Context Learning with the Self–Organizing Map

Markus Varsta
Jukka Heikkonen
José del R. Millan

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Otaniemi 1997 / B4

Context Learning with the Self–Organizing Map

Markus Varsta and Jukka Heikkonen
Laboratory of Computational Engineering

José del R. Millan
Europen Commission, Joint Research Centre. Institute for Systems Informatics and Safety

Research Reports B4
April 1997

Laboratory of Computational Engineering
Helsinki University of Technology
Miestentie 3
P.O. Box 9400
FIN-02015 HUT
FINLAND

ISSN 1455-0474
ISBN 951-22-3543-9
LIBELLA OY

Abstract

In this paper a Recurrent Self-Organizing Map (RSOM) algorithm is proposed for temporal sequence processing. The RSOM algorithm is close in nature to the Kohonen's Self-Organizing Map, except that in the RSOM context of the temporal sequence is involved both in the best matching unit finding and in the adaptation of the weight vectors of the map via an introduced recursive difference equation associated for each unit of the map. The experimental results in the paper demonstrate that the RSOM is able to learn and distinguish temporal sequences, and that the RSOM algorithm can be utilized, for instance, in electroencephalogram (EEG) based epileptic activity detection.

1 Introduction

Temporal sequence processing (TSP) has gained a lot of interest in a number of applications varying from weather forecasting to time series processing, speech recognition and remote sensing. As a result many different computational techniques have been proposed for TSP. Statistical methods based on linear (e.g., AR and ARMA) and non-linear (e.g. NARMAX and MARS) time series models have been effectively used in many applications (see e.g. [2]). Recently neural networks have proven to be a promising alternative to them [12]. Much of the allure of neural networks in TSP has been their non-linearity, learning and generalization ability, and their power in handling a large volume of temporal data.

The simplest way of utilizing neural networks in TSP is to convert the temporal sequence into a concatenated vector via a tapped delay line and to feed the resulting vector as an input to a network [12]. This time-delay neural network approach, however, has its well know drawbacks, one of the most serious ones of being the difficulty to determine the proper length for the delay line. Therefore a number of dynamic neural networks models are designed for TSP to capture inherently the essential context of the temporal sequence without the need of external time delay mechanics. In these models the node and learning equations are often described by differential or difference equations and the interconnections between the nodes may include a set of feedback connections, i.e., the networks are recurrent in nature (see [12, 16]).

Most recurrent neural networks are trained via supervised learning rules. Only quite rare unsupervised neural networks models have been proposed for TSP, although, it can be foreseen that in temporal sequence analysis unsupervised neural networks could reveal useful information from the temporal sequences at hand in analogy to unsupervised neural networks' reported power in cluster analysis, dimensionality reduction and visualization of their 'static' input spaces. Moreover, in many TSP applications unsupervised learning could utilize more effectively the available temporal data than supervised learning methods, because no preclassification or prelabeling of the input data is needed. Based on the above the needs for the unsupervised learning methods in TSP are immense.

One of the most interesting unsupervised approaches for TSP is the Temporal Kohonen Map (TKM) [1] derived from the Kohonen's Self-Organizing Map [5, 8] algorithm. In the TKM the involvement of the earlier inputs in each node activity is defined as a recursive difference equation which defines the current node activity as a function of the previous node activations and the new input. In this paper an enhanced for the

TKM algorithm is proposed, and the resulting algorithm is named as the Recurrent Self-Organizing Map (RSOM). In brief the RSOM defines a difference vector for each unit of the map tp the best matching unit finding and adaptation of weights as a recursive difference equation of the new input shown to the map, the successive previous difference vectors, and the weight vector of the unit—otherwise the RSOM algorithm is close in nature to the SOM algorithm.

The rest of the paper is organized as follows. Section 2 provides a short literature review on the use of the Self-Organizing Map in TSP. The novel RSOM algorithm is described in detail in section 3. Section 4 shows that a RSOM of 9×9 nodes exposed repeatedly to nine possible inputs is able to distinguish the 81 possible sequences of length 2. Moreover, experiments in grouping of electroencephalogram (EEG) spectral features are used to demonstrate the potentials of the proposed RSOM algorithm.

2 Temporal Sequence Processing by the SOM

The Self-Organizing Map (SOM) [5, 8] is a vector quantization method to map patterns from an input space V_I onto typically lower dimensional space V_M of the map such that the topological relationships between the inputs are preserved. The topology preserving ability of the SOM means that the inputs which are close to each other in V_I tend to be represented by units close to each other on the map space V_M. The closeness between the units in the map is described by defining the map as a 1-, 2-, or multidimensional discrete lattice of units.

The training algorithm of the SOM is based on unsupervised learning, where one sample, the input vector $x(t)$ from the input space V_I, is selected randomly and compared to the weight vectors w_i of the units i in the map V_M. The best matching unit b to given input pattern $x(t)$ is selected using some criterion, such as

$$\|x(t) - w_b(t)\| = \min_{i \in V_M} \{\|x - w_i(t)\|\} , \tag{1}$$

where the parallel vertical bars denote to Euclidean vector norm. Initially all weight vectors are set randomly to their initial positions into the input space. During the learning phase the weights in the map are updated towards the given input pattern $x(t)$ according to

$$w_i(t+1) = w_i(t) + \gamma(t) h_{ib}(t)(x(t) - w_i(t)) , \tag{2}$$

where $i \in V_M$ and $\gamma(t)$, $0 \leq \gamma(t) \leq 1$, is a scalar valued adaptation gain. The "neighborhood kernel" function $h_{ib}(t)$ gives the excitation of unit i when the best matching unit is b. A typical choice for h_{ib} is a Gaussian function $h_{ib}(t) = exp(-\|r_i - r_b\|^2/\sigma(t)^2)$, where σ controls the width of the function, and r_i and r_b are the coordinates of units i and b in the map lattice, respectively. During learning the function h_{ib} normally approaches towards delta function, i.e., σ approaches slowly towards zero in time. To guarantee convergence of the algorithm, the gain $\gamma(t)$ should decrease as a function of time according to conditions [14]: $\lim_{t\to\infty} \int_0^t \gamma(t')dt' = \infty$ and $\lim_{t\to\infty} \gamma(t) = 0$. After a number of input vector presentations the mapping has formed, i.e., the weight vectors will specify the centers of clusters covering the input space and the point density function of these centers tend to approximate closely the probability density function of the input space V_I.

The known properties of the Self-Organizing Map has been excessively used in many applications for visualization of the multi dimensional pattern spaces. In TSP the mapping can be applied to find a chain of best matching units for subsequent patterns to produce a time varying trajectory on the map space. This kind of trajectory approach has been employed, for instance, for visualization of speech signals and speech signal variations in time [9]. Moreover, the approach has found its way to process monitoring and control [15].

There have also been attempts to utilize contextual information of pattern sequences in connection to the SOM algorithm. The time delay approach is the most simplest example of that kind in which the pattern sequences are concatenated to static input patterns before feeding them to the SOM. In [4] such approach was reported to improve phonemic recognition accuracies compared to cases when only one sample pattern vector at time was utilized in recognition.

Two of the interesting variations of the SOM algorithm for TSP are the Hypermap [6] and Self-Organizing Operator Maps [7] architectures. The Hypermap can use several levels of information in pattern sequences for best matching unit finding. For instance, the context in which the pattern occurred is first used to define a subset of units from which subset the final best matching unit is searched by the pattern itself.

The Self-Organizing Operator Map permits to analyze the temporal sequence directly by associating an operator to each unit of the map. These operators should be defined according to the applications, however, the derivation of learning laws for arbitrary operators can be rather complicated. In [7] a genetic algorithm type method was proposed to adapt operator parameters. If an error function can be defined for an operator, such as in [4] where the Linear Prediction Coding was used as the operator function, the parameters of the operator can be trained via gradient optimization approach. Operator maps have been used, for instance, in speech analysis (see e.g. [4]).

The Self-Organizing Maps with leaky integration of signals have the closest resemblance to our RSOM algorithm to be described in the next section. In [3] a three stage computational approach was proposed to TSP. First, input pattern $x(t)$ is mapped through the first SOM V_1 whose transformed unit responses $r(i,t)$ are collected into a "leaky integrator" memory, where the response values were adapted by equation $s(i,t) = s(i,t-1) + \beta(r(i,t) - s(i,t))$, where i is a unit of the map V_1, $s(i,t)$ is the memory value for the unit i at time t, and β is the adaptation factor of the memory. Stored unit samples are then used in the computation of the response of the second SOM V_2. This integration approach has been shown to improve phoneme recognition accuracies. Some analogy to the above approach can be found in the Temporal Kohonen Map algorithm [1].

The Temporal Kohonen Map (TKM) [1] determines the response of each unit i of the map for the given input $x(t)$ in terms of a potential $V_i(t) = dV_i(t-1) - (1/2)\|x(t) - w_i(t)\|^2$, where $0 < d < 1$. The involvement of successive previous inputs is explicit by stating a general solution for the potential equation: $V_i(t) = -(1/2)\sum_{k=0}^{n-1} d^k \|x(t-k) - w_i(t-k)\|^2 + d^n V_i(t-n)$. The weight update at time t in the TKM is analogical to the normal SOM learning, except that the best matching unit b at time t is searched according to $V_b(t) = \max_i\{V_i(t)\}$. In [1] it was demonstrated that a TKM of 4×4

units is able to classify the 16 possible sequences of length 2 by repeatedly showing the map with four possible binary inputs. Also some preliminary results in the formation of semantic word maps from sentences by the TKM algorithm was reported in [1].

3 The Recurrent Self-Organizing Map

The Temporal Kohonen Map [1] does not directly use the temporal contextual information of input sequences in weight updating at time t: The involvement of the previous inputs is only taken into account in the best matching unit searching. Our goal, and also the main novelty of the developed Recurrent Self-Organizing Map (RSOM) algorithm, is to utilize the temporal nature of the given input sequence both in the best matching unit searching and in the adaptation of the weights. This can be accomplished by associating the following recursive difference equation for each unit i of the map V_M to compute difference vector $y_i(t)$ from the given input $x(t)$ at time t:

$$y_i(t) = (1-\alpha)y_i(t-1) + \alpha(x(t) - w_i(t)) \ , \tag{3}$$

where $w_i(t)$ is the weight vector of the unit i, $i \in V_M$, and α, $0 < \alpha \leq 1$, is the weighting factor determining the effect of the earlier difference vectors and the new input vector in the computation of $y_i(t)$. When α is close to 1, the system of Eq. 3 corresponds to a short term memory, whereas when α decreases towards zero Eq. 3 describes a long term memory. It is possible to state a general solution for the Eq. 3 in a form

$$y_i(t) = \alpha \sum_{k=0}^{n-1} (1-\alpha)^k (x(t-k) - w_i(t-k)) + (1-\alpha)^n y_i(t-n) \ , \tag{4}$$

which explicitly shows the involvement of the earlier inputs. It should also be noted that it is possible to replace Eq. 3 by any recursive difference equation more suitable for the given application in computing of $y_i(t)$.

By replacing $x'_i(t) = x(t) - w_i(t)$, Eq. 3 can be written in a familiar form

$$y_i(t) = (1-\alpha)y_i(t-1) + \alpha x'_i(t) \ , \tag{5}$$

which describes an exponentially weighted linear IIR filter with the impulse response $h(k) = \alpha(1-\alpha)^k$, $k \geq 0$. For further analysis of the Eq. 5, see e.g. [13].

Now the best matching unit b at time t is searched by

$$y_b = min_i\{\|y_i(t)\|\} \ , \tag{6}$$

where $i \in V_M$. Note again that by setting $\alpha = 1$ Eq. 6 simplifies to the best matching unit searching equation (Eq. 1) of the SOM algorithm corresponding to the case when no history is used in learning.

The adaptation of the weight vectors $w_i(t)$ is based on the difference vectors $y_i(t)$:

$$w_i(t+1) = w_i(t) + \gamma(t) h_{ib}(t) y_i(t) \ , \tag{7}$$

where $i \in V_M$ and $\gamma(t)$ and $h_{ib}(t)$ has the same meaning as in Eq. 2. Again, by setting $\alpha = 1$, Eq. 7 reduces to the normal SOM weight adaptation law. In learning the same principles of the normal SOM learning should be used, so that after a number of input sequence presentations the mapping will form. After learning the weight vectors are frozen meaning that the functionalities of units $i \in V_M$ with constant weight vectors w_i can be described and analysed by Eq. 5.

Some practical algorithmical things related to the RSOM learning of temporal sequences should be noticed. The first ones are related to the presentation of the input sequence $\{x(1), x(2), \ldots, x, (N-1), x(N)\}$. At each learning cycle one can randomly select a time t_c, $t_c \geq n$ for sampling a subsequence $\{x(t_c - n + 1), x(t_c - n), \ldots, x(t_c)\}$ of n samples to be presented to the RSOM starting at sample $x(t_c - n + 1)$. The value of n has to be sufficient high so that the sample $x(t_c - n + 1)$ has only minimal effect to the total value of $y_i(t_c)$, and at each learning cycle the memory should be reset to zero via setting $y_i(t_c - n) = 0$, $i \in V_M$. Moreover, it is preferable after each weight vector adaptation at time t to update the difference vectors according to $y_i(t) = (1-\alpha) y_i(t-1) + \alpha(x(t) - w_i(t+1))$, which corresponds to the updated status of the map, although in practice this may only have minor effect to the final organization of the map due to the fine tuning of the weights at the later stage of the learning.

4 Experimental Results

Two experiments are reported to show the potentials of the proposed RSOM algorithm. The first one is related to learning of sequences from a synthetic data, whereas the second one demonstrates the usefulness of the RSOM in an EEG spectral feature clustering.

In the first experiment a RSOM map of 9×9 units is repeatedly exposed with the following 9 input pairs: $(1,1), (1,3), (1,5), (3,1), (3,3), (3,5), (5,1), (5,3), (5,5)$. If the input pairs $(1,1)$, $(1,3)$, etc. are denoted by numbers $1, 2, \ldots, 9$, respectively, it is expected that the RSOM should learn to map all possible XY, $X = 1, \ldots, 9$ and $Y = 1, \ldots, 9$, sequence combinations into separate units, where Y corresponds to the most recently given input. The RSOM learning parameter α was set to value 0.7, and Luttrell's [10] map building method was employed. It should be noted that only in the beginning of the learning the memory was reset to zero via setting $y_i = 0$, $\forall i$.

Fig. 1 shows the resulting map with the XY marking of the best matching units of the possible 81 different XY combinations. Each XY in figure is drawn on the location of the two dimensional weight vector of the best matching unit of the corresponding XY input pair. Fig. 1 demonstrates that the best matching units are determined by the combination of the two most recent inputs, and that the map can perform a perfect classification of all the 81 possible sequences of length 2. Moreover, it can be observed that the last input, Y, of the XY pair determines the map block of 3×3 units from which block the winning unit for XY is distinguished according to the input X.

The second RSOM experiment is related to EEG spectral feature clustering for epileptic activity detection. EEG is an important clinical tool for diagnosing, monitoring and managing neurological disorders related to epilepsy. Since epileptiform activity can be noticed in EEG as a clearly distinguishable transient waveforms, wavelets have been used efficiently to extract suitable features for epilepsy detection. It can also be foreseen that the temporal context in which the epileptic activity occurs in EEG should provide valuable information for detecting and analysing of the activity, and hence the RSOM algorithm should be a potential tool for EEG based epileptic activity analysis.

The sampling rate of the EEG data used in the test was 200 Hz. For spectral feature extraction at time t a feature extraction window W^t of 256 samples was collected from the original EEG sequence S_{EEG} as follows: $W^t(i) = S_{EEG}(t-127+i), i = 0, \ldots, 255$. The wavelet transform of a continuous signal $g(t)$ with a wavelet $\Psi((t-b)/a)$ is given as: $C^t_{a,b} = \int_{-\infty}^{\infty} g(t)\Psi((t-b)/a)\, dt$, where a is the scaling and b is the dilation factor of the wavelet Ψ. In our case two Daubechies' mother wavelets, Daub_4 and Daub_{12}, were employed and the discrete wavelet transform for each window W^t was done using Mallat's "with holes" [11] algorithm. Finally, total of sixteen energy features f^t were determined for each W^t by computing the squared sum of both Daub_4 and Daub_{12} wavelet coefficients at each scale.

The extracted EEG features were first clustered by four SOMs of 3×3, 5×5, 9×9 and 17×17 units in two dimensional lattice. The Luttrell's method [10] was used in training. The training data contained total of 150987 16-dimensional feature vectors, among which 5430 patterns correspond to epileptic activity. After training each unit of the map was labeled according to the plurality rule to belong either to "normal" or "epileptic" activity, and the SOMs were used as classifiers to evaluate the discrimination potentials of the feature clusters. Note that for labeling the number of epileptic activity samples mapped into each unit i was multiplied by the factor 145557/5430 for equal weighting of both classes. Table 1 shows the confusion matrices of the four SOM classifiers; the diagonal entries give the correctly classified samples. These results were compared to the RSOMs of same sizes. The α parameter was set to value 0.6, and in each learning cycle 5 past feature values f^{t_c-64}, f^{t_c-48}, f^{t_c-32}, f^{t_c-16}, and f^{t_c} were shown to the map with randomly selected time t_c. The RSOM maps were taught with the Luttrell's approach [10]. When the labeled RSOMs were employed to classify training samples, the results were better than those of normal SOMs with the same number of units (see Table 1). These clustering results suggest that the use of the context in the EEG based epileptic activity detection might improve the analysing/classification results, and the RSOM may be a valuable tool for the purpose.

5 Conclusions

We have proposed a Recurrent Self-Organizing Map (RSOM) algorithm for temporal sequence processing (TSP). The RSOM can be used as a flexible tool in diverse tasks involving temporal data. The preliminary results reported have shown the potentials of the method, although, more detailed experiments in different TSP tasks with varying RSOM learning parameter values have to be conducted before solid usability conclusions of the proposed method.

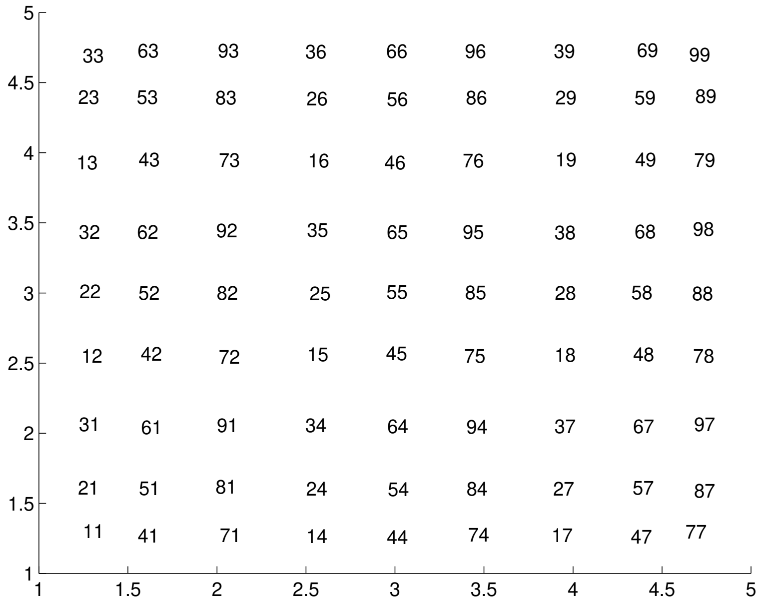

Figure 1: The RSOM map of 9×9 units (see text for details).

# of units	SOM		RSOM	
3×3	126966	18591	129269	16224
	405	5025	321	5109
5×5	131794	13763	133389	12104
	443	4987	380	5050
9×9	133820	11737	134660	10833
	394	5036	317	5113
17×17	134465	11092	135935	9558
	338	5092	284	5146

Table 1: SOM and RSOM confusion matrices obtained in the clustering of EEG spectral features (see text for details).

Acknowledgements

We wish to thank Dr.Tech. Alpo Värri from Tampere University of Technology, Tampere, Finland, for generously providing us the scored EEG recordings of the ANNDEE project.

References

[1] G.J. Chappell and J.G. Taylor. The Temporal Kohonen Map. Neural Networks,

Vol. 6, pages 441–445, 1993.

[2] N.A. Gershenfeld, A.S. Weigend. The Future of Time Series: Learning and Understanding. In *Times Series Prediction: Forecasting the Future and Understanding the Past*, Eds. A.S. Weigend and N.A. Gershenfeld, Addison-Wesley, pages 1–70, 1993.

[3] J. Kangas. Time-Dependent Self-Organizing Maps for Speech Recognition. In *Artificial Neural Networks*, Eds. T. Kohonen, K. Mäkisara, O. Simula, J. Kangas, North-Holland, pages 1591–1594, 1991.

[4] J. Kangas. On the Analysis of Pattern Sequences by Self-Organizing Maps. *Dr. Tech. thesis*, Helsinki University of Technology, 1994.

[5] T. Kohonen. Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, Vol. 43, pages 59–69, 1982.

[6] J. Kohonen. The Hypermap Architecture. In *Artificial Neural Networks*, Eds. T. Kohonen, K. Mäkisara, O. Simula, J. Kangas, North-Holland, pages 1357–1360, 1991.

[7] T. Kohonen. *Things You Haven't Heard about the Self-Organizing Map.* In *Proc. IEEE Int. Conf. on Neural Networks*, pages 1147–1156, 1993.

[8] T. Kohonen. *Self-Organizing Maps.* Springer-Verlag, 1995.

[9] L. Leinonen, J. Kangas, K. Torkkola and A. Juvas. Dysphonia Detection by Pattern Recognition of Spectral Composition. Journal of Speech and Hearing Research, Vol. 35, pages 287–295, 1992.

[10] S.P. Luttrell. Image Compression Using a Multilayer Neural Network. *Pattern Recognition Letters*, Vol. 10, pages 1–7, 1989.

[11] S.G. Mallat. A Theory of Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 11, pages 674–693, 1989.

[12] M.C. Mozer. Neural Networks Architectures for Temporal Sequence Processing. In *Times Series Prediction: Forecasting the Future and Understanding the Past*, Eds. A.S. Weigend and N.A. Gershenfeld, Addison-Wesley, pages 243–264, 1993.

[13] J.G. Proakis and D.G. Manolakis. *Digital Signal Processing: Principles, Algorithms, and Applications.* Macmillan Publishing Company, 1992.

[14] H. Ritter and K. Schulten. Convergence Properties of Kohonen's Topology Preserving Maps: Fluctuations, Stability, and Dimension Selection. *Biological Cybernetics*, Vol. 60, pages 59–71, 1988.

[15] V. Tryba and K. Goser. Self-Organizing Feature Maps for Process Control in Chemistry. In *Artificial Neural Networks*, Eds. T. Kohonen, K. Mäkisara, O. Simula, J. Kangas, North-Holland, pages 847–852, 1991.

[16] A.C. Tsoi and A.D. Black. Locally Recurrent Globally Feedforward Networks: A Critical Review of Architectures. IEEE Trans. on Neural Networks, Vol. 5, No. 2, pages 229–239, 1994.