ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Negatively Correlated Echo State Networks

Ali Rodan and Peter Tiňo

School of Computer Science, The University of Birmingham
Birmingham B15 2TT, United Kingdom
E-mail: {a.a.rodan, P.Tino}@cs.bham.ac.uk

Abstract. Echo State Network (ESN) is a special type of recurrent neural network with fixed random recurrent part (reservoir) and a trainable reservoir-to-output readout mapping (typically obtained by linear regression). In this work we utilise an ensemble of ESNs with diverse reservoirs whose collective read-out is obtained through Negative Correlation Learning (NCL) of ensemble of Multi-Layer Perceptrons (MLP), where each individual MPL realises the readout from a single ESN. Experimental results on three data sets confirm that, compared with both single ESN and flat ensembles of ESNs, NCL based ESN ensembles achieve better generalisation performance.

1 Introduction

It has been extensively shown that ensemble learning can offer a number of advantages over a single learning machine (e.g. neural network) training. It has a potential to e.g. improve generalisation and decrease the dependency on training data [3]. One of the key elements for building ensemble models is the "diversity" among individual ensemble members. Negative correlation learning (NCL) [5] is an ensemble learning technique that encourages diversity among ensemble members through their negative correlation. It has been successfully applied in a number of applications, including regression problems [2], classification problems [7], or time series prediction using simple auto-regressive models [5].

In this paper we apply the idea of NCL to the ensemble of Echo State Networks (ESNs). Each ESN operates with a different reservoir, possibly capturing different features of the input stream. On each reservoir we build a non-linear readout mapping. Crucially, the individual readouts of the ensemble are coupled together by a diversity-enforcing term of the NCL training, which stabilises the overall collective ensemble output. There have been studies of simple ESN ensembles [9], or Multi-Layer Perceptron (MLP) readouts [1, 4], but to the best of our knowledge, this is the first study employing a NCL style training in ensembles of state space models, such as ESNs.

The paper has the following organisation: Section 2 gives a background on Echo State Network and Negative Correlation Learning. In Section 3 we introduce negatively correlated ensembles of ESNs. Experimental studies are presented in Section 4. Finally, our work is concluded in Section 5.

2 Background

Echo state Network (ESN) [6] (shown in Fig.1 (right)) is a discrete-time recurrent neural network with K input units, N recurrent (reservoir) units and L output units. The activation vectors of the input, internal, and output units at time step t are denoted by $s(t)$, $x(t)$, and $y(t)$, respectively. The connections between the input units and the recurrent units are given by an $N \times K$ weight matrix V, connections between the internal units are collected in an $N \times N$ weight matrix W.

The recurrent units are updated according to[1]:

$$x(t+1) = f(Vs(t+1) + Wx(t)), \tag{1}$$

where f is the reservoir activation function (tanh in this study). The readout is computed as:

$$y(t+1) = g(x(t+1)), \tag{2}$$

where g is the readout function and can either be linear (typical case for ESN), or non-linear (e.g. a MLP). The readout mapping can be trained in an offline or online mode by minimising the Mean Square Error, $MSE = \langle(\hat{y}(t) - y(t))^2\rangle$, where $\hat{y}(t)$ is the readout output, $y(t)$ is the desired output (target), and $\langle\cdot\rangle$ denotes the empirical mean.

Elements of W and V are fixed prior to training with random values drawn from a uniform distribution over a (typically) symmetric interval. The reservoir connection matrix W is typically scaled as $W \leftarrow \alpha W/|\lambda_{max}|$, where $|\lambda_{max}|$ is the spectral radius of W and $0 < \alpha < 1$ is a scaling parameter [6].

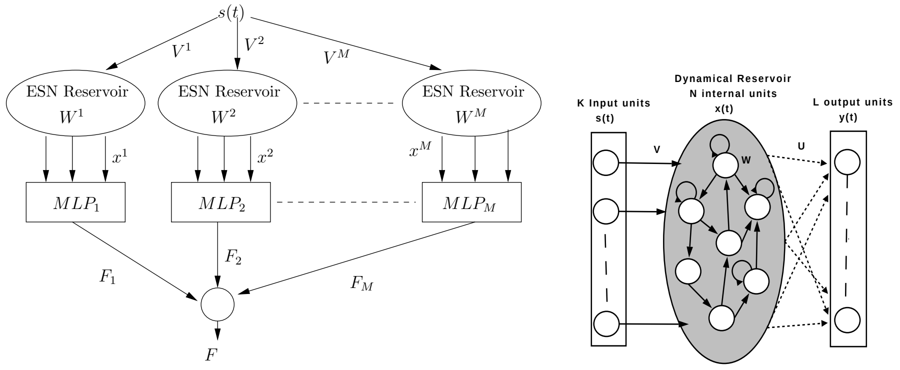

Fig. 1: Ensemble of ESN with MLP readouts (left) and a single ESN (right). The single ESN has linear readout with weight matrix U (reservoir activation vector is extended with a fixed element accounting for the bias term).

Negative Correlation Learning (NCL) has been successfully applied to training MLP ensembles [2, 3, 5, 7]. In NCL, all the individual networks are

[1] There are no feedback connections from the output to the reservoir and no direct connections from the input to the output.

trained simultaneously and interactively through the correlation penalty terms in their error functions. The procedure has the following form: Given a set of M networks and a training input $x(t)$, the ensemble output $F(x((t))$ is calculated as a flat average over all ensemble members $F_i(x(t))$,

$$F(x((t)) = \frac{1}{M}\sum_{i=1}^{M}(F_i(x(t))). \tag{3}$$

In NCL the penalised error functional to be minimised reads:

$$E = \frac{1}{2}(F_i(x(t)) - y(x(t)))^2 + \lambda p_i(x(t)), \tag{4}$$

where

$$p_i(x(t)) = (F_i(x(t)) - F(x(t)))\sum_{i\neq j}(F_j(x(t)) - F(x(t))), \tag{5}$$

and $\lambda > 0$ is an adjustable strength parameter for the negative correlation enforcing penalty term p_i. It can be shown that

$$E = \frac{1}{2}(F_i(x(t)) - y(x(t)))^2 - \lambda(F_i(x(t)) - F(x(t)))^2. \tag{6}$$

Note that when $\lambda = 0$, we obtain a standard de-coupled training of individual ensemble members. Standard gradient-based approaches can be used to minimise E by updating the parameters of each individual ensemble member.

3 Negatively Correlated Ensemble of ESNs

To apply NCL to ensembles of ESN, we replace the linear readout (of standard ESN) with non-linear Multi-Layer Perceptron (MLP)[2]. The training of negatively correlated ensemble of M ESNs consists of first, driving the individual ESN reservoirs with the input stream and collecting the reservoir states $x^i(t) = (x^i_1(t)......x^i_N(t))$, where $x^i(t)$ is the reservoir activation vector of the i-th ESN, $i = 1, 2, ..., M$, at time t. Each ESN i has N reservoir units with reservoir weight matrix W^i and input matrix V^i.

We then use the reservoir states $x^i(t)$ as an input for the MLP readouts F_i (see figure 1 (left)). The readout MLPs had a single hidden layer of logistic sigmoid units (the hidden layer size was determined through cross-validation) and were trained using Negative Correlation learning described above.

We remark that in contrast to standard NCL, in ensemble of ESNs, the maps F_i each receive a different input $x^i(t)$ that provide diverse representations of the common input stream $...s(t-1)s(t)$ observed up to time t. However, one can treat the reservoir activations $x^i(t)$ as internal representations of the i-th ensemble model receiving the common input $s(t)$. From this point of view, all the ensemble models receive the same input, as is the case in the standard NCL.

[2] To exploit the power of negative correlation the ensemble members should be non-linear models. Negatively correlated linear mappings cannot implement the idea of globally correct mappings by all ensemble members, while being locally diverse.

4 Experimental Studies

We employ three timeseries used in the ESN literature covering a spectrum of memory structure [8, 9, 10, 11]. For each data set, we denote the length of the training, validation and test sequences by L_{trn}, L_{val} and L_{tst}, respectively. The first L_v values from training, validation and test sequences are used as the initial washout period. In what follows we briefly introduce the data sets.

- 10th order NARMA system [8, 11]:

$$y(t+1) = 0.3\ y(t) + 0.05\ y(t) \sum_{i=0}^{9} y(t-i) + 1.5\ s(t-9)\ s(t) + 0.1, \tag{7}$$

where $y(t)$ is the system output at time t, $s(t)$ is the system input at time t (an i.i.d stream of values generated uniformly from an interval $[0, 0.5]$). The current output depends on both the input and the previous outputs. In general, modelling this system is difficult, due to the non-linearity and possibly long memory. The input $s(t)$ and target data $y(t)$ are shifted by -0.5 and scaled by 2 as in [8]. The networks were trained on system identification task to output $y(t)$ based on $s(t)$, with $L_{trn} = 2000$, $L_{val} = 3000$, $L_{tst} = 3000$ and $L_v = 200$.

- Chaotic Laser Dataset [8, 10]: The time series is a cross-cut through periodic to chaotic intensity pulsations of a real laser. The task is to predict the next laser activation $y(t+1)$, given the values up to time t; $L_{trn} = 2000$, $L_{val} = 3000$, $L_{tst} = 3000$ and $L_v = 200$.

- Sunspot series [8, 9]: The dataset[3] contains 3100 sunspots numbers from Jan 1749 to April 2007, where $L_{trn} = 1600$, $L_{val} = 500$, $L_{tst} = 1000$ and $L_v = 100$. The task was to predict the next value $y(t+1)$ based on the history of y up to time t.

Experimental setup: The ensemble used in our experiments consists of $M = 10$ ESNs with MLP readouts. In all experiments we use ESNs with reservoirs of $N = 100$ units. Hence, each individual MLP readout has 100 inputs. We used NCL training of readouts via gradient descent on E with learning rate $\eta = 0.1$. The output activation function of the MLP readout was linear for NARMA task and sigmoid logistic for the laser and sunspot tasks.

We optimised the penalty factor λ and the readout complexity (number of hidden nodes in F_i) using the validation set, λ was varied in the range $[0, 1]$ (step size 0.1) [3]. The number of hidden nodes was varied from 1 to 20 (step 1).

The single ESN model architecture described by hyperparameters such as input weight scale, spectral radius and reservoir sparsity, was determined on the validation set. Linear readout was trained via ridge regression [8, 12]. The performance of this model was determined in 10 independent runs (e.g. 10 realisations of ESN based on the best performing hyperparameters).

For ensemble ESN (Ens-ESN-MLP), we used the 10 ESN reservoirs generated in the single ESN experiment as the ensemble members. Due to random

[3] obtained from National Geophysical Data Center (NGDC)

initialisation of MLP readouts, we report the average performance (plus the minimum, maximum and standard deviation values) over 10 random initialisations of MLPs.

Experimental Results: Table 1 summarises the results of the single ESN model, Negatively Correlated ensemble of ESNs and independent ensemble of ESNs ($\lambda = 0$) for the three time series considered in this paper. To assess the improvement achieved by using a genuine NCL training vs. independent training of ensemble members ($\lambda = 0$), the MLP readouts were initialised with the same weight values in both cases. In all datasets, the ESN ensemble trained via NCL outperformed the other models, with the most significant performance gain for NARMA modelling task.

Note that the two ESN ensemble versions we study share the same number of free parameters, with the sole exception of the single diversity-imposing parameter λ in NCL based learning. The single ESN has been used as a natural baseline against which to compare the ensemble performance.

Dataset	Test	ESN	Ens-ESN-MLP	Ens-ESN-MLP
		linear regression	Indep. learning	NCL
NARMA	MSE	0.00102	0.000795	**0.000297**
	STD	0.000101	0.0000142	0.0000237
	Min	0.000865	0.000768	0.000270
	Max	0.00118	0.000810	0.000349
Laser	MSE	0.000197	0.000187	**0.000138**
	STD	0.0000724	0.00000767	0.00000205
	Min	0.0000998	0.000172	0.0000987
	Max	0.000315	0.000197	0.000170
Sunspots	MSE	0.00163	0.00136	**0.00115**
	STD	0.000122	6.385E-06	1.054E-05
	Min	0.00143	0.00136	0.00110
	Max	0.00191	0.00138	0.00116

Table 1: Performance of the single ESN model and the ESN ensemble models.

5 Conclusions

We have empirically demonstrated that coupling ESN models through negatively correlated non-linear readouts can lead to performance improvements over the simple ESN ensemble. In contrast to traditional negatively correlated ensembles, the readouts receive different inputs. However, when considering our model as ensemble of ESNs, each receiving the same input stream, the reservoir activations represent internal feature representations of the inputs and the model can be viewed as a novel generalisation of NCL to state space models.

References

[1] S. Babinec and J. Pospichal. Merging echo state and feedforward neural networks for time series forecasting. In *Proceedings of the 16th International Conference on Artificial Neural Networks (ICANN 2006), volume 4131 of LNCS, pages 367-375. Springer*, 2006.

[2] G. Brown, J. L. Wyatt, , and P. Tino. Managing diversity in regression ensembles. *Journal of Machine Learning Research*, 6:1621–1650, 2005.

[3] G. Brown and X. Yao. On the effectiveness of negative correlation learning. In *First UK Workshop on Computational Intelligence (UKCI'01), Edinburgh, Scotland*, 2001.

[4] K. Bush and C. Anderson. Modeling reward functions for incomplete state representations via echo state networks. In *Proceedings of the International Joint Conference on Neural Networks, Montreal, Quebec*, July 2005.

[5] Y. Liu and X. Yao. Ensemble learning via negative correlation. *Neural Networks*, 12:1399–1404, 1999.

[6] M. Lukosevicius and H. Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[7] R. Mckay and H. Abbass. Analysing anticorrelation in ensemble learning. In *In Proceedings of 2001 conference on Artificial Neural Networks and Expert systems, pp. 22-27*, 2001.

[8] A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, accepted, 2010.

[9] F. Schwenker and A. Labib. Echo state networks and neural network ensembles to predict sunspots activity. In *ESANN 2009 proceedings, European Symposium on Artificial Neural Networks -Advances in Computational Intelligence and Learning, Bruges*, 2009.

[10] J. Steil. Online reservoir adaptation by intrinsic plasticity for backpropagation-decorrelation and echo state learning. *Neural Networks*, 20:353–364, 2007.

[11] D. Verstraeten, B. Schrauwen, M. D'Haene, and D. Stroobandt. An experimental unification of reservoir computing methods. *Neural Networks*, 20:391–403, 2007.

[12] F. Wyffels, B. Schrauwen, , and D. Stroobandt. Stable output feedback in reservoir computing using ridge regression. In *Proceedings of the 18th international conference on Artificial Neural Networks, pp.808-817, Lecture Notes in Computer Science, LNCS 5163, Springer-Verlag*, 2008.