This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30044542

Reproduced with the kind permissions of the copyright owner.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Copyright: 1999, IEEE
An Empirical Study of Neighbourhood Decay in Kohonen’s Self Organising Map

Russell Keith-Magee, Svetha Venkatesh
School of Computing, Curtin University of Technology, Perth, Western Australia
{keithmag|svetha}@cs.curtin.edu.au
Masahiro Takatsuka
Department of Geography, The Pennsylvania State University
masa@gis.psu.edu

Abstract

In this paper, empirical results are presented which suggest that size and rate of decay of region size plays a much more significant role in the learning, and especially the development, of topographic feature maps.

Using these results as a basis, a scheme for decaying region size during SOM training is proposed. The proposed technique provides near optimal training time. This scheme avoids the need for sophisticated learning gain decay schemes, and precludes the need for a priori knowledge of likely training times. This scheme also has some potential uses for continuous learning.

1 Introduction

The theory underlying self organisation was originally grounded in a biological process - the topological organisation of stimuli observed in the cortex. The work of [8] provided an algorithmic model of the mechanism by which this topographical organisation can occur. Malsberg’s model was biologically motivated; however, in recent times, the focus of SOM research has moved away from the biological origins of the self organising processes, to concentrate on studies of performance of Kohonen’s computational approximation of the underlying biological process.

Kohonen’s algorithm involves two key parameters for operation; the learning gain (also known as learning rate), and the neighbourhood size (also known as region size). A wide range of decay schemes have been proposed for these parameters; however, to date, analysis of SOM performance has concentrated on optimising the decay of learning gain within Kohonen’s algorithm. These schemes generally fall into one of two categories; those that are computationally intensive, and those that require a priori knowledge of the data set. However, optimisation of neighbourhood size has been largely ignored in the literature, being largely limited to defining conditions for ordering, not optimising convergence time.

In this paper, empirical results are presented which suggest that size and rate of decay of region size plays a much more significant role in the learning and development, of topographic feature maps. A scheme for decaying region size during SOM training is presented. The proposed technique provides near optimal training time, without the need for extensive parameter tuning. The proposed scheme avoids the need for sophisticated schemes for learning rate decay, and precludes the necessity for a priori knowledge of the likely required training time. This scheme is demonstrated using a range of data sets, on a range of map sizes.

1.1 Kohonen’s SOM Model

As a computational technique, Kohonen’s SOM algorithm [5, 10], and variants thereof (for a summary, see [7], Ch. 5) have been found to be an extremely useful method of arriving at a topological organisation for a wide range of applications (See Ch. 7 of [7] for an extensive list).

Kohonen’s SOM algorithm is computationally quite simple; instead of requiring a complicated system of equations balancing individual neuron dynamics and the dynamics of the map as a whole (as was required by previous SOM algorithms [8, 1]), it can be stated as a single iterative weight update rule [6, 7]. This computational approximation provides a significant computational advantage, but do not significantly undermine the utility of Kohonen’s algorithm (as is evident by the plethora of SOM application literature).

Kohonen’s algorithm involves two key parameters for operation; the learning gain (also known as learning rate), and
the neighbourhood size (also known as region size). A wide range of decay schemes have been proposed for these parameters. These schemes generally fall into one of two categories. Firstly, there are parameter schemes that require a priori knowledge of the data set. These decay schemes were amongst the first to be suggested, as they use a simple arithmetic or geometric decay over time. Initially, schemes featuring linear or inverse decay schemes [6] were proposed; other schemes feature adjustments based on semi-empirical measures [9]. Schemes have also been proposed which perform a posteriori statistical analysis of a learning regime to establish future training parameters [11]. Secondly, computationally intensive schemes have been proposed; these schemes use online filtering mechanisms such as Kalman filters to adjust learning parameters during the learning process [3]. However, while these schemes permit learning in fewer training epochs, the computational cost of the adjustment process can exceed the computational benefit of a shorter training regime.

It is worthy of note that the majority of schemes proposed for parameter decay concentrate on optimising the decay of learning gain within Kohonen’s algorithm. Optimisation of neighbourhood size has been largely ignored in the literature, being largely limited to defining conditions for ordering, not optimising convergence time.

2 The role played by region size

In this section, a series of experiments are performed to establish the role played by region size in the learning process. The data used for each experiment in this section are a set of two dimensional real valued Cartesian coordinate vectors defining a grid on a plane. In this paper, an $n \times n$ grid refers to the data set consisting of the points $(1, 1), (1, 2), \ldots, (1, n), (2, 1), \ldots, (n, n)$

The SOM used in these experiments consists of a square grid of neurons, with 2 real valued input neurons (i.e., inputs are not thresholded or normalised by a sigmoid function). Weights are instantiated with random values in the range $[0; 0.1]$.  

The algorithm used in these experiments is the same as that presented in Section 3.1 of [7]. In all cases, the learning gain parameter was set to $\alpha(t) = 0.3$. In this way, the role played by region size decay can be established independent of learning gain decay.

In order to test the role played by region size on the learning process, the effect of large and small region sizes on the convergence process must be established. To establish these roles, grid data sets with sizes ranging from $5 \times 5$ to $20 \times 20$ were learnt on maps ranging in size from the smallest map that could accommodate all the data, to maps requiring a significant computational load to evaluate.

2.1 Results

Although experiments performed in this experiment covered a wide range of map and data set sizes, results observed during training demonstrate two distinct behaviours, each with unique benefits. Large region sizes rapidly generate a consistent global topology, but do not generate a precise local topological mapping. Small region sizes, on the other hand, allow for the fine tuning of local topologies, but are not good at arranging global topologies.

Figure 1 is an example of the use of a large region size; using this technique, data points are rapidly spread to the extent of the map, resulting in an initially good topological mapping. However, little improvement is seen on this initial topology; the rough topology reached in the first few epochs of training does not improve with further training.

Figure 2 is an example of the use of a small region size. While global topologies do not develop quickly using the small region size, local topologies are evident almost immediately. Winning neurons can be divided up into small groups, each of which exhibits a locally consistent topology. However, the topologies of adjacent groups do not necessarily correlate; the global topology is therefore twisted.

Attempts to generate precise global topologies quickly require a combination of these two characteristics; the rapid spreading of points across the map afforded by a large region size, and the fine control afforded by a small region size. A learning scheme which utilises a combination of these behaviours should provide near optimal learning behaviour.

This result is supported by the observations made by [8] in his experiments on biologically faithful self organising map models. Two important qualities are demonstrated by Malsberg’s experiments. Firstly, learning was achieved without the need for a decaying learning gain parameter. In most
theoretical and experimental studies since, variation in gain has been considered the most significant factor controlling learning.

Secondly, global topographical organisation was only observed on small maps; in large map simulations, local topographical organisation occurred, but global topology was erratic. The cause of this result was theoretically demonstrated by [1]. In this proof, it was demonstrated that a single peak of excitation will result, provided the domain of the lateral connection function spans the entire map; a lateral connection function with a smaller range will result in multiple peaks occurring on a map. During Malsberg’s experiments, the range of the lateral connection function was only a small number of neurons, and was not decayed. Consequently, on small maps, the connection function spanned the entire map, and global organisation was observed, on larger maps, the small connection function did not span the map, and many areas of excitation occurred, resulting in erratic global topology.

3 Optimizing Kohonen’s SOM

In this section, the results obtained in the previous sections are applied to Kohonen’s algorithm, by way of a novel method for decaying region size during training which provides near optimal training times on every operation of the algorithm; this technique has the additional benefit of removing the need for extensive tuning of key learning parameters.

3.1 Implementation

SOM learning can be broadly characterised by two behaviours: global topological organisation, and local topological fine tuning. These two behaviours can be obtained by using a large and small region size, respectively. However, both of these behaviours are desirable during the learning process. It therefore seems reasonable to propose a scheme for decaying region size during training which utilises both of these characteristics.

In these experiments, a ‘step’ region decay scheme is used. This scheme consists of 2 phases:

**Phase 1:** $\sigma = \lceil N/2 \rceil$, for N epochs

**Phase 2:** $\sigma = R$, for as many epochs as are required to converge on a solution.

In this formulation, $R$ is the resolving distance of the SOM: the distance between peaks that would be observed if every data point was spread evenly over them map.

The SOM used in these experiments consists of a square grid of neurons, with 2 real valued input neurons (i.e., inputs are not thresholded or normalised by a sigmoid function). Weights are instantiated with random values in the range $0-0.1$.

The algorithm used in these experiments is the same as that presented in Section 3.1 of [7]. When using the step region decay function, the learning gain parameter was set to $\alpha(t) = 0.3$.

To test for convergence in these experiments, the goodness of fit measure of [4] was used. This measure evaluates both continuity in the mapping, and quantization error (the Euclidean distance between the training vector and the weight vector), and can be applied to any data set. Smaller values indicate better mappings. It should be noted that this measure is not a metric; values of goodness are not comparable between data sets.

3.2 Experimental Results

Two experiments were performed. Firstly, grid data sets of various sizes were tested on a range of map sizes, to establish an appropriate value for $N$, the length of phase 1 of training.

Secondly, a range of data sets were tested to demonstrate the flexibility of the step region size decay scheme, using this value of $N$. The proposed decay scheme was tested using a range of data sets. Two of these data sets are presented in this paper: 5 x 5 grid data, and a minimal spanning tree.

3.3 Evaluating decay epoch

In this experiment, grid data sets of various sizes were learnt by large and small SOM’s. These experiments represent a
Despite the wide range of map and data set sizes in these experiments, a consistent behaviour is demonstrated in all of these experiments. As a result of training with the large region size, the goodness of fit of each map rapidly decays to a suboptimal level, and becomes asymptotic at this value. When the region size is reduced, the goodness of fit of the map again decreases rapidly, becoming asymptotic at a goodness of fit value representative of a well formed map.

The first asymptotic value is reached within a small number of epochs; in all experiments, the asymptotic value has been achieved by the tenth epoch. The second asymptotic value is reached approximately 10 epochs after the reduction of region size. It should therefore be possible to train grid data sets in approximately 20 training epochs; the consistency found in the results of this section suggest that this training time is largely independent of the size of the data set and the map.

3.4 Flexibility testing

3.4.1 Grid Topology

In this experiment, a 5 × 5 grid data set was used. This data set was then learnt by a SOM of size 9 × 9. Four schemes of parameter decay were used:

1. \( \alpha(t) = 1.0 - 0.9/300t, \sigma(t) = \text{round}(5.0 - 5.0/300t), \)

2. \( \alpha(t) = 1.0 - 0.9/100t, \sigma(t) = \text{round}(5.0 - 5.0/100t), \)

3. \( \alpha(t) = 0.3, \sigma(t) = \text{round}(5.0 - 5.0/300t), \)

4. \( \alpha(t) = 0.3, \sigma(t) = 5, t < 10, \text{otherwise } \sigma(t) = 2. \)

Schemes 1 and 2 are classic Kohonen linear decay schemes, over a period of 300 and 100 epochs respectively. Scheme 3 uses a classic linear decay for \( \sigma \), but keeps \( \alpha \) constant throughout training. Scheme 4 is the proposed step decay scheme. Following the results of the previous section, a large region size training time \((N)\) of 10 epochs is used. A goodness of fit of \( \approx 0.03 \) evenly spaces the grid data over the SOM. As can be seen in Figure 4, all four schemes attain this goodness of fit; Schemes 1 and 3 after \( \approx 150 \) epochs, Scheme 2 after \( \approx 75 \) epochs, and Scheme 4, the step decay, after \( \approx 20 \) epochs.

This is not to say that a linear decay scheme is incapable of matching the performance of the step algorithm. However, linear (and many other decay schemes) require \textit{a priori} knowledge of the time required for decay, or an appropriate rate of decay. This knowledge is not available for most problems without empirical study. The step algorithm requires no such \textit{a priori} knowledge; only a knowledge of the size of the map, and the probable density of data points on that map. Given this knowledge, which is available prior to training, without empirical testing, it is possible to use the step decay technique to attain near time-optimal SOM training.

Of significant note is the fact that Schemes 1 and 3 have almost identical convergence times. In addition, Scheme 3 is observed to demonstrate much less erratic behaviour during convergence. This suggests that the role played by decaying learning gain during training is much less significant than the literature suggests; the erratic convergence of Scheme 1 could even lead to the conclusion that decaying \( \alpha \) impedes learning.

The same procedure was then repeated for the 5 × 5 grid data set on a 17 × 17 SOM. The decay schemes used in this test were almost identical to those in the previous test, except that \( \sigma(t) \) was decayed from a value of 9.0, and was decayed to a value of 3 in scheme 4.

These schemes are the analogous to the previous experiments. Similar behaviour is observed in the large map case as in the small map case; whilst all four algorithms eventually attain the desired topology, the step region size method permits a more rapid convergence onto the optimal topology, without the need for \textit{a priori} knowledge. Again, the difference in convergence times between Scheme 1, with decaying \( \alpha \), and Scheme 3, with constant \( \alpha \), is negligible.

It is worthy of note that of the three training regimes, the
Figure 4: SOM Learning of grid data set, on a $9 \times 9$ map. Plot shows goodness of fit during training for two linear region decay mechanisms, and the step region decay mechanism. A goodness level of $\approx 0.02$ evenly spaces the grid data over the SOM.

step method is the only regime to not converge to a goodness measure of 0. This is a result of the linear decay algorithms training with region sizes of 1 and 0 at the end of their cycle. At this point, the narrow neighbourhood function strongly reinforces the winning nodes, excluding all other map neurons. If a significantly reduced goodness measure was a desirable outcome for the user, a third step of winner-take-all training could be applied once the user had ascertained convergence at the second step. However, since the goodness of fit level of 0.02 which is observed in this case is sufficient to provide a stable topology, this step is not required.

3.4.2 Minimal Spanning Tree

The minimal spanning tree data set, drawn from [6], consists of 32 training vectors, each consisting of 5 real values. Figure 6 shows the results of training, using the step decay method. The SOM consists of a $15 \times 15$ grid of neurons; an a region size of 8, decaying to 2 at epoch 20 is used. This result is vastly superior to the results reported by [6] - in his experiment, 10000 epochs are required to converge a smaller ($7 \times 10$), hexagonal map. This highlights the advantage of the stepped learning technique; it is quite possible that conventional decay techniques could be optimised to learn the tree data in a comparable time, but empirical investigation would be required to perform this optimisation. The stepped

Figure 5: SOM Learning of grid data set, on a $17 \times 17$ map. Plot shows goodness of fit during training for two linear region decay mechanisms, and the step region decay mechanism. A goodness level of $\approx 0.02$ evenly spaces the grid data over the SOM.

Figure 6: Training a SOM on a minimal spanning tree. (a)-(f): progression of topology during training. (g) Goodness of fit during training.
3.5 Discussion

These experiments clearly demonstrate the principle benefit of the stepped region size; a near optimal training time can be achieved, using parameters that can be easily derived from the data set and map to be used. No a priori knowledge or empirical studies of likely training time are required to establish the key parameters of the step decay scheme.

Perhaps the strongest evidence supporting use of the step decay mechanism comes through observing the goodness of fit curves for the 300 epoch linear decay in Figures 4 and 5. This curve demonstrates two behaviours; erratic oscillation (during epochs 1-150), and stable decay (during epochs 150-300). The switch between states occurs at the point where region size decays to the value of $R$; once the region size reaches this point, a stable topographical representation is achieved rapidly. Rather than engage in a long period of erratic representations, the step decay mechanism uses a large region size for the minimum possible time - just long enough to establish a primitive global topology. Region size is then dropped to $R$, to allow precise local ordering.

It is also interesting to note that this near optimal training time was achieved without the need for a sophisticated scheme of learning gain decay. This seems to contradict the vast majority of literature (for example, [7, 2, 9]), which emphasizes the importance of learning gain, leaving neighbourhood size as a peripheral concern. Sophisticated schemes of learning rate decay would therefore seem to be unnecessary.

4 Conclusion

In recent times, analysis of SOM performance has concentrated on optimising gain decay, rather than the size, form, and decay of the neighbourhood function. In this paper, the importance of neighbourhood size to the learning and development process was established. Using empirical studies, the role played by large and small region sizes was established. It was found that a combination of large and small region sizes is required during the training process to establish a good global topology with local precision.

Based upon these findings, a scheme for decaying region size during SOM training was proposed. This "step" region size decay scheme provides near optimal training time for arbitrary training sets. This decay scheme precludes the need for a priori knowledge of likely training times. In addition, the use of this scheme removes the need for sophisticated learning gain decay schemes, demonstrating that learning gain is not the most significant attribute contributing to the learning of SOM topologies.

An additional benefit of the proposed region decay scheme is its potential for use in continuous learning situations; a feature not available to conventional parameter decay schemes. This feature will be explored in future work.

References