A

Abstract

A novel neural model made up of two self-organizing map nets — one on top of the other — is introduced and analysed experimentally. The model makes an effective use of context information, and that enables it to perform sequence classification and discrimination efficiently. It was successfully applied to a set of contrived sequences, and also to a real sequence — the third voice of the sixteenth four-part fugue in G minor of the Well-Tempered Clavier (vol. I) of J. S. Bach. The model has application in cognitive domains which demand classifying either a set of sequences of vectors in time or sub-sequences into a unique and large sequence of vectors in time.

1 Introduction

Several researchers have extended the self-organizing feature map model (Kohonen, 1989) to classify sequential information. The problem involves either classifying a set of sequences of vectors in time or recognizing

in their model. In the following sections, we shall describe the model, and present two experiments. In the first one, the model is applied to a small scale problem in order to analyse its behaviour. In the second, it is applied to a large scale real example.

2 The model

The

$$\mathbf{X}(t) = \mathbf{V}(t) + \delta_1 \mathbf{X}(t-1) \tag{1}$$

where $\delta_1 \in (0,1)$ is the decay rate. The winning unit i^* in the map[2] is the unit which has the smallest distance $\Psi(i^*,t)$. For each output unit i, the distance $\Psi(i,t)$ between the input vector $\mathbf{X}(t)$ and the unit's weight vector $\mathbf{W}_i$ is given by

$$\Psi(i,t) = \sum_{j=1}^{m} [x_j(t) - w_{ij}(t)]^2 \tag{2}$$

Each output unit i in the neighbourhood N^* of the winning unit i^* has its weight $\mathbf{W}_i$ updated by

$$w_{ij}(t+1) = w_{ij}(t) + \alpha \Upsilon(i)[x_j(t) - w_{ij}(t)] \tag{3}$$

where $\alpha \in (0,1)$ is the learning rate. $\Upsilon(i)$ is the *neighbourhood interaction function* (Lo & Bavarian, 1991), a gaussian type function, and is given by

$$\Upsilon(i) = \kappa_1 + \kappa_2 e^{-\frac{\kappa_3 [\Phi(i,i^*)]^2}{2\sigma^2}} \tag{4}$$

where κ_1, κ_2, and κ_3 are constants which confer the shape to the function. We have set κ_1, κ_2, and κ_3 to be 0.1, 0.7, and 10 in our experiments. σ is the radius of the neighbourhood N^*, and $\Phi(i,i^*)$ is the distance in the map between the unit i and the winning unit i^*. The distance $\Phi(i',i'')$ between any two units i' and i'' in the map is calculated according to the maximum norm,

$$\Phi(i',i'') = max\left\{\left|l' - l''\right|, \left|c' - c''\right|\right\} \tag{5}$$

where (l',c') and (l'',c'') are the coordinates of the units i' and i'' respectively in the map.

The neighbourhood interaction function has proved to be useful, indeed. It provokes two main effects. Fi(8.3n)TjR2770.24Tf5.000.7(39g8444.6-4.55977Td(0)TjRtdigh)999.

by reducing the number of epochs required. Second, it improves the quality of the map by

the radius in 1. The coarse-mapping phase took 20%, and the fine-tuning phase took 80% of the total number of epochs. The initial weights were given randomly, in the range between 0 and 0.1, to all SOMs.

Different decay rates were tried. In the bottom SOM of model II, they ranged from 0.4 to 0.7, and in the top SOM, from 0.7 to 0.95. In the model I, the decay rate ranged from 0.7 to 0.95. The input layer of the model I and of the bottom SOM of model II held two units. Bits 0's of the sequences were represented as (1,0), whereas bits 1's were represented as (0,1).

Model I was tested with three different map sizes, 9×9, 15×15, and 21×21, trained in 400, 700, and 1000 epochs respectively. In model II, the map sizes were set to 6×6 (trained in 250 epochs) and 9×9 (trained in 400 epochs) to the bottom and top SOM respectively. The transfer function Λ was given by equation 7, with $N^* = \{i^*\}$.

The best results of models I and II are displayed in the tables 1 and 2 respectively. A sequence $\mathbf{S_a}$ is said to have the same classification as that of the referential sequence $\mathbf{S_r}$ if the distance $\Phi(i_a^*, i_r^*) < 2$, where i_a^* and i_r^* are the (last) winning units of $\mathbf{S_a}$ and $\mathbf{S_r}$.

Table 1: Results for model I (first experiment)

Map Size	Decay Rate	No. Miscl.
9×9	0.7	

bits of a sequence because the contribution of these first bits to the classification of the sequence is very low. For instance, let $\mathbf{S_a}$ = 100000 and $\mathbf{S_b}$ = 010000 be two sequences. Considering a decay rate of 0.8, the activations of the two input units would be 3.362 and 0.328 after the entrance of the last bit of $\mathbf{S_a}$. The activations would be 3.280 and 0.410 for $\mathbf{S_b}$. The differences in the activations between $\mathbf{S_a}$ and $\mathbf{S_b}$ are not relevant, and probably the sequences would be classified as identical by model I.

The problem with model I is that the SOM sees just bits in its input. Yet, its performance would be much improved if the input not only represented bits, but also the context where they appeared. Different input units would then be activated depending upon the order that the bits were input. For example, considering a representation that includes three bits at most, $\mathbf{S_a}$ and $\mathbf{S_b}$ would be represented by table 3. As the representation makes a clear distinction between the beginnings of $\mathbf{S_a}$ and $\mathbf{S_b}$, it helps model I to distinguish between the two sequences as well.

Table 3: Context representation for two binary sequences

Seq.	time: 1	time: 2	time: 3	time: 4	time: 5	time: 6
$\mathbf{S_a}$	(1)	(10)	(100)	(000)	(000)	(000)
$\mathbf{S_b}$	(0)	(01)	(010)	(100)	(000)	(000)

The idea of encoding context in the representation to distinguish variations in sequences is not original. Wickelphones (Wickelgren, 1969) and Wickelfeatures (Rumelhart & McClelland, 1988) are examples of such a representation. Model II also makes use of the representation, and that is the reason why its performance is much superior than that of model I. The top SOM of model II sees bits and over all, context in its input. As opposed to Wickelphones and Wickelfeatures, the representations in the input layer of the top SOM are not handmade beforehand, but instead, they are built up by the bottom SOM. The advantage of this approach is twofold. First, one does not need to worry about encoding context once the bottom SOM is in charge of making an internal representation of context in its map. Second, only the representations required by the application will be built up by the bottom SOM reducing thus, the necessary number of units in the input layer of the top SOM.

The size of context is the size of memory of past inputs, that means,

the maximum number of past input bits that the bottom SOM may recognize. The size of context is directly dependent of the decay rat-341.03913.6801Td13.20The

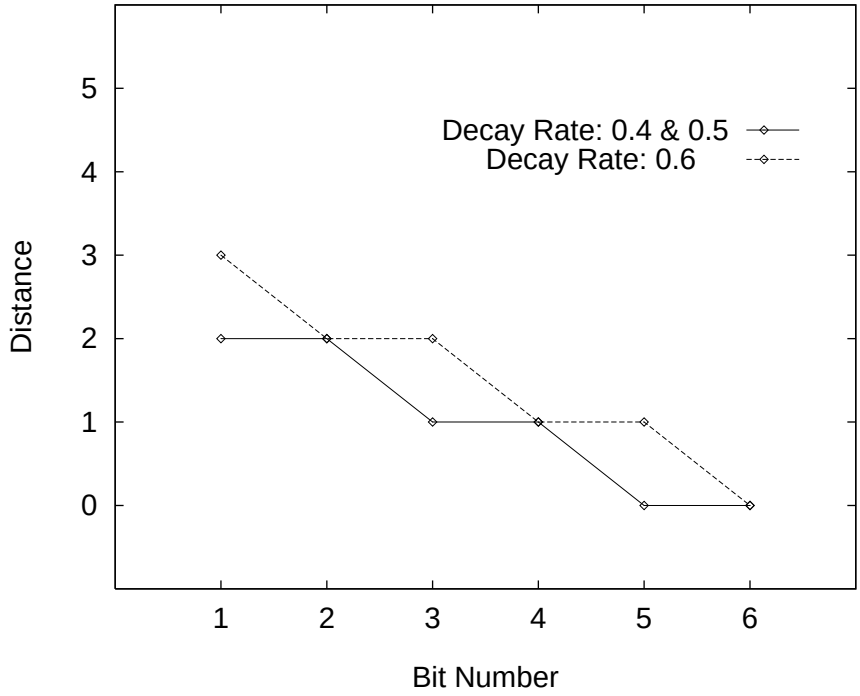


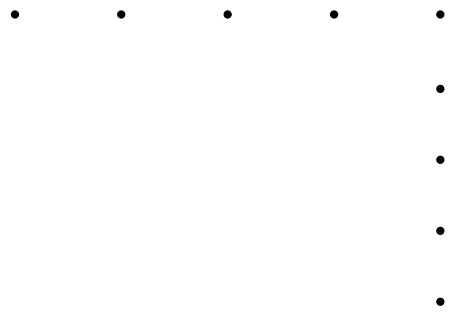
Figure 2: Distances between the winning units of two binary sequences

4 Second experiment

The second experiment was on recognizing sub-sequences into a large and unique input sequence. The input data consisted of a sequence of musical intervals that corresponded to the third voice of the four-part fugue in G minor of Bach (Bach, 1989). The *theme* of the fugue (figure 4), a *referential sub-sequence*, was divided into two parts — *theme I* and *theme II*. Several perfect and modified instances of theme I and II occur in the third voice of the fugue.

The experiment pursued two aims. First, to verify whether models I and II recognize all instances of theme I and II in the third voice of the fugue. Second, to verify whether any other sub-sequence, which was not an instance, was not misclassified as theme I or II.

The training of the two SOMs of model II and the SOM of model I



the top SOM, from 0.7 to 0.9. In model I, the decay rate ranged from 0.7 to 0.9. We present here only the results using decay rates of 0.5 and 0.85 respectively for the bottom and top SOM of model II, and 0.85 for the SOM of model I.

The SOM of model I was tested with map size of 18×18, and was trained in 850 epochs. In model II, the map sizes were set to 15×15 (trained in 700 epochs)

were also present in the theme.

5 Conclusion

A novel neural model made up of two self-organizing map networks — one on top of the other — is presented. It has application in domains which demand classifying either a set of sequences of vectors in time or sub-sequences into a unique and large sequence of vectors in time. The model makes an effective use of context information, and that enables it to perform sequence classification and discrimination efficiently. Despite the good results, it is still open to further research. In principle, the model could have any number of self-organizing map nets — the more nets, the more similar and longer the sequences of vectors in time which could be recognized.

Acknowledgements

This research was fully supported by CAPES, Brazil.

Reference

Bach, J. S. (1989). *Das Wohltemperierte Klavier*. Vol. 1. BWV 846–869. Bärenreiter Kassel, Basel, Germany.

Carpenter, G. A., & Grossberg, S. (1987). ART2: self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, *26*(23), 4919–4930.

Chappell, G. J., & Taylor, J. G. (1993). The temporal Kohonen map. *Neural Networks*, *6*, 441–445.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Gjerdingen, R. O. (1991). Using connectionist models to explore complex musical patterns. In Todd, P. M., & Loy, D. G. (Eds.), *Music and Connectionism*, pp. 138–149. The MIT Press, Cambridge, MA.

James, D. L., & Miikkulainen, R. (1995). SARDNET: a self-organizing feature map for sequences. In Tesauro, G., Touretzky, D. S., & Leen,