Ensemble Separable Recursive Techniques for MLP Networks

Vijanth S. Asirvadam¹ and Sean F. McLoone²

¹School of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, 31750, Bandar Seri Iskandar, Perak D. Ridzuan, Malaysia. vijanth_sagayan@petronas.com.my
²Department of Electronic Engineering, National University of Ireland Maynooth, Maynooth, Co. Kildare, Ireland

Abstract. Ensemble of hybrid recursive training strategies (or also known as separable learning) are derived for the training of neural networks particularly for MLP networks. This new technique takes the best of hybrid learning by bi-partition weights update to nonlinear recursive training for the optimization of nonlinear weights and training of linear weights in one routine. The proposed ensemble of hybrid weight update takes two form of separable training mechanisms and select the best (binary selection) to update the MLP networks weights. The simulation results on chaotic time series show the ensemble technique approach show slight advantage compared to the two hybrid variants.

Keyword. Recursive Prediction Error, Multilayer Perceptron (MLP), Adaptive Learning, System Identification

1 Introduction

The training for neural networks (e.g. MLP networks) can be considered as an optimization problem where the objective is to minimize cost function, such as the sum-squared error (SSE), with respect to the network parameter \( w \). The training strategy employed then depends on whether the sum squared error is minimized with respect to all the weights using full nonlinear optimization or by using separable approach in which the nonlinear optimization is applied to nonlinear weights and the linear optimization is applied to the single output neuron which is linear-in-parameter [9]. It is shown in batch neural network learning procedures that a better minimization in training curve obtained by dividing the optimization problem respect to the nature of the weights [6][9]. In the case of recursive learning techniques for MLP network it is reported in the literature separable schemes show better performance compared to those non-separable approaches in system identification applications [10][2].

Recently the author of the paper proposed a switching form separable learning technique [3] which discriminate nonlinear weight updates when the subsequent input pattern is correlated. This paper proposed a ensemble of dual separable recursive training techniques for nonlinear systems, as the present author attempted similar approach on partial-linear networks using RBF networks recently [1].

Organization of the paper is as follows, section 2 explains briefly on learning algorithms for neural networks. Section 3 outlines different types of separable recursive or instantaneous training for MLP-network and the ensemble form which is proposed in this paper. Section 4 presents simulation results using dynamic chaotic time-series problem to depict robustness of proposed ensemble separable recursive training schemes and finally section 5 gives concludes the discussion of the paper.

2 LEARNING NONLINEAR SYSTEM

When batch of data \( \{y_k, d_k\}_{k=1}^{N_v} \) is available then the near optimal weight \( \mathbf{w}^* \) vector may then be obtained by minimizing the cost function. This form of minimization known as prediction error estimation [7] is attained by continuous iteration of weight update procedure given by

\[
\mathbf{w}_{t+1} = \mathbf{w}_t + \lambda_t \mathbf{s} (\mathbf{w}_t)
\]

where \( \lambda_t \) and \( \mathbf{s} (\mathbf{w}_t) \) is the learning rate and search direction for the \( t \)th iteration set as the negative gradient of the cost function, \( E(\mathbf{w}) \).

\[
\mathbf{s} (\mathbf{w}_t) = -\nabla E(\mathbf{w}_t) \quad \text{where} \quad \nabla E(\mathbf{w}_t) = \sum_{k=1}^{N_v} \nabla y_k^2(\mathbf{w}_t) = \sum_{k=1}^{N_v} \nabla y_k(\mathbf{w}_t) \cdot \mathbf{e}_k
\]
The backpropagation algorithm (steepest descent), normally yields poor convergence rate. A second order properties in form of matrix, $P(w_t)$ - approximation of the inverse of the Hessian matrix, $H(w_t)$, added to the gradient direction to improve the search direction, $s(w_t)$. The popular second order approximation of inverse Hessian is known as Gauss-Newton method which is given as

$$P(w_t) = H^{-1}(w_t) \quad \text{where} \quad H(w_t) = \sum_{k=1}^{N_v} \nabla y_k(w_t)\nabla y_k^T(w_t)$$

(3)

### 2.1 Recursive (online) Training

The recursive training of MLP-network is based on minimization of cost (error) function, $E(w)$, by accumulating information given by successive presentation of the training data on-line. Based on single data point available at each $t^{th}$ sample instant, the recursive estimate of $E(w)$ is derived as

$$\epsilon^2(w_t) = (y(w_t) - d_t)^2$$

(4)

The on-line or stochastic backpropagation (SBP) algorithm, a first order estimate, performs weight update, $w$, as

$$w_{t+1} = w_t - \lambda_t \epsilon(w_t)$$

(5)

where $\lambda_t$ is the learning rate which can be a constant or time varying value. The disadvantage of SBP is that it shows slow convergence similar BBP but SBP training form the basis for second order recursive training algorithm which will be next.

### 2.2 Recursive Prediction Error - Second Order Approximation

The second order algorithms for recursive prediction error (RPE) method are well investigated for network with linear-in-weight[5][7]. The time varying nonlinear system can also be identified by linearizing the network output $y_t$ about the weight $w_t$ [5],

$$\nabla y(w_t) = \frac{\partial}{\partial w_t} g(w_t, x_t)$$

(6)

where $g(w_t, x_t)$ is the nonlinear function or MLP network. The Hessian matrix for $t^{th}$ training vector, can be derived in recursive form as

$$H_t = \alpha_t H_{t-1} + (1 - \alpha_t)(\nabla y(w_t)\nabla y^T(w_t))$$

(7)

$$w_{t+1} = w_t + H_t^{-1} \nabla y(w_t)\epsilon(w_t)$$

(8)

where $\alpha_t$ is the forgetting factor which is usually set to the value between $0.9 \leq \alpha_t \leq 1$. The inverse of $H$ is computationally intensive to compute direct recursively, thus matrix inversion lemma [7] normally used to calculate the matrix ($P = R^{-1}$) directly as

$$w_{t+1} = w_t + P_t \nabla y(w_t)\epsilon(w_t)$$

(9)

$$S(w_t) = \alpha_t + \nabla y^T(w_t)P_{t-1}\nabla y(w_t) \quad P_t = \frac{1}{\alpha_t}[P_{t-1} - P_{t-1}\nabla y(w_t)S^{-1}(w_t)\nabla y^T(w_t)]P_{t-1}$$

(10)

### 3 Hybrid Training

The hybrid or separable form of recursive training techniques are formed by separating the training into two categories based on the nature of the neuron (nonlinear/linear) in the neural networks. One is to optimize the nonlinear weight using recursive nonlinear training, described in the earlier section. The second is to implement recursive linear square (RLS) optimization on the output neuron weights (linear-in-parameter). From equation (6), the output gradient, $\nabla y(w)$, can be partition to

$$\nabla y(w) = \begin{bmatrix} \nabla y^L(w) \\ \nabla y^L(w) \end{bmatrix} = \begin{bmatrix} \nabla y^N(w) \\ r \end{bmatrix}$$

(11)
where \( r \) is the output and the bias of the hidden layer of MLP which takes the form

\[
r = \begin{bmatrix} r_1 & \cdots & r_{N_h} \end{bmatrix}^T
\]

(12)

The decomposed inverse Hessian, \( P \), or covariance matrix of estimate \( w \) is given as

\[
P_{N_w \times N_w} \cong \begin{bmatrix} P_{NL}^{NL} & 0 \\ 0 & P_{NL}^L \\ \end{bmatrix}
\]

\[
N_{NL} = N_i + 1 \cdot N_h
\]

\[
N_L = N_h + 1
\]

(13)

where \( N_i \) is the number of input to the nonlinear network (MLP) and \( N_h \) is the number of neurons. By decomposition of \( \nabla y(w) \) and \( P \) matrix based on MLP network layer, which separate the learning to linear/nonlinear parameter, the hybrid training techniques can be derived.

3.1 Hybrid or Separable Training Approach

The bi-partition form of neural network training methods, which is proposed by the present author[2] and Ngia et al.[10], does the weight both which are linear and nonlinear-in parameter update simultaneously. The hybrid recursive prediction error (HRPE) is compose of RPE and RLS algorithms for training of nonlinear and linear weights respectively. The nonlinear weights update using RPE can be summarized as

\[
w_{NL}^{t+1} = w_{NL}^t + P_{NL}^t \nabla y(w_{NL}^t) e(w_t)
\]

(14)

\[
S(w_{NL}^t) = \alpha_{NL}^t + \nabla y(w_{NL}^t) P_{NL}^t \nabla y^T(w_{NL}^t)
\]

(15)

\[
P_{NL}^t = \frac{1}{\alpha_{NL}^t} (P_{NL}^{t-1} - P_{NL}^t \nabla y(w_{NL}^t) S^{-1}(w_{NL}^t) \nabla y^T(w_{NL}^t) P_{NL}^{t-1})
\]

(16)

while the linear weights adaptation using RLS is described as

\[
w_{L}^{t+1} = w_{L}^t + P_{L}^t r_{t} e(w_t)
\]

(17)

\[
S(w_{L}^t) = \alpha_{L}^t + r_{t} P_{L}^{t-1} r_{t}^T
\]

\[
P_{L}^t = \frac{1}{\alpha_{L}^t} [P_{L}^{t-1} - P_{L}^t r_{t} S^{-1}(w_{L}^t) r_{t}^T P_{L}^{t-1}]
\]

(18)

3.2 Switching Separable Training Approaches

A new form of hybrid training technique recently being proposed by the present author[3] by extending hybrid recursive training (HRPE) with switching module. The switching mode HRPE (HRPE-sw) update the neural network weights according to the correlation of the input data. The correlation of subsequent input vector to the MLP network is give as,

\[
\zeta = \cos^{-1}\left(\frac{x_{k-1}^T x_k}{\|x_{k-1}\| \|x_k\|} \right) < \varepsilon
\]

(19)

The correlation index in equation (21) will test the correlation of the input data over each sample instant, hence the MLP weights update will be based on the following module.

\[
\text{if } \zeta < \varepsilon \text{ then} \\
\quad \text{update only the linear weights} \\
\text{else} \\
\quad \text{update the both linear AND nonlinear neuron associated weights}
\]

end
3.3 Dual Ensemble Hybrid/Separable Training Approaches

The ensemble hybrid training algorithms (E-HRPE) is proposed in this paper which ensemble both full hybrid training approach (HRPE) and the switching module (HRPE-sw) describe in the previous section. A binary form of ensemble separable training being implemented which selects the weights update that yield minimal cost function $E(w)$ at a particular epoch or sample time, described in procedural code as follows.

$$\text{if } \{E(w_{t+1})\}_{\text{HRPE}} \leq \{E(w_{t+1})\}_{\text{HRPE-sw}} \text{ then}$$

$$w_{t+1} = w_t + [P]_{\text{HRPE}} \nabla y(w_t) \varepsilon(w_t)$$

$$\text{else}$$

$$w_{t+1} = w_t + [P]_{\text{HRPE-sw}} \nabla y(w_t) \varepsilon(w_t)$$

$$\text{end}$$

4 Simulation Results

The separable training approaches are tested using Mackey-Glass chaotic time series which consists of both correlated and random data. Fig. 1 show correlation angle (both in radians and degree) for Mackey-Glass testcase which consists of large number redundant subsequent input vector. The correlation threshold, $\varepsilon$ is to 0.05 radians based on the observations. The chaotic Mackey-Glass[8] time series defined by the differential delay equation

$$\frac{dy_t}{dt} = -0.1y_t + 0.2 \frac{y_{t-1}}{1 + y_{t-17}}$$

$$x_t = [y_{t-1}, y_{t-6}, y_{t-12}, y_{t-18}]$$

and sampled at a rate of one second. A four input, known as nonlinear auto regressive exogenous input (NARX) in system identification literature, are chosen as the model structure for MLP network - equation (20).

4.1 Results Summary and Discussion

The forgetting factor, $\alpha$, for both recursive linear and nonlinear algorithms are set to 0.999. The initial weights for MLP are set as 20 set of random weights. The number of hidden neuron for MLP are tested for three sets (5, 10 and 20) and the simulation results are shown in Tables 1 to 3 and box plots Fig. 2 to Fig. 4 for average of 20 simulation runs. In each cases the proposed ensemble hybrid technique (EHRPE) show distinct advantage compared to other hybrid forms.

<table>
<thead>
<tr>
<th>Table 1: Evaluation for Nh =5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>HRPE</td>
</tr>
<tr>
<td>HRPE-sw</td>
</tr>
<tr>
<td>EHRPE</td>
</tr>
</tbody>
</table>

Figure 1. The correlation angle (both in radians and degree) for Mackey-Glass problem

Figure 2. Boxplots depict NMSE -Nh =5
Fig. 2 to Fig. 4 depict boxplots obtained showing the mean of normalized sum of squares error (NMSE) for 20 initial weights for MLP with 5, 10 and 20 hidden neurons. The column 1: describes the performance the complete hybrid update, 2: switchable form and the 3: is the proposed ensemble technique. The proposed ensemble recursive neural-network training algorithm shows the best results compared to conventional hybrid technique[2][9] and the switchable version[3] for three different network sizes. The switching form of recursive training edges better compared to full separable form of update for 20 random simulations which validates the previous results[3]. The ensemble not only give good mean value but also the deviation from the mean are also greatly minimized.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>%NMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nh = 10</td>
<td>Mean</td>
</tr>
<tr>
<td>HRPE</td>
<td>0.1710</td>
</tr>
<tr>
<td>HRPE-sw</td>
<td>0.1298</td>
</tr>
<tr>
<td>EHRPE</td>
<td>0.0721</td>
</tr>
</tbody>
</table>

**Table 2: Evaluation for Nh = 10**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>%NMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nh = 20</td>
<td>Mean</td>
</tr>
<tr>
<td>HRPE</td>
<td>0.2012</td>
</tr>
<tr>
<td>HRPE-sw</td>
<td>0.2777</td>
</tr>
<tr>
<td>EHRPE</td>
<td>0.0824</td>
</tr>
</tbody>
</table>

**Table 3: Evaluation for Nh = 20**

Figure 3. Boxplots depict NMSE - Nh = 10

Figure 4. Boxplots depict NMSE - Nh = 20

Figure 5. Prediction of Mackey-Glass Chaotic Time Series with Hybrid Variants

5 Conclusions

The ensemble form separable recursive least square algorithms for training of feedforward neural network have been proposed in this work. Two new form of separable learning methods with a switching module, tested on MLP-networks with a layer of sigmoidal hidden neurons and single linear output neuron, show superior performance measure compared to conventional hybrid (or separable) training techniques. Future work will look into adoption of separable recursive training algorithms on dynamic kernel based regression using RBF network[4]. The instantaneous update on RBF
Gaussian kernel tend to give good generalization due to prior setting of initial weights using unsupervised learning techniques.

Acknowledgement

This research project is being funded by E-Science (01-02-02-SF0053) under the Ministry of Science, Technology and Innovations (MOSTI) of Malaysia.

Reference


[2] Asirvadam V. S., McLoone S. F. and Irwin G. W.”Separable Recursive Training Algorithms for Feedforward Neural Networks” IEEE World Congress on Computational Intelligence, pp. 1212-17, May 12-17, 2002, Honolulu, Hawaii,


