hspace{1cm} (17)

where $\eta$ is the learning rate parameter. The equation (17) can be rewritten as follows for brevity,

$$w_{ij}^{\text{out}}(n + 1) = w_{ij}^{\text{out}}(n) - \eta \delta_j(n) x_i(n)$$  \hspace{1cm} (18)
where
\[ \delta_j(n) = \frac{\partial J_{CESN}(n)}{\partial y_j(n)} g'(z_j(n)) \] (19)

The function \( g \) is the activation function of output layer, which is selected as identity in our experiments. Then, the term \( \delta_j(n) \) can be computed as

\[ \delta_j(n) = \frac{\partial J_{CESN}(n)}{\partial y_j(n)} g'(z_j(n)) = \frac{\partial \text{CLF}(t(n), y(n))}{\partial y_j(n)} \]
\[ = \beta_1 \cdot \exp\left(-\frac{(t(n) - y(n))^2}{2\delta^2}\right)(t(n) - y(n)) \] (20)

where \( \beta_1 = \beta / \delta^2 \). The equations (18) and (20) can be applied to update the ESN under CLF criterion.

**Algorithm 1** Weights updating algorithm of ESN using CLF criterion

**Input:** Input matrix \( U = (u_{p,i})_{N \times K} \), target matrix \( T = (t_{p,j})_{N \times L} \).

**Initialization:** Kernel parameters \( \delta \), update step size \( \eta \), number of reservoir units \( N \), maximum number of iterations \( I_m \), termination tolerance \( \epsilon \).

**Output:** The optimal weight matrix \( \tilde{W}^{\text{out}} \).

**Step 1:** Select appropriate parameters of dynamic reservoir including the size, the sparseness and the spectral radius of reservoir. Then randomly initialize reservoir matrix \( W^x \) the and input matrix \( W^{\text{in}} \).

**Step 2:** Compute the output matrix \( X \) of the reservoir layer using equations (1) or (9).

**Step 3:** Update the output matrix \( W^{\text{out}} \) under CLF criterion.

repeat

1. Compute gradient term \( \delta(n) \) by equation (20).
2. Update output matrix \( W^{\text{out}} \) by equation (18).

until \( |\text{CLF}(n) - \text{CLF}(n - 1)| < \epsilon \)

The whole procedure of training ESN with the CLF criterion is summarized in Algorithm 1. In the training procedure, the ESN is firstly constructed with
suitable parameters including size, sparseness and spectral radius of the reservoir. The outputs of reservoir $X$ can be obtained. Then the output weights $W_{\text{out}}$ of ESN with CLF can be computed stochastic gradient descent method. In practice, the kernel width $\delta$ of CLF and learning rate $\eta$ in equation 18 are decided empirically.

4. Experiments and Discussions

In this section, we evaluate the performance of CLF-based ESN (CESN) in terms of accuracy, generalization and robustness. ESN based on CLF criterion is verified over dynamical problems for human activity classification, chaotic time series prediction and nonlinear system identification. For comparison, regularized ESN (RESN), ESN with Bayesian regression (BESN) and Huber loss function (HESN) are also implemented.

The basic ESN model need select several parameters for reservoir including the size of neural units, the sparseness of internal connections and the spectral radius of the matrix $W_x$. We can follow the suggestions in [36] for the reservoir constructing. Firstly, the size of reservoir is usually chosen in the range of 50-1000, which can be much larger than hidden units of traditional RNN. Secondly, we select the value of the sparseness from 1% to 5%, which leads to many loosely coupled subsystems in reservoir. Finally, the spectral radius of $W_x < 1$ can keep ESN stable [8].

The experiments are conducted on a computer with Intel Core i5 CPU with 4 GB of RAM, running on Ubuntu Linux 14.04.

For evaluating the robustness of proposed method, outliers and noise adding to training data are modeled using gross error model:

$$D_\varrho = \{D|D = (1 - \varrho)G + \varrho F, 0 \leq \varrho \leq 1\}$$ (21)

where $\varrho$ determines the probability of occurrence of outliers, $G$ and $F$ are stochastic variables which occur with probabilities $1 - \varrho$ and $\varrho$ respectively. Different probabilities of outliers are added to training data with gross model. Then
 CESN is trained using the contaminated training data set. Finally, performance is evaluated using the testing data set which is free of outliers.

In the experiments, the normalized root mean squared error (NRMSE) between a target signal $t(n)$ and a output signal $y(n)$ for $n = 0, \ldots, N−1$ is used for evaluating the performances of comparison methods. NRMSE is calculated as

$$NRMSE = \sqrt{\frac{1}{N\sigma_{target}^2} \sum_{n=0}^{N-1} (t(n) - y(n))^2}$$

where $\sigma_{target}^2$ is the variance of target signal. For the test performance with outliers free data, the lower NRMSE implies the better generalization capacity.

4.1. Human activities classification

In this example, we employ CESN on the task of human activities classification and its performance is evaluated on the CAD-120 benchmark dataset. The benchmark dataset comprises 120 activity sequences which belong to 10
different high-level activities including making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects and having a meal. The high-level activities are performed by four different subjects performing each activity multiple times. Three-dimensional coordinates of 15 joints and ground-truth category labels of manipulated objects are used as inputs in this experiment.

Table 2: Experimental results obtained for nonlinear dynamical system with training data containing outliers. The outliers are generated by gross error model. Performances are measured by NRMSE on 20 Monte Carlo runs of testing data without outliers.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>RESN</th>
<th>BESN</th>
<th>HESN</th>
<th>CESN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>70.97</td>
<td>74.19</td>
<td>74.19</td>
<td><strong>77.42</strong></td>
</tr>
<tr>
<td>Precision</td>
<td>72.19</td>
<td>80.83</td>
<td>74.33</td>
<td><strong>83.50</strong></td>
</tr>
<tr>
<td>Recall</td>
<td>70.00</td>
<td>73.33</td>
<td>73.33</td>
<td><strong>76.67</strong></td>
</tr>
</tbody>
</table>

The ESN setup is given in Table 1. CESN is trained using videos of three subjects and tested by the data of an unseen subject. Noise and outliers are artificially added to training data. Performances are then evaluated by three criterion: accuracy, precision and recall.

Comparison results of different models trained by data with 10% outliers are provided in Table 2. This table shows that HESN has a close performance with BESN while BESN shows better precision. HESN and BESN are more robust than RESN. CLF-based ESN provides better performance on accuracy, precision and recall in the noisy condition.

In this experiment, models are only trained by the skeleton data and labels of manipulated objects. It is assumed the performance of classification can be improved when other information is introduced.
4.2. Mackey-Glass chaotic time series prediction

The Mackey-Glass time series has been recognized as one of the benchmark problems for assessing performance of learning algorithms. The Mackey-Glass is a delay differential equation and given with the following form

\[
\frac{dx}{dt} = ax(t) + \frac{bx(t - \tau)}{1 + x(t - \tau)^{10}}
\]

where \(x(t)\) is the state signal. The Mackey-Glass system has a chaotic attractor if \(\tau > 16.8\). The parameters in equation (23) are selected with \(a = -0.1\), \(b = 0.2\) and \(\tau = 17\). In a discrete manner, the Mackey-Glass delay differential equation can be approximated as:

\[
x(n + 1) = x(n) + \kappa(-0.1x(t) + \frac{0.2x(n - \tau/\kappa)}{1 + x(n - \tau/\kappa)^{10}})
\]

with a stepsize \(\kappa\). Here the stepsize is set to the same value \(\kappa = 1/10\) as in [8].

The goal in our experiment is to predict the state signal \(x(t + \Delta t)\) using historical information. This can be formulated as a function given by:

\[
x(t + \Delta t) = h(x(t), x(t - 6), x(t - 12), x(t - 18))
\]

where the delay time and the embedded dimension are determined by the mutual information [43] and the false nearest method [44]. The two parameters are 6 and 4 in our task.

For ESN setup in Table 1, the number of reservoir units and the spectral radius of \(W^x\) are selected as 700 and 0.02, respectively.

For each training sequence, 1000 data samples with zero-mean Gaussian noise (\(SNR = 0\)) are generated. Then, we add outliers to original training set with different levels including 15%, 25% and 35%. The prediction performance of CLF criterion in terms of NRMSE is tested over 20 Monte Carlo noiseless time series each with the length of 1000.

All the methods for performance evaluation are trained by the time series with outliers generated in the same condition. The performance results of training process are shown in figure [8]. From this figure, it can be observed performance of ESN is suffered by outliers badly. With the noise become serious,
Figure 3: Comparison of training performance of CESN with ESN, RESN, BESN and HESN. The original data is polluted by 15% (the first column), 25% (the second column) and 35% (the third column) of outliers for training. The CESN shows superior performance.
its performance is decreasing fast. With suitable regularization parameter $\lambda$, RESN presents better robustness than ESN. Training performances of HESN and BESN are more accurate than the previous two models. ESN trained by CLF shows similar training performance to HESN and BESN which is robust to training data with different levels of outliers.

NRMSEs of five methods for Mackey-Glass system are presented in Table 3. In this table, traditional ESN shows poor performance of prediction. As Huber loss function is robust to outliers, the accuracy of the corresponding prediction model is acceptable. In the HESN, $\eta$ is an important parameter for model training, according to which the training data are partitioned into two subsets, and different loss functions, such as the quadratic loss function and robust loss function are applied to each subset. When the levels of outliers are 15% and 25%, BESN has the similar prediction accuracy to HESN. With the outliers increasing, HESN has a better performance than BESN. CESN and HESN shows more stable performance in different levels of outliers. Compared with HESN, prediction NRMSE of CESN is lower than that of HESN.

Table 3: Experimental results obtained for Mackey-Glass chaotic system with training data containing outliers. The outliers are generated by gross error model. Performances are measured by NRMSE on 20 Monte Carlo runs of testing data without outliers.

<table>
<thead>
<tr>
<th>Noise</th>
<th>RESN</th>
<th>BESN</th>
<th>HESN</th>
<th>CESN</th>
</tr>
</thead>
<tbody>
<tr>
<td>p=15%</td>
<td>0.2383</td>
<td>0.1517</td>
<td>0.1550</td>
<td><strong>0.1259</strong></td>
</tr>
<tr>
<td>p=25%</td>
<td>0.2601</td>
<td>0.1587</td>
<td>0.1672</td>
<td><strong>0.1284</strong></td>
</tr>
<tr>
<td>p=35%</td>
<td>0.3406</td>
<td>0.2391</td>
<td>0.1682</td>
<td><strong>0.1457</strong></td>
</tr>
</tbody>
</table>

Figure 4 shows the convergence curves of the training process of CESN. We choose four different learning rates $\eta$ ($1e^{-5}$, $5e^{-6}$, $1e^{-6}$ and $5e^{-7}$) to test the convergence performance. One can clearly observe that the CESNs show
different convergence speed and steady-state. In this simulation, it is evident that CESN trained with $\eta = 1e^{-6}$ may achieve the desirable performance.

![Training curves of CESN with different learning rates](image)

Figure 4: Training curves of CESN with different learning rates: $\eta = 1e^{-5}$, $\eta = 5e^{-6}$, $\eta = 1e^{-6}$ and $\eta = 5e^{-7}$. In this experiment the learning rate is selected as $\eta = 1e^{-6}$.

4.3. Nonlinear dynamical system identification

An unknown nonlinear dynamical system is used for modeling its dynamics in this example. We evaluated the performance of CESNs trained by data with outliers. The dynamics of this nonlinear system is given as:

$$y(n + 1) = \frac{y(n)y(n-1)y(n-2)u(n-1)(u(n-2)-1) + u(n)}{1 + y^2(n-2) + y^2(n-1)}$$ (26)

where $u(n)$ is the control signal. 1000 data points are generated according to equation (26) for training, which are contaminated by outliers with $u(n)$ following uniform distribution $U [-2, 2]$. Testing set free of outliers and defined with control signal $u(n)$ as:

$$u(n) = \begin{cases} 
\sin(3\pi n/250), & n \leq 500 \\
0.25 \sin(2\pi n/250) + 0.2 \sin(3\pi n/50), & n > 500 
\end{cases}$$ (27)
The goal of this task is to build dynamics identification model of system equation (26) using CESN to suppress the negative effect of outliers. The input of system is \( \mathbf{x}(n) = [y(n), u(n)] \), whereas output is given as \( \mathbf{t}(n) = [y(n + 1)] \).

For all of the time series, an echo state network with 300 neural units in the reservoir is trained. Specifications of the ESN setup are given in Table 1. The length of the training sequence is selected as 900, in which the first 100 data are eliminated. As in previous nonlinear systems, outliers then artificially added to the training data. Testing sequences with 1000 samples each are used to for performance measurement. Figure 5 shows experimental results averaged over 20 Monte Carlo runs. Experiments are performed with four different percents of training data contaminated by outliers, i.e. 15%, 25%, and 35% of outliers. As it can be seen, CESN obtains higher regression accuracy than ESN, RESN, HESN and BESN while BESN shows comparable performance with CESN. It can be inferred that CESN is able to learn the inherent mechanism of system equation (26) in a noisy situation. Further, CESN shows stable performance in different percentages of outliers in the data. Table 4 presents the details of performances of different methods. It can be seen that ESN, RESN and HESN obtained the higher NRMSE compared with BESN and CESN. The results also illustrate that models trained by CLF show a better robustness in different levels of training data contamination by outliers. We also show the total training time of 20 Monte Carlo runs of different methods in Table 4. It can be observed that methods using regularization (RESN, BESN) consume less computational resource but are suffered by outliers. HESN costs much training time while obtaining unsatisfied results. In contrast, CESN can provide better performance with acceptable computational time.

In order to achieve good performance of ESN, the selection of parameters including the size of the reservoir and the sparseness of matrix \( \mathbf{W}^x \) are important. Then, with different values of the two parameters, we evaluated the performances of proposed CESN compared against traditional RESN, BESN and HESN. Figure 6(a) shows the average test errors of four methods with different sizes of the reservoir. For each size, 20 Monte Carlo runs of simulations
Figure 5: Testing data for Nonlinear dynamical system identification and 5 methods estimate given training data polluted by (a). 15 percent (the left column), (b). 25 percent (the middle column), (c). 35 percent (the right column) of outliers. Test data is absent of outliers.

Table 4: Experimental results obtained for nonlinear dynamical system with training data containing outliers. The outliers are generated by gross error model. Performances are measured by NRMSE on 20 Monte Carlo runs of testing data without outliers.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training time(s)</th>
<th>NRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>p=15%</td>
<td>p=25%</td>
</tr>
<tr>
<td>ESN</td>
<td>0.87</td>
<td>2.5472</td>
</tr>
<tr>
<td>RESN</td>
<td>0.91</td>
<td>0.6065</td>
</tr>
<tr>
<td>BESN</td>
<td>2.68</td>
<td>0.2443</td>
</tr>
<tr>
<td>HESN</td>
<td>51.82</td>
<td>0.4899</td>
</tr>
<tr>
<td>CESN</td>
<td>18.06</td>
<td>0.1997</td>
</tr>
</tbody>
</table>
with 35% of outliers are tested. From figure 6(a) it can be observed that the proposed BESN and CESN achieves better accuracy and robustness than those of RESN and HESN with different values of size of the reservoir. One can also find that the testing performance of CESN is a little better than BESN in most sizes of the reservoir. Figure 6(b) shows the test results of four different models with respect to the values of sparseness. With the values of sparseness increasing, it is observed that NRMSEs of RESN and HESN change drastically from 0.01-0.04. On the contrary, test errors of CESN and BESN keep stable with the value of sparseness changing. Moreover, the proposed CESN performs better than the other methods. In figure 6, the results of ESN is not plotted due to their much larger test errors.

The kernel sizes $\sigma$ is another important parameter that can influence the performance of proposed CESN. A suitable selection of $\sigma$ is very important for CLF loss function. The performance of identification task affected by different kernel sizes is shown in Fig 7. One can find that if $\sigma$ is too small, the model produces large NRMSEs. It is also observed that the NRMSEs are tending toward stability with $\sigma$ increasing. In this example, CESN obtains the best performance for system identification when $\sigma = 3.0$. The suitable selection of
Figure 7: Performance of CESN with different values of $\sigma$. If $\sigma$ is too small, the test NRMSE is large. The NRMSEs are tending toward stability with $\sigma$ increasing. In this example, CESN obtains the best performance for system identification when $\sigma = 3.0$.

Kernel width should depend on the application scenario. The method for setting optimal kernel size is an open issue for CESN.

5. Conclusion

For real-world nonlinear dynamical systems, outliers often cause harmful influence to echo state network and lead to its poor generalization ability. In this paper, we propose an efficient and robust echo state network by replacing the well-known MSE loss function with CLF. CLF has a clear theoretical foundation and is robust to outliers, which can be used to enhance the generalization performance of traditional ESN. The proposed model combines the advantages of both ESN and the CLF criterion. Experimental results on three different tasks have shown that the ESNs training with CLF criterion can achieve higher generalization and robustness performance in comparison with traditional ESN,
regularized ESN, ESN with Bayesian regression and Huber loss function. We also analyze the effects of different learning rates and kernel sizes to CESN. Online CESN with regularization for certain tasks is an interesting topic for future research.

Acknowledgements

This work is partially supported by 973 Program (No.2015CB351703), Program of introducing talents of discipline to university (No.B13043) and National Natural Science Foundation of China (No.61231018). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References


[16] G. K. Venayagamoorthy, Online design of an echo state network based wide area monitor for a multimachine power system, Neural Networks 20 (3) (2007) 404 – 413.


