Population Based Convergence Criterion for Self-Organizing Maps

Benjamin Ott, Gregory Breard, and Lutz Hamel, Department of Computer Science and Statistics, University of Rhode Island

What is a Self-Organizing Map?

Self-organizing maps (SOMs) are a common type of artificial neural network in which training data is run iteratively through a training algorithm. SOMs are used extensively as a method of analysis in a broad variety of fields including bioinformatics, financial analysis, signal processing, and experimental physics as they provide a simple yet effective algorithm for clustering via unsupervised learning [5]. The simple nature of the SOM algorithm and the way in which the visualization of the SOM can be easily and intuitively interpreted make it appealing as an analysis tool. However, with any analysis tool, and especially iterative learning-based tools, questions pertaining to the reliability of the interpretation of the results and the convergence of the algorithm naturally emerge.

The SOM Algorithm

```
Repeat until Done
  For each observation in the training data Do
    Find the neuron that best describes the
    observation.
    Make that neuron look more like the
    observation.
    Smooth the immediate neighborhood of that
    neuron.
  End For
End Repeat
```

Quantization Error

The quantization error is the distance (error) from a training observation to its best matching unit (neuron) in the SOM. It was proposed by Teuvo Kohonen as a measure for the quality of a SOM, with lower quantization errors implying better maps. However, the quantization error can be driven to zero by constructing a larger map, by training the map longer, and/or by increasing the learning rate of the SOM (i.e. by increasing the complexity of the model). Hence, quantization error is not an objective measure for determining when a SOM has been sufficiently trained since there is no measure for determining when a given quantization error is 'good enough'.

Population Based Convergence Criterion

The population based convergence criterion statistically compares the input sample space (training data) to the output space (the neurons in the SOM) using a simple two sample test. The two sample test can be computed quickly and does not lead to over-fitting. What we have observed is that the population based convergence criterion tracks the quantization error well in that as it increases, the quantization error decreases. Furthermore, full convergence can be achieved when the quantization error is not zero (i.e. this approach avoids over-fitting and provides an indicator of when a SOM is 'good enough'. The advantage of the population based convergence criterion over Cottrell's stability and reliability measure [2] is that it can be computed very quickly on a single map while the stability and reliability measures take longer to compute and need to be computed over at least 200 maps created using bootstrapped samples of training data. The advantage over Bishop's generative topographic mapping (GTM) [1], Verbeek's generative self-organizing map (GSOM) [6], and approaches minimizing energy function which have been imposed on the SOM [4,3] is that the population based convergence criterion does not modify the SOM algorithm and is much simpler, seemingly in line with the simple SOM algorithm.

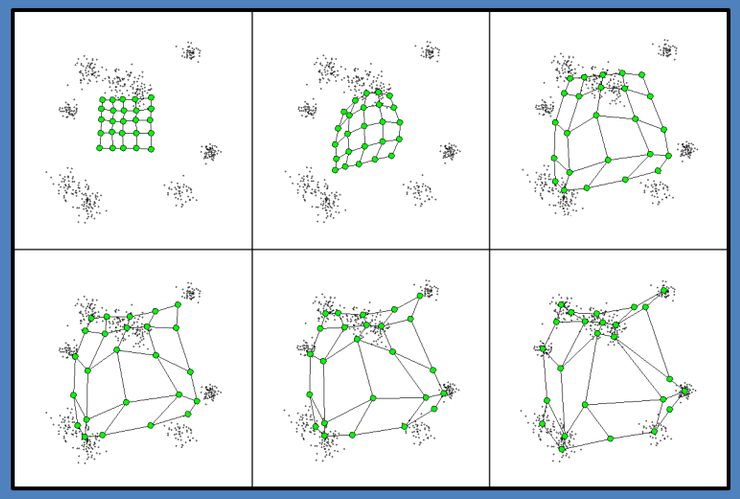

Figure 1: How a grid of neurons samples a training data space.

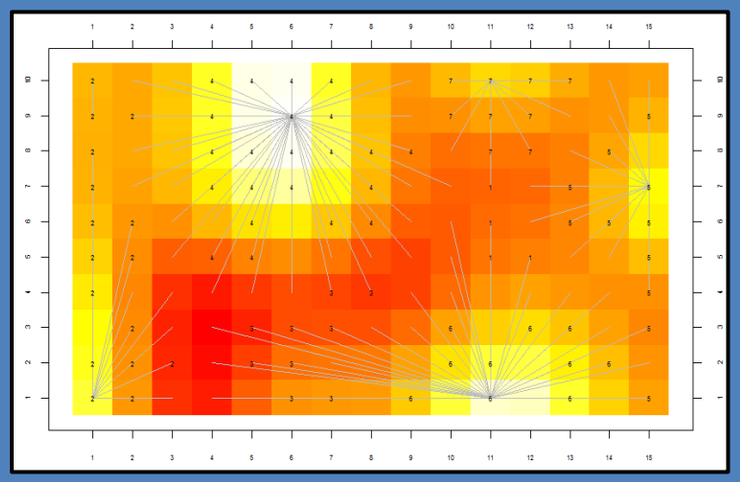

Figure 2: Starburst of map trained using the Hepta data set for 500 iterations, achieving a convergence score of 67.5%

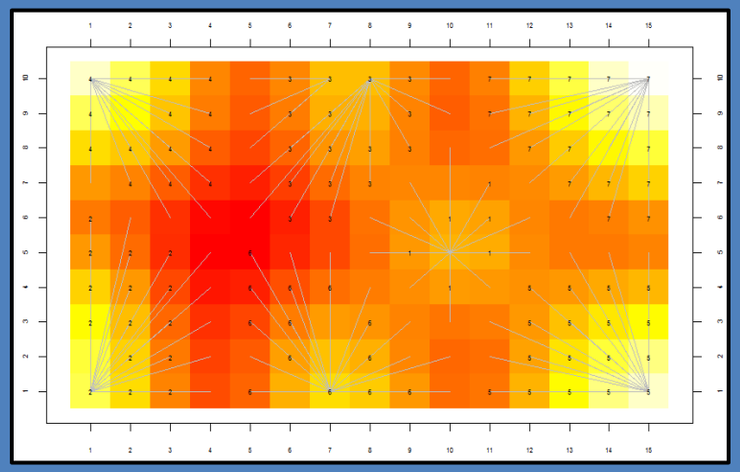

Figure 3: Starburst of map trained using the Hepta data set for 5000 iterations, achieving a convergence score of 100%

popsom: An R Package for SOMs

Our `popsom` package contains a set of routines which are useful in constructing and evaluating SOMs. The utilities are built around the `som` library available from the CRAN archive and the main utilities available in the package are the following: `map.build` constructs a map from a data set, `map.convergence` reports the accuracy of the map in terms of modeling the underlying data distribution, `map.significance` graphically reports the significance of each feature with respect to the SOM model, `map.umat` displays the unified distance matrix (U-matrix) of the SOM model, `map.starburst` displays the starburst representation of the SOM model, `map.projection` prints a table with the associations of labels with map elements, and `map.feature` computes and displays the enhanced U-matrix for a feature of the training data.

A Brief Tutorial of popsom

For this demonstration, we used the **Hepta** data set from the Fundamental Clustering Problems Suite [7]. The first step is to load the data frame and labels for our training data:

```
R> df <- read.table('hepta.lrn')
R> labels <- read.table('hepta.cls')
```

Using these it is simple to build a map with the following function:

```
R> m <- map.build(df, labels)
```

Also, the dimensions of the map, the learning rate, and the number of iterations can also be specified with `map.build`. Once the map is generated, it can be analyzed and displayed with the other methods included in the package. For example, use the following to calculate the convergence index of the map (that is, the variance captured by the map so far):

```
R> map.convergence(m)
```

Or, to compute and plot the relative significance of each feature:

```
R> map.significance(m)
```

It is, however, the visual representations of the SOMs that are the most spectacular feature of our package. In addition to the standard U-matrix, an enhanced U-matrix, or starburst, can also be generated as follows:

```
R> map.starburst(m)
```

A Population Based Convergence Criterion for Self-Organizing Maps, Lutz Hamel and Benjamin Ott. Proceedings of the 8th International Conference on Data Mining (DMIN'12).

[1] C. Bishop, M. Svensen, and C. Williams. Gtm: A principled alternative to the self-organizing map. *Artificial Neural Networks-ICANN 96*, pages 165–170.

[2] M. Cottrell, E. De Bodt, and M. Verleysen. A statistical tool to assess the reliability of self-organizing maps. *Advances in self-organising maps*, pages 7–14, 2001.

[3] E. Erwin, K. Obermayer, and K. Schulten. Selforganizing maps: ordering, convergence properties and energy functions. *Biological cybernetics*, 67(1):47–55,1992.

[4] T. Heskes. Energy functions for self-organizing maps. Kohonen maps, pages 303–316.

[5] T. Kohonen. *Self-organizing maps*. Springer series in information sciences. Springer, 2001.

[6] J. J. Verbeek and N. Vlassis. The generative selforganizing map.

[7] **Ultsch, A.**: Clustering with SOM: U*C, In *Proc. Workshop on Self-Organizing Maps, Paris, France*, (2005), pp. 75-82