<table>
<thead>
<tr>
<th>Title</th>
<th>An Online Self-Constructive Normalized Gaussian Network with Localized Forgetting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Backhus, Jana; Takigawa, Ichigaku; Imai, Hideyuki; Kudo, Mineichi; Sugimoto, Masanori</td>
</tr>
<tr>
<td>Citation</td>
<td>IEICE transactions on fundamentals of electronics communications and computer sciences, E100A(3): 865-876</td>
</tr>
<tr>
<td>Issue Date</td>
<td>2017-03</td>
</tr>
<tr>
<td>Doc URL</td>
<td><a href="http://hdl.handle.net/2115/65552">http://hdl.handle.net/2115/65552</a></td>
</tr>
<tr>
<td>Rights</td>
<td>copyright©2017 IEICE</td>
</tr>
<tr>
<td>Type</td>
<td>article</td>
</tr>
<tr>
<td>File Information</td>
<td>e100-a_3_865.pdf</td>
</tr>
</tbody>
</table>

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP
An Online Self-Constructive Normalized Gaussian Network with Localized Forgetting

Jana BACKHUS†a, Ichigaku TAKIGAWA†,††, Nonmembers, Hideyuki IMAI††, Mineichi KUDO††, and Masanori SUGIMOTO††, Members

SUMMARY In this paper, we introduce a self-constructive Normalized Gaussian Network (NGnet) for online learning tasks. In online tasks, data samples are received sequentially, and domain knowledge is often limited. Then, we need to employ learning methods to the NGnet that possess robust performance and dynamically select an accurate model size. We revise a previously proposed localized forgetting approach for the NGnet and adapt some unit manipulation mechanisms to it for dynamic model selection. The mechanisms are improved for more robustness in negative interference prone environments, and a new merge manipulation is considered to deal with model redundancies. The effectiveness of the proposed method is compared with the previous localized forgetting approach and an established learning method for the NGnet. Several experiments are conducted for a function approximation and chaotic time series forecasting task. The proposed approach possesses robust and favorable performance in different learning situations over all testbeds.

key words: Normalized Gaussian Networks, dynamic model selection, online learning, chaotic time series forecasting

1. Introduction

In applications where data samples are received sequentially, incremental learning schemes have to be applied to train neural networks. In truly sequential learning schemes [16], only one data sample is observed at any time and directly discarded after learning. So, no prior knowledge is available on the number of training data or the data distribution. Furthermore, training data are often not independent and identically distributed (i.i.d.) in real world applications. This is a problem when network models are updated in favor of newly arriving training data. If the data distribution is not i.i.d., then networks are prone to forget already learned information. This phenomenon is called negative or in severe cases catastrophic interference when it is not wanted.

Normalized Gaussian networks (NGnet) are feedforward three layer networks that are related to Radial Basis Function (RBF) neural networks. They have localized learning behavior by partitioning the input space with local units. Only a few units are updated for a newly received data sample. Local model networks are often applied to sequential learning schemes, because their model structure eases the effects of negative interference. NGnets differ from RBF networks in the normalization of the Gaussian activation function. The normalization switches the traditional roles of weights and activities in the hidden layer, and NGnets therefore exhibit better generalization properties [1].

When NGnets are applied to online learning tasks, two major points have to be considered. The learning performance of the NGnet should be robust even in cases where the data distribution is not i.i.d., and since an accurate model size selection at initialization is difficult, the network model should be adapted dynamically. In this paper, we propose an NGnet with localized forgetting that applies self-constructive dynamic model selection for the first time. Our contributions include:

- Revising a previously proposed localized forgetting approach [2].
- Applying dynamic model selection self-constructively.
- Improving model selection in regard to negative interference.
- Dealing with model redundancies by merging units.

The dynamic model selection is applied self-constructively to avoid the necessity of a model initialization and therefore ease application to online learning tasks.

The effectiveness of the proposed method is demonstrated for two experiments: a function approximation task, applying both i.i.d. and non i.i.d. data distributions, and a chaotic time series forecasting task. In regard to the revision of the localized forgetting approach in [2], we compare our proposed method to it, discuss the differences in learning performance, and show that the application of dynamic model selection results in improved learning performance. Also, we compare the proposed method with an established learning method [15] for the NGnet that applies dynamic model selection and show that ours is more robust in different learning environments, especially for non i.i.d. learning tasks. This improved robustness makes our proposed method suitable for online learning tasks.

The rest of the paper is divided as follows: Section 2 explains some basic definitions for the NGnet while the proposed method is explained in Sect. 3 and evaluated in Sect. 4. A conclusion is drawn in Sect. 5.

2. Normalized Gaussian Network and Online EM

2.1 Normalized Gaussian Network

The Normalized Gaussian network (NGnet) is a universal
function approximator that was first proposed by Moody and Darken [12]. It transforms an \( N \)-dimensional input vector \( x \) to a \( D \)-dimensional output vector \( y \) with

\[
y = \sum_{i=1}^{M} N_i(x) \tilde{W}_i \tilde{x}.
\]

The model softly partitions the input space into \( M \) local units with the normalized Gaussian functions \( N_i \). \( \tilde{W}_i \equiv (W_i, b_i) \) is a \( D \times (N + 1) \)-dimensional linear regression matrix with \( \tilde{x}^T \equiv (x^T, 1) \) where prime (‘) denotes a transpose.

### 2.2.2 Online EM Algorithm

A stochastic interpretation was first proposed for the NGnet by Xu et al. [19]. The unknown model parameters are estimated by maximum likelihood estimation based on the log-likelihood of the observed in- and output data \((x, y)\). Here, the Expectation-Maximization (EM) algorithm is applied for parameter estimation. An offline approach has been proposed by Xu et al. [19] that was later adopted to an online EM algorithm by Sato and Ishii [15].

The stochastic model is defined by the following probability distribution for a complete event \((x, y, i)\) [19]

\[
P(x, y, i|\theta) = (2\pi)^{-D/2} \sigma_i^{-D} |\Sigma_i|^{-1/2} M^{-1} \times \exp \left[ -\frac{1}{2} (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - \frac{1}{2} \sigma_i^2 (y - \tilde{W}_i \tilde{x})^2 \right],
\]

where \( \theta \equiv \{ \mu_i, \Sigma_i, \sigma_i^2, \tilde{W}_i | i = 1, \ldots, M \} \) is the set of model parameters that have to be estimated. For the online EM, the parameters are updated with the following E- and M-step.

#### 2.2.1 E (Estimation) Step:

Given the current estimator \( \theta(t-1) \), the posterior probability \( P_i(t) \equiv P(i|x(t), y(t), \theta(t-1)) \) evaluates how likely the \( i \)-th unit generates the current observation \((x(t), y(t))\):

\[
P_i(t) \equiv \frac{P(x(t), y(t), i|\theta(t-1))}{\sum_{j=1}^{M} P(x(t), y(t), j|\theta(t-1))}.
\]

#### 2.2.2 M (Maximization) Step:

The model parameters \( \theta \) are updated with

\[
\mu_i(t) = \frac{\langle x \rangle_i(t)}{\langle 1 \rangle_i(t)},
\]

\[
\Sigma_i^{-1}(t) = \frac{1}{\langle x x^T \rangle_i(t) - \mu_i(t) \mu_i^T(t)} - 1,
\]

\[
W_i(t) = \frac{\langle y x^T \rangle_i(t) - \mu_i(t) \mu_i^T(t)}{\langle x x^T \rangle_i(t) - \mu_i(t) \mu_i^T(t)} - 1,
\]

\[
\sigma_i^2(t) = \frac{\langle y y^T \rangle_i(t) - Tr(W_i(t) \langle x x^T \rangle_i(t))}{D}.
\]

The parameter updates (6)–(9) include symbols \( \langle \rangle_i \) that denote weighted sums of accumulated observations \((x(t), y(t))\) until the current time step \( t \).

A weighted sum is incrementally updated at each time step \( t \) with respect to the posterior probability \( P_i(t) \).

\[
\langle f \rangle_i(t) = \lambda(t) \langle f \rangle_i(t-1) + P_i(t)f_i,
\]

where \( f \equiv f(x, y) \) and \( f_i \equiv f(x(t), y(t)) \) are used with abbreviations. A time-dependent discount factor \( \lambda(t) \) has been introduced to (10). It plays an important role in discarding the effect of old learning results that were employed to an earlier inaccurate estimator. The factor has to be chosen such that \( \lambda \to 1 \) when \( t \to \infty \) to fulfill the Robbins-Monro condition for convergence of stochastic approximations [9].

In the online EM [15], \( \langle f \rangle_i(t) \) has been defined as a weighted mean by employing a normalization coefficient \( (1 + \lambda(t) / \eta(t-1))^{-1} \). The weighted mean is calculated as follows:

\[
\langle f \rangle_i(t) = (1 - \eta(t)) \langle f \rangle_i(t-1) + \eta(t) P_i(t)f_i.
\]

In the following, we will however omit the normalization by employing a weighted sum (10) as defined in [2].

### 3. Self-Constructive Normalized Gaussian Network with Localized Forgetting

In this paper, we introduce an improvement for a parameter estimation method with localized forgetting and then discuss the application of dynamic model selection to an NGnet with localized forgetting for the first time.

#### 3.1 Parameter Update with Localized Forgetting

Previously, Celaya and Agostini have proposed a parameter update with localized forgetting [2] to improve learning robustness. The weighted sum (10) employs a discount factor \( \lambda(t) \) to forget old learning results over time. The time-dependent discount is a problem when data are not i.i.d., and learning is prone to negative interference. In [2], forgetting is based on weights so that only as much old information is forgotten as new information is received for a unit \( i \). For this, (10) has been changed to

\[
\langle f \rangle_i(t) = \lambda(t) \langle f \rangle_i(t-1) + \frac{1 - \lambda(t)}{1 - \lambda(t)} f_i.
\]

This update method is preferable because of its stable learning performance. We consider it however bad practice that the update with new information \( f_i \) in (12) is dependent on discount \( \lambda(t) \). \( \lambda(t) \) controls forgetting and has been originally introduced to speed up convergence [14]. The
constraint for \( \lambda(t) \) is \((0 \leq \lambda(t) \leq 1) \) [15], where \( \lambda(t) = 1 \) is a special case with no forgetting of old learning results. Yet, the update factor \( \frac{1 - \lambda(t)P_i(t)}{1 - 2\lambda(t)P_i(t)} \) becomes undefined over all \( t \) when \( \lambda(t) = 1 \), and the NGnet is then unable to learn.

We propose a modified localized forgetting approach by reconsidering its derivation from (10) with additional aspects that have not been considered previously. For our derivation, (10) is rewritten to

\[
\langle f \rangle_t = \Lambda(P_i(t))\langle f \rangle_{t-1} + \Omega(P_i(t))f_i,
\]

where a new forgetting factor \( \Lambda(P_i(t)) \) and update factor \( \Omega(P_i(t)) \) have to be determined by considering the following conditions. For a full update, \( P_i(t) = 1 \) and (13) should reduce to the same as (10), given by \( \Lambda(1) = \lambda(1) \) and \( \Omega(1) = 1 \). For the NGnet, the units divide the input space between each other according to their representation ability of a current sample \((x(t), y(t))\). A full update only happens at a same time step \( t \) for a weighted sum \( \langle f \rangle_{t,M} \) that accumulates information over all units \( M \), or when a unit represents the data sample to 100% which rarely happens. Additionally, an update with value \( f_i \) should be consistent for a shared weighted sum \( \langle f \rangle_{i+j}(t-1) \) by unit \( i \) and \( j \) when updated once with weight \( P_i(t)+P_j(t) \) or twice with weights \( P_i(t) \) and \( P_j(t) \) at a time step \( t \). For a single update with \( (P_i(t)+P_j(t)) \), we get

\[
\langle f \rangle_{i+j}(t) = \Lambda(P_i(t)+P_j(t))\langle f \rangle_{i+j}(t-1)
+ \Omega(P_i(t)+P_j(t))f_i.
\]

If \( \langle f \rangle_{i+j}(t-1) \) is updated twice with \( P_i(t) \) and \( P_j(t) \), then

\[
\langle f \rangle_{i+j}(t) = \Lambda(P_i(t))\langle f \rangle_{i+j}(t-1) + \Omega(P_i(t))f_i,
\]

\[
\langle f \rangle_{i+j}(t+1) = \Lambda(P_j(t))\langle f \rangle_{i+j}(t) + \Omega(P_j(t))f_i.
\]

We insert (15) into (16) and get

\[
\langle f \rangle_{i+j}(t+1) = \Lambda(P_j(t))\langle f \rangle_{i+j}(t-1) + (\Lambda(P_j(t))\Omega(P_i(t)) + \Omega(P_j(t)))f_i.
\]

(17) is however one step ahead of time and ignores the fact that units divide a current data sample \((x(t), y(t))\) between each other at a time step \( t \). Furthermore, the update \( \Omega(P_i(t))f_i \) is already partially forgotten by \( \Lambda(P_i(t)) \) which should be prevented for updates on the same \( t \). Then, (17) reduces to

\[
\langle f \rangle_{i+j}(t) = \Lambda(P_j(t))\langle f \rangle_{i+j}(t-1)
+ (\Omega(P_i(t)) + \Omega(P_j(t)))f_i.
\]

Finally, we get two functional equations from (14) and (17)

\[
\Lambda(P_i(t)+P_j(t)) = \Lambda(P_i(t))\Lambda(P_j(t)),
\]

\[
\Omega(P_i(t)+P_j(t)) = \Omega(P_i(t)) + \Omega(P_j(t)).
\]

For (19), the solution is well known to be of the form \( \Lambda(P_i(t)) = cP_i(t) \), where \( c \) is determined by using \( \Lambda(1) = \lambda(1) \). So, \( \Lambda(P_i(t)) = \lambda(t)P_i(t) \) is derived while \( \Omega(P_i(t)) = P_i(t) \) is the solution for the update factor. The new stepwise update for the weighted sum is obtained as

\[
\langle f \rangle_t = \lambda(t)P_i(t)\langle f \rangle_{t-1} + P_i(t)f_i.
\]

The new approach complies with all considered dependencies and adds new information \( f_i \) independent from \( \lambda(t) \). It is then able to update its parameters over all \( 0 \leq \lambda(t) \leq 1 \).

3.2 Model Selection

The selection of an accurate model size is another important problem when applying neural networks. Especially for online learning tasks, it is difficult to choose a good model size because domain knowledge is limited. Furthermore, choosing an accurate model size by hand necessitates extensive trial-and-error studies. An alternative is dynamic adaptation of the network model during learning. Dynamic model adaptation avoids the need to set a static model size in advance and can be achieved with some methods to increase or reduce model complexity. Also, the model can be built from scratch during learning by applying these methods self-constructively. Several works for RBF networks have proposed self-constructive unit adaptation, Platt’s Resource Allocating Network (RAN) [13] as well as its extensions (RANEKF [8], MRAN [11]) and the GGAP-RBF network that has been proposed by Huang et al. [6].

3.2.1 Unit Manipulation Mechanisms

Some unit manipulation mechanisms have been introduced for the NGnet [15] and are adapted to use with our localized forgetting approach. The adapted unit manipulation mechanisms include a produce, delete and split mechanism. However, the delete mechanism needs to be revised to be applicable to the local forgetting NGnet. In addition, the manipulation decision of split is changed to improve robustness against negative interference, and a merge mechanism is added to reduce redundancy of units and further improve model compactness.

(1) Produce

\( P(x(t), y(t))\theta(t-1) \) is a probability that indicates how properly the current model parameters \( \theta(t-1) \) can estimate the newly received data sample \((x(t), y(t))\). When the probability is smaller than a certain threshold \( T_{\text{produce}} \), a new unit is created according to the produce mechanism in [15].

(2) Delete

A weighted sum \( \langle 1 \rangle_t \) indicates how much the i-th unit has been in charge of the observed data until the current time step \( t \). When the delete mechanism has been first proposed in [15], \( \langle 1 \rangle_t \) was assumed to be a weighted mean scaled between zero and one by a normalizer \( \eta(t) = (1 + \lambda(t)/\eta(t-1))^{-1} \). The normalizer \( \eta(t) \) replaces the normalizer \( 1/T \) used in the offline EM-algorithm, where \( T \) is the total number of samples, which is unknown in the online EM-algorithm.

In our approach, \( \langle 1 \rangle_t \) is however an unscaled
weighted sum and cannot be used directly as a reference. We
introduce therefore a local unit update counter as a normal-
izer to overcome this problem. To take each unit’s received
discount into account, we define the update counter $c_{i}^{(\text{update})}$
similarly to $\eta(t)$ stated above, but here each unit $i$ manages
its own counter locally. The update counter is incremented
by one at every time step where the unit’s update is numer-
ically important. In other words, when the update weight
$P_{i}(t) > 10^{-16}$, the update counter is incremented by
$$c_{i}^{(\text{update})} = 1 + \lambda(t) P_{i}(t) c_{i}^{(\text{update})}.$$  \hspace{1cm} (22)
A unit $i$ is deleted if $\langle 1 \rangle_{i}(t)/c_{i}^{(\text{update})} < T_{\text{Delete}}$, where
$T_{\text{Delete}}$ is a delete threshold.

(3) Split
The output variance of a unit $i$, $\sigma_{i}^{2}(t)$, represents the
accumulated squared error between the unit’s predictions and the
real outputs. High variance values are related to the unit
being in charge of a too large partition of the input space, and
splitting such units can improve learning performance.
Our split decision compares $\sigma_{i}^{2}(t)$ with the output variances
of the other units using a local evaluation where only output
variances of some nearest neighbors are considered. When
the unit’s output variance is considerably bigger than the
biggest variance of its neighbors, the unit is split according to
the split mechanism in [15].

(4) Merge
A merge manipulation is introduced to reduce redundancies
in the network model. Redundancy refers to two network
units approximating a similar partition of the input-output
space. For finding possible merge candidates, the grade of
overlap between units has to be evaluated. The overlap
is calculated here with the Bhattacharyya coefficient (BC)
for that a closed form solution exists for two multivariate
Gaussian distributions $G_{1}(\mu_{1}, \Sigma_{1})$ and $G_{2}(\mu_{2}, \Sigma_{2})$ [5].
$$d_{B}(G_{1}, G_{2}) = \frac{1}{8} \cdot (\mu_{1} - \mu_{2})^{\Sigma^{-1}}(\mu_{1} - \mu_{2})$$
$$+ \frac{1}{2} \cdot \log \frac{\Sigma}{\sqrt{\det \Sigma_{1} \cdot \det \Sigma_{2}}}.$$ \hspace{1cm} (23)
$$BC(G_{1}, G_{2}) = \exp(-d_{B}(G_{1}, G_{2})).$$ \hspace{1cm} (24)
Here, $d_{B}$ is the Bhattacharyya distance and $\Sigma = (\Sigma_{1} + \Sigma_{2})/2$.
The Bhattacharyya coefficient has the advantage that merge
can be applied online based on currently available informa-
tion of each unit, which is for example not possible for the
merge mechanism in [18].

For a similarity $S(i, j)$ between two units $i$ and $j$,
we have to calculate the overlap of the units’ input and
output distributions. We denote the input distribution of
a unit $i$ as $G_{i}^{\text{input}}(\mu_{i}, \Sigma_{i})$ and the output distribution as
$G_{i}^{\text{output}}(W_{i}, \sigma_{i}^{2} I)$. The similarity $S(i, j)$ is calculated by
$$S(i, j) = BC(G_{i}^{\text{input}}, G_{j}^{\text{input}}) \cdot BC(G_{i}^{\text{output}}, G_{j}^{\text{output}}).$$ \hspace{1cm} (25)
If $S(i, j) > T_{\text{Merge}}$ for a threshold $T_{\text{Merge}}$, then the units are
possible merge candidates. The flow of the merge mecha-
nism is described in the following:
1. Calculate the similarity $S(i, j)$ for all pairs $[i, j]$.
2. Choose the pair $[i_{\text{max}}, j_{\text{max}}]$ with maximal similarity.
3. If $S(i_{\text{max}}, j_{\text{max}}) > T_{\text{Merge}}$ then merge units, go to 1.
4. Otherwise, stop routine.
The merge mechanism becomes computationally heavy with
increasing model complexity $M$. Furthermore, merge can-
didates are not found frequently, and it is sufficient to ap-
ply merge in intervals of a few hundred time steps. Yet,
the calculation of the output $BC$ depends on input $x$ for
the output center $W_{i}x$. Calculating similarities with cur-
cent input $x(t)$ is inappropriate when performed in intervals.
a possible alternative would be to use the weighted sum
$\langle y \rangle_{i}(t) \equiv \langle \langle y \rangle_{i}(t), \langle 1 \rangle_{i}(t) \rangle$, however preliminary
experiments have shown that this approach is overestimating
the similarity between output distributions.

We revise the overlap calculation to avoid the inclusion of
$x$ in the output center for the BC calculation. A multivari-
ate theorem for Gaussian distributions can be used to conduct
an affine transformation of the output distribution. Here, we
want to transform the output distribution $y \sim N(\bar{W}_{i}x, \sigma_{i}^{2} I)$
so that input $x$ is not included in the center. For con-
venience, we consider the transformation of the transpose
$y' \sim N(\bar{x}W'_{i}, \sigma_{i}^{2} I)$ instead. After affine transformation, the
output distribution becomes $\tilde{W}'_{i} \sim N(\bar{x}W'_{i}, \sigma_{i}^{2} I)$ instead.
and input $x$ is excluded from the center. But $W'_{i}$ is an
$(N + 1) \times D$-dimensional matrix, and the left term of (23)
becomes a $D \times D$-dimensional matrix dependent on the output
dimension $D$. Therefore, we update (23) to
$$d_{B}(G_{1}, G_{2}) = \frac{1}{8D} \cdot Tr \left( (\mu_{1} - \mu_{2})^{\Sigma^{-1}}(\mu_{1} - \mu_{2}) \right)$$
$$+ \frac{1}{2} \cdot \log \frac{\Sigma}{\sqrt{\det \Sigma_{1} \cdot \det \Sigma_{2}}}.$$ \hspace{1cm} (26)
where $Tr$ is the trace of the matrix.

Finally, we have to merge units $i$ and $j$ when $S(i, j) > T_{\text{Merge}}$. Again, we consider the units’ Gaussian distribu-
tions for the input and output space and merge the centers
and covariances of input and output distributions in the same
matter. New centers $\mu_{\text{new}}$ and covariances $\Sigma_{\text{new}}$ are cal-
culated by
$$\mu_{\text{new}} = \omega_{i} \mu_{i} + \omega_{j} \mu_{j},$$ \hspace{1cm} (27)
$$\Sigma_{\text{new}} = \sum_{k} \omega_{k} (\Sigma_{k} + (\mu_{k} - \mu_{\text{new}})(\mu_{k} - \mu_{\text{new}})^{T})$$ \hspace{1cm} (28)
where $\omega_{k} = \langle \langle \rangle \rangle_{k}(t) / \langle \langle \rangle \rangle_{j}(t)$ are functioning as weights
with $\langle \langle \rangle \rangle_{j}(t)$ representing the importance of unit $i$ and $j$ during
training until current time step $t$, calculated with (21).
3.2.2 Self-Constructive Model Adaptation

In online learning schemes, domain knowledge is limited and an accurate model initialization before training is often difficult. Therefore, we build the network model dynamically and self-constructively during learning. The model is initialized with zero units ($M = 0$), and an initial unit is produced upon receiving the first training data sample. Some preset initialization is necessary for the first unit, where parameters are set to

$$\mu_1 = x(t)$$  \hspace{1cm} (29)

$$\Sigma_1 = I_N$$  \hspace{1cm} (30)

$$\sigma^2 = 0.1$$  \hspace{1cm} (31)

$$\tilde{W}_1 = (0, y(t)).$$  \hspace{1cm} (32)

All other units are then created in relation to the first unit with the self-constructive model selection algorithm described below.

1. \textbf{if} $M = 0$ \textbf{then}
2. \hspace{0.5cm} Produce first unit
3. \hspace{0.5cm} \textbf{else}
4. \hspace{1cm} Compute Probabilities:
5. \hspace{1cm} $P(x(t), y(t))\theta(t - 1)) = \Sigma_{i=1}^{M} P(x(t), y(t), i)\theta(t - 1))$
6. \hspace{1cm} \textbf{if} $P(x(t), y(t))\theta(t - 1)) < T_{\text{produce}}$ \textbf{then}
7. \hspace{1.5cm} Produce new unit
8. \hspace{1.5cm} \textbf{else}
9. \hspace{2cm} Update units
10. \hspace{2cm} Test delete, split, merge
11. \hspace{1.5cm} \textbf{end if}
12. \hspace{0.5cm} \textbf{end if}

4. Experiments

We evaluate the effectiveness of our proposed method, an NGnet with local forgetting (LF) that applies self-constructive model selection for the first time, and compare it with two older learning approaches (GF, GFdisc) for a function and chaotic time series approximation task.

4.1 Preparations

First of all, we discuss several preparative decision steps that have been taken before the experiments.

4.1.1 Compared Methods

In our experiments, we want to evaluate our proposed method in regard to the methods in [2] and [15].

In a first step, we compare our method with [2] to discuss differences between the newly proposed local forgetting update approach in Sect. 3.1 and the previous one [2]. In [2], model selection was yet to be considered. Since it is not applicable without proposing some changes in the delete manipulation, we do not apply dynamic model selection for this comparison. Instead, we compare the two methods with a static model size and identical initialization.

In a second step, we compare our method with an NGnet with global forgetting $GF$ that has been proposed in [15] and where dynamic model selection has been considered. Here, we include dynamic model selection for our approach and compare it with two versions of $GF$: one is equal to [15] including a normalization coefficient ($GF_{\text{norm}}$) in its update function as in (11), while the other one employs only a discount factor ($GF_{\text{disc}}$) as in (10). For $GF_{\text{disc}}$, we need a normalizer for the delete mechanisms because of the same reasons described in Chapter 3.2.1 Paragraph (2). For global forgetting, the normalizer is equal to a unit life time counter, because units forget and need to be updated at every time step. Both global forgetting approaches apply dynamic model selection self-constructively in all experiments.

4.1.2 Dynamic Model Selection

Threshold parameters have to be set for the four manipulation mechanisms: produce (Prod.), delete (Del.), split (Spl.) and merge (Mrg.). We have conducted many preliminary experiments to find accurate parameters for all compared methods, and the best performing ones were selected separately for each experiment and method.

An overview of the selected parameters can be found in Table 1. We have summarized the two global forgetting methods as $GF$ when the same parameters have been selected after extensive testing. For the function approximation task, we test different learning scenarios that are all marked with $FA$ in the table. Some split parameters have the form $3\times$ or $10\times$, which means three or ten times. In those cases, the split threshold is dynamic and set relative to another value. For the chaotic experiment, we also apply a dynamic threshold decision for $GF_{\text{disc}}$’s split mechanism, because the static one used in [15] has unstable performance. A unit is then split when its output variance is ten times bigger than the average of the other units’ variances.

The produce threshold is set to the same parameter for all compared methods within an experiment. This is reasonable because the same produce mechanism is used by

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Manipulation parameter settings for all experiments.</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA: Balanced</td>
<td>LF (Prop.)</td>
</tr>
<tr>
<td>FA: Imbalanced</td>
<td>LF (Prop.)</td>
</tr>
<tr>
<td>FA: Dynamic</td>
<td>LF (Prop.)</td>
</tr>
<tr>
<td>FA: Chaotic</td>
<td>LF (Prop.)</td>
</tr>
<tr>
<td>GFdisc</td>
<td>LF (Prop.)</td>
</tr>
</tbody>
</table>
all methods and differences in model selection behavior are more dependent on the other manipulations. After testing several values, we set $T_{\text{produce}} = 0.1$ for the $FA$ experiments and $T_{\text{produce}} = 0.00001$ for the $chaotic$ experiment.

4.1.3 Performance Evaluation

For all experiments, we evaluate the learning performance of the tested methods with the Root Mean Square Error (RMSE). The RMSE is defined by

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{n_{\text{test}}}},$$

where $n_{\text{test}}$ is the number of test data samples.

Additionally, one-tailed permutation tests are conducted for the function approximation task. In the result tables, we present obtained p-values in the column $p$-Val to evaluate the significance of the learning results, comparing $LF$ with the other tested methods.

4.1.4 Selection of Discount Factor

The discount factor $\lambda(t)$ that is applied in (21) is scheduled over time $t$ with

$$\lambda(t) = 1 - \frac{1 - a}{at + b},$$

depending on the two parameters $a$ and $b$. $a$ controls how fast $\lambda(t) \to 1$ and $b$ sets the initial value of $\lambda$. An example of $a$’s influence is shown in Fig.1. In our experiments, we apply several values for $a$ and $b$ to evaluate the effect of different discount schedules on the performance of the compared methods.

We also conduct experiments without forgetting. Here, we apply a discount factor $\lambda(t) = 1$ which is equal to any value $b$ and $a = 1$. Then, $LF$ updates (21) reduce to the same as $GFdisc$’s (10) making a fair comparison of their different model selection approaches possible.

4.2 Function Approximation Task

In the first experiment, we consider a commonly used function approximation task ([2], [15], [16]). The function has the input dimension $N = 2$, the output dimension $D = 1$, and is defined by

$$g(x_1, x_2) = \max\{e^{-10x_1^2}, e^{-50x_2^2}, 1.25e^{-5(x_1^2 + x_2^2)}\}. \quad (35)$$

A normally distributed random noise $\epsilon(t) \sim N(0, 0.01)$ is added to each function output $g(x_1(t), x_2(t))$, and $y(t) = g(x_1(t), x_2(t)) + \epsilon(t)$ is obtained as the noisy sample output.

4.2.1 Preparations

We consider three different test cases for the function approximation task. In the following, we discuss how we have obtained training and test data for these cases.

(1) Training Data

We use different input data distributions to evaluate the robustness of the compared methods in a variety of learning scenarios. When data are not i.i.d., neural networks are more prone to negative interference and achieving good performance is difficult. Three test cases are considered

- Balanced
- Imbalanced
- Dynamic

For the $balanced$ test case, the training data are i.i.d over the whole input space $(-1 \leq x_1, x_2 \leq 1)$. For the $imbalanced$ test case, we use non-identically distributed data. Here, 95% of the data samples are extracted from a sub-region of the input domain with $(0 \leq x_1, x_2 \leq 0.25)$, and the remaining data are i.i.d in $(-1 \leq x_1, x_2 \leq 1)$. For these two cases, we obtain 10,000 data samples for each test run. For the $dynamic$ test case, the input data distribution is slowly changing for $x_1$ over time. $x_1$'s sample distribution starts at the interval $[-1, -0.2]$ and ends at the interval $[0.2, 1]$ after 250,000 training samples. $x_2$ is sampled in $[-1, 1]$ over all $t$.

A similar experiment has been conducted in [2] and [15] to evaluate learning performance in dynamic environments.

In all test cases, we have obtained training data sets for 50 test runs, and the same data sets are applied to all compared methods.

(2) Test Data

Similarly to [2], we have obtained test data from a regular grid formed of $21 \times 21$ points in the input domain. These data are used to evaluate the learning performance of the compared methods over the whole input space in all three test cases. Additionally, we have obtained test data in the last input interval of the $dynamic$ test case. Here, performance is evaluated by two test sets, representing the whole input space ($RMSE\ All$) and the last training interval ($RMSE\ Last$) respectively. For the function approximation task, all presented results are the averages of 50 test runs.

4.2.2 Experiments without Model Selection

In a first step, we compare our proposed method LF ($Prop.$)
with the former local forgetting method (\textit{LF} \cite{2}) for the \textit{balanced} test case. This experiment evaluates the effectiveness of the newly proposed update function (21) in regard to the previous one (12). The proposed method’s main purpose is to eliminate the inability of \textit{LF} \cite{2} to update its parameters when the discount factor \( \lambda (t) \) is one over all \( t \).

Two different static model sizes are tested. A static model with 25 units is equal to the size used in \cite{2}, and we add some experiments with 100 units. Performance results are stated in Table 2 with \textit{Init} (25) and \textit{Init} (100) respectively. In both cases, units are initialized on a grid of the input space. For the initialization, \((x, y)\) values are obtained according to the location of a unit on the grid, and the unit’s parameters are then set according to (29) - (32).

Experimental results are presented in Table 2, with the best results marked in bold. We have applied different discount schedules for forgetting in the first nine cases, and \textit{LF (Prop.)} has achieved better performance in all cases. Yet, the \textit{LF} \cite{2} difference in learning performance is very small and results \textit{LF (Prop.)} has achieved better performance in all cases. Yet, the \textit{LF} \cite{2} count schedules for forgetting in the first nine cases, and best results marked in bold. We have applied different discount factors \( a=0.001 \) to update its parameters as stated for \textit{FA: Balanced} test case without forgetting. Experimental results are presented in Table 3 for two different manipulation parameter settings. For both settings, we have chosen the same parameters as stated for \textit{FA: Balanced} in Table 1 except for delete. The best performing results are marked in bold in the table. We have also included information about the number of manipulation occurrences.

For \textit{Setting 1}, the delete parameter has been set to \( T_{\text{Delete}} = 0.0001 \) so that both methods do not delete units. Comparing the performance of \textit{LF} and \textit{GFdisc}, we can see that they are performing approximately the same. \textit{GFdisc} is slightly better, but the high p-value shows that this difference is not significant. On the other hand, the network size is smaller for \textit{LF} so that it performs equally well with a smaller network size.

For \textit{Setting 2}, we set \( T_{\text{Delete}} = 0.01 \) for \textit{GFdisc} and \( T_{\text{Delete}} = 0.02 \) for \textit{LF} to compare the two methods when they have similar deletion behavior. The improved learning performances show that deleting units has a positive effect for both methods. Yet, while \textit{LF} improves a lot, \textit{GFdisc} is only improving slightly. The difference in learning performance is significant, supported by the p-value equaling zero. This emphasizes the effectiveness of the localized deleting approach applied to \textit{LF}.

Finally, we compare the results of \textit{LF} applying model selection with the results of \textit{LF (Prop.)} without model selection in Table 2. Without model selection, \textit{LF} has achieved \textit{RMSE} = 0.0557 as the best performance. With model selection, \textit{LF} performs better while having less network complexity for both settings presented in Table 3.
(2) Balanced Test with Forgetting

For the balanced test case with forgetting, we compare LF with the two global forgetting approaches GFnorm and GFdisc. Experimental results are presented in Table 4 with the best results for each discount schedules marked in bold.

Here, LF is not the overall superior approach but shows the most robust performance over all discount schedules in regard to learning performance and network size. Also, it performs best for discount schedules where the learning behavior of the global forgetting methods is unstable. On the other hand, GFdisc is able to perform best for discount schedules where its learning is stable. This applies for discount schedules that are near to one from the start (b = 1000) or reaching one very fast (a = 0.1). For higher discounts, the performance of GFdisc steadily decreases and network sizes increase, resulting in non robust behavior. The learning behavior of GFnorm has the same tendencies as GFdisc’s over the different discount schedules. Yet, GFnorm is unable to perform best for any discount schedule.

The permutation tests were conducted comparing LF with either GFnorm or GFdisc. Regardless of whether the local or global forgetting method has performed better, the resulting p-values were approximately equal to zero in all cases proving the significance in the performance differences.

Finally, we compare the results of LF applying model selection with the results of LF (Prop.) without model selection in Table 2. Without model selection, LF has achieved learning performances between RMSE = 0.0725 and RMSE = 0.0435 depending on the network size and discount schedule. For LF with model selection, the performance is ranging between RMSE = 0.0372 and RMSE = 0.0287. So even in the worst performing case, LF performs still better and possesses smaller network sizes when model selection is applied. This emphasizes the advantage of applying dynamic model selection to LF.

(3) Imbalanced Test without Forgetting

For the imbalanced test case without forgetting, experimental results are presented in Table 5 with best results marked in bold. Here, LF and GFdisc trained NGnets are compared for discount \( \lambda = 1 \). Except for delete, manipulation parameters are set as stated in Table 1 for FA: Imbalanced. For delete, two threshold parameter settings are tested. In Setting 1, the threshold is set to \( T_{Delete} = 0.0001 \) so that no units are deleted, while in Setting 2 we have relaxed the thresholds to \( T_{Delete} = 0.001 \) for both methods to show the differences in performance when deleting occurs.

LF shows superior performance for both tested parameter settings with high significance as proved by the low p-values. In Setting 1, updates and deletions are the same for both methods so that LF’s favorable performance is related to the new split and merge approach. In Setting 2, some units have been deleted. This has led to a decrease in learning performance for both methods, but the decrease is visible lower for LF. The negative effect of delete is here smaller, likely because of the localized deleting approach applied to LF.

(4) Imbalanced Test with Forgetting

For the imbalanced test case with forgetting, the experimental results are presented in Table 6. Again, several discount schedules are applied and the best result for each schedule is marked in bold. Because of the imbalanced data distribution, this testbed is more prone to negative interference.

The results show favorable performance for our method LF in almost all cases with high significance proven by p-values equaling zero. There is one exception for the discount schedule (a = 0.1, b = 1000), where LF and GFdisc have approximately the same learning performance. While there is no significant difference between the performances, the net-
shortly better than LF. However, its performance decreases again. On the other hand, LF is able to steadily improve its approximation. The final performance results in Table 7 also show that LF is performing much better over the whole function. Interestingly, GFdisc performs better for the last training interval RMSE Last, but is unable to perform well over the whole function. Reasons for the bad performance of GFdisc are global forgetting and likely the high number of unit productions and deletions. LF also experiences a performance decrease in comparison with the non-forgetting case but is still able to perform similarly well. These results emphasize again the robustness of LF.

The delete thresholds of both methods are equal to the ones in the dynamic without forgetting case, so it is interesting to compare the deletion behavior. For LF, the number of unit deletions is only slightly increasing for the forgetting case, having a stable deletion behavior. On the other hand, while GFdisc has not deleted any units before, the number of deletions increased when discount was applied. This shows the high impact of forgetting on the unit manipulations in GFdisc and is likely one reason for its learning instabilities. LF is therefore preferable for online learning tasks, especially when data distributions are not i.i.d.

### 4.3 Chaotic Time Series Approximation Task

In the second experiment, we apply the NGnet to a chaotic time series approximation task to examine the learning performances for a realistic and difficult problem. The Lorenz attractor [10] is applied as testbed and is defined by

$$\dot{x} = a(y - x), \quad \dot{y} = bx - y - xz, \quad \dot{z} = xy - cz.$$  \hfill (36)

For the commonly used parameters \(a = 10, b = 28, c = 8/3\), the attractor has chaotic behavior. Trajectories from the attractor are obtained by fourth order Runge-Kutta integration with integration time step \(t_i = 0.001\). The vector notation of the attractor is \(X = F(X)\) where \(X \equiv (x, y, z)'\) and \(F\) denotes the vector field. Additionally, some noise can be added to each observation with

$$Y(t) = X(t) + \xi(t),$$  \hfill (37)
where the noise $\xi(t)$ is a white Gaussian noise with zero mean and some standard deviation $SD$.

Commonly, a time delay embedding method is used for the application of chaotic time series. According to Takens’ theorem [17], the chaotic dynamics can be reconstructed from one dimension, e.g. dimension $x$. A next state $x(t + P)$ is then defined by the following delay coordinate

$$x(t + P) = \{x(t), x(t - \Delta t), \ldots, x(t - (d - 1)\Delta t)\},$$

(38)

where $d$ is an embedding dimension and $\Delta t$ is an embedding delay. The reconstruction of the chaotic dynamics is dependent on an appropriate choice of $d$ and $\Delta t$. For the Lorenz attractor, $\Delta t = 0.15$ and $d = 3$ are sufficient to reconstruct the chaotic dynamics [7]. $P$ is the number of prediction steps ahead and is set to $P = 1$.

4.3.2 Experimental Results

The effectiveness of the proposed method $LF$ is compared with $GFdisc$ for several discount schedules with $b = 150$ and $b = 1000$ in Table 8. Presented results are obtained from test runs with different grades of noise added to the training data, denoted by the noise standard deviation $SD$. The results show that $LF$ has robust short-term performances and long-term characteristics in all test cases. Especially, $LF$ shows very good long-term characteristics with no CD beyond 1.8. There are a few cases where $GFdisc$ has better short-term performance or long-term characteristics mainly for $b = 1000$, but it misses the robustness shown by $LF$. Also, better short-term prediction is not necessarily an indicator for good long-term characteristics. For some cases, $LF$ has better long-term characteristics even when short-term performance is worse than $GFdisc$’s. This phenomenon has also been noted in [3]. Similarly to the first experiment, $GFdisc$ again has difficulties to show stable performance for any discount schedule except ($a = 0.1$, $b = 1000$). It performs especially bad for two out of three discount schedules with $b = 150$. Also, $GFdisc$ shows no robustness for its long-term characteristics with only five CD values over 1.8.

Overall, $LF$ is performing robust in all test cases and its overall performance is clearly better than $GFdisc$.

In Fig. 4(a)–4(d), we show some examples for how different CD values relate to the long-term characteristics. 30,000 data points are presented in all four figures, but Fig. 4(d) seems sparse because the recursive prediction al-

| Discount/ | b=150 | b=1000 |
| Data Noise | LF (Prop.) | GFdisc | LF (Prop.) | GFdisc |
| a=0.1 | RMSE1 | RMSE2 | CD | RMSE1 | RMSE2 | CD | RMSE1 | RMSE2 | CD |
| SD=0.0 | 0.1676 | 0.1892 | 159 | 1.86 | 0.1922 | 0.2298 | 319 | 1.63 | 0.1740 | 0.1915 | 148 | 1.87 | 0.1410 | 0.1797 | 155 | 1.70 |
| SD=0.1 | 0.1679 | 0.1894 | 159 | 1.86 | 0.1696 | 0.2111 | 328 | 1.39 | 0.1750 | 0.1927 | 149 | 1.86 | 0.1407 | 0.1794 | 150 | 1.87 |
| SD=0.2 | 0.1713 | 0.1921 | 159 | 1.85 | 0.1530 | 0.1760 | 307 | 1.70 | 0.1766 | 0.1936 | 150 | 1.81 | 0.1446 | 0.1738 | 158 | 1.83 |
| SD=0.3 | 0.1844 | 0.1974 | 155 | 1.80 | 0.2003 | 0.2139 | 323 | 1.78 | 0.1851 | 0.1975 | 154 | 1.81 | 0.1671 | 0.1857 | 159 | 1.89 |
| a=0.01 | | | | | | | | | | | | | | | | |
| SD=0.0 | 0.1671 | 0.1890 | 157 | 1.88 | 0.5779 | 0.6212 | 421 | 1.13 | 0.1737 | 0.1914 | 148 | 1.87 | 0.1885 | 0.2113 | 166 | 1.65 |
| SD=0.1 | 0.1678 | 0.1887 | 156 | 1.86 | 0.5511 | 0.5580 | 408 | 0.34 | 0.1748 | 0.1926 | 149 | 1.86 | 0.2017 | 0.2379 | 170 | 1.65 |
| SD=0.2 | 0.1697 | 0.1910 | 159 | 1.85 | 0.6094 | 0.6480 | 425 | 0.81 | 0.1765 | 0.1935 | 150 | 1.81 | 0.1603 | 0.1909 | 171 | 1.82 |
| SD=0.3 | 0.1824 | 0.1956 | 155 | 1.80 | 0.6758 | 0.6848 | 442 | 1.25 | 0.1849 | 0.1973 | 154 | 1.81 | 0.1857 | 0.1997 | 184 | 0.00 |
| a=0.001 | | | | | | | | | | | | | | | | |
| SD=0.0 | 0.1666 | 0.1888 | 157 | 1.86 | 0.7612 | 0.8457 | 388 | 1.13 | 0.1737 | 0.1913 | 148 | 1.87 | 0.1797 | 0.2059 | 169 | 1.57 |
| SD=0.1 | 0.1672 | 0.1884 | 156 | 1.87 | 0.7746 | 0.8036 | 385 | 0.19 | 0.1748 | 0.1926 | 149 | 1.86 | 0.1942 | 0.2114 | 173 | 1.61 |
| SD=0.2 | 0.1692 | 0.1905 | 160 | 1.86 | 0.8037 | 0.8328 | 383 | 0.00 | 0.1765 | 0.1935 | 150 | 1.82 | 0.1634 | 0.1897 | 170 | 1.80 |
| SD=0.3 | 0.1818 | 0.1951 | 155 | 1.80 | 0.8262 | 0.8468 | 415 | 0.60 | 0.1849 | 0.1973 | 154 | 1.81 | 0.1867 | 0.2030 | 186 | 0.00 |

where $\xi(t)$ is a white Gaussian noise with zero mean and some standard deviation $SD$.
results show that our method is able to employ overall robust ploying global forgetting and dynamic model selection, and compared it with an established method for the NGnet em-
tages of applying dynamic model selection. Then, we have localized forgetting method to discuss differences and advan-
the proposed method when prone to negative interference.

data distributions have been tested to show the robustness of task. For the function approximation task, i.i.d. and non i.i.d.
tion approximation task and a chaotic time series forecasting proposed method has been evaluated by two testbeds, a func-
ronments and for handling redundancies in the model. The robustness of the NGnet in negative interference prone envi-
are made for the dynamic model selection to achieve better to it for the first time. In addition, several improvements

5. Conclusion

In this paper, we have proposed an online self-constructive NGnet with localized forgetting. Our major contributions are a modification of a previously proposed localized forgetting method and the application of dynamic model selection to it for the first time. In addition, several improvements are made for the dynamic model selection to achieve better robustness of the NGnet in negative interference prone environments and for handling redundancies in the model. The proposed method has been evaluated by two testbeds, a function approximation task and a chaotic time series forecasting task. For the function approximation task, i.i.d. and non i.i.d. data distributions have been tested to show the robustness of the proposed method when prone to negative interference. We have compared our method with a previously proposed localized forgetting method to discuss differences and advantages of applying dynamic model selection. Then, we have compared it with an established method for the NGnet employing global forgetting and dynamic model selection, and results show that our method is able to employ overall robust learning behavior while the global approach’s performance is highly dependent on the given learning environment.

Possible future work includes the development of an automatic parameter selection method to further ease the model adaptation of NGnets in online learning tasks, which is especially helpful when domain knowledge is limited.

References

Jana Backhus received her Master of Information Science and Technology in the field of Computer Science at the Graduate School of Information Science and Technology, Hokkaido University in 2014, and is currently working towards the Ph.D. degree at the same institution.

Ichigaku Takigawa received his D.E. from Hokkaido University in 2004 and is currently an Associate Professor at the Department of Computer Science and Information Technology, Graduate School of Information Science and Technology, Hokkaido University. His research interests include machine learning and data mining.

Hideyuki Imai received the D.E. from Hokkaido University in 1999. He is currently an Professor at the Department of Computer Science and Information Technology, Graduate School of Information Science and Technology, Hokkaido University. His research interests include statistical inference.

Mineichi Kudo received his Dr. Eng. degree in Information Engineering from Hokkaido University in 1988. Starting as an instructor at Hokkaido University in 1988, he is a professor since 2001. In 2001, he received the twenty-seventh annual pattern recognition society award together with professor Jack Sklansky. He was elected as a fellow of the International Association for Pattern Recognition on December 10, 2008. His current research interests include design of pattern recognition systems, image processing, data mining and computational learning theory. He is a member of the Pattern Recognition Society and the IEEE.

Masanori Sugimoto received the B.E., M.E., and D.E. degrees from the University of Tokyo, Tokyo, Japan, in 1990, 1992 and 1995, respectively. He is currently a Professor in the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan. His research interests include acoustic engineering, signal processing, artificial intelligence, and human-computer interaction technologies for designing smart systems and environments.