WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

A Statistical Tool to Assess the Reliability of Self-Organizing Maps

M. Cottrell[1], E. de Bodt[2], M. Verleysen[3]

[1] Université Paris I, SAMOS-MATISSE, UMR CNRS 8595
90 rue de Tolbiac, F-75634 Paris Cedex 13, France
[2] Université catholique de Louvain, IAG-FIN, 1 pl. des Doyens,
B-1348 Louvain-la-Neuve, Belgium and
Université Lille 2, ESA, Place Deliot, BP 381, F-59020 Lille, France
[3] Université catholique de Louvain, DICE, 3, place du Levant,
B-1348 Louvain-la-Neuve, Belgium

Abstract

Making results reliable is one of the major concerns in artificial neural networks research. It is often argued that Self-Organizing Maps are less sensitive than other neural paradigms to problems related to convergence, local minima, etc. This paper introduces objective statistical measures that can be used to assess the stability of the results of SOM, both on the distortion and on the topology preservation points of views.

1 Introduction

Neural networks are powerful data analysis tools. Part of their interesting properties comes from their inherent non-linearities, in contrast to classical, linear tools. Nevertheless, the non-linear character of the methods has also its drawbacks: most neural network algorithms rely on the non-linear optimization of a criterion, leading to well-known problems or limitations concerning local minima, speed of convergence, etc.

It is commonly argued that vector quantization methods, and in particular self-organizing maps, are less sensitive to these limitations than other classical neural networks, like multi-layer perceptrons and radial-basis function networks. For this reason, self-organizing maps (SOM) [1] are often used in real applications, but rarely studied on the point of view of their reliability: one usually admits that, with some "proper" choice of convergence parameters (adaptation step and neighborhood), the SOM algorithm converges to an "adequate", or "useful", state.

WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

This paper aims at defining objective criteria that may be used to measure the "reliability" of a SOM in a particular situation. The bootstrap methodology is used to measure the variability of both the quantization error and the organization of the map. Section 2 summarizes the main idea of the bootstrap, section 3 defines a measure of the variability of the quantization error, section 4 a measure of the variability of the organization of the map, and section 5 applies the concepts to a few simple distributions.

2 Bootstrap

The main idea of the bootstrap [2] is to use the so-called "plug-in principle". Let F be a probability distribution depending on an unknown parameter vector $\boldsymbol{\theta}$. Let $\boldsymbol{x}=x_1, x_2,..., x_n$ be the observed sample of data and $\hat{\boldsymbol{\theta}} = T(\boldsymbol{x})$ an estimate of $\boldsymbol{\theta}$. The bootstrap consists in using artificial samples (called *bootstrapped samples*) with the same empirical distribution as the initial data set in order to guess the distribution of $\hat{\boldsymbol{\theta}}$. Each *bootstrapped sample* consists in n uniform drawings with repetitions from the initial sample. If $\boldsymbol{x}^*$ is a bootstrapped sample, $T(\boldsymbol{x}^*)$ will be a bootstrap replicate of $\hat{\boldsymbol{\theta}}$.

This main idea of the bootstrap may be declined in several ways. In particular, when the evaluation of $T(\boldsymbol{x})$ requires non-linear optimization, the well-known problems, or limitations, related to local minima and convergence are encountered. It may thus happen that different local minima are reached when $T(\boldsymbol{x}^*)$ is evaluated for different bootstrapped samples. This is clearly not what we are looking for: our purpose is to examine the variability (or the sampling distribution) of some parameters when they are evaluated through different (bootstrapped) samples, but keeping all other conditions unchanged. In order to overcome this problem, local bootstrap methods may be used, where the initial conditions for each evaluation of $\hat{\boldsymbol{\theta}}$ are kept fixed. In the following, we will speak about:

- Common Bootstrap (**CB**) when each evaluation of $\hat{\boldsymbol{\theta}}$ is initialized at random;
- Local Bootstrap (**LB**) when the initial values of each evaluation are kept fixed;
- Local Perturbed Bootstrap (**LPB**) when a small perturbation is applied to the initial conditions obtained as with the Local Bootstrap.

If we want to evaluate the influence of the convergence (only) during the evaluation of $\hat{\boldsymbol{\theta}}$, we will not bootstrap samples, but reiterate the evaluation of $T(\boldsymbol{x})$ with the same sample $\boldsymbol{x}$ and different initial conditions. In this case, we will speak about Monte-Carlo simulations instead of bootstrap, and we will use the same three variants as above: Common Monte-Carlo (**CMC**), Local Monte-Carlo (**LMC**) and Local Perturbed Monte-Carlo (**LPMC**).

WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

3 Stability of the Quantization in the SOM

One of the two main goals of Self-Organizing Maps is to quantize the data space into a finite number of so-called *centroids* (or *code vectors*). Vector quantization is used in many areas to compress data over transmission links, to remove noise, etc. The distance between an observed data x_i and its corresponding (nearest) centroid is te *quantization error*. Averaging this quantization error over all data leads to the *distortion* or *intra-class sum of squares* (which are different names for the same error, used respectively in the information theory domain and by statisticians):

$$SSIntra = \sum_{i=1}^{U} \sum_{x_j \in V_i} d^2(x_j, G_i) \quad (1)$$

where U is the number of units in the SOM, G_i is the i-th centroid, d is the classical Euclidean distance, and V_i is the Voronoi region associated to G_i, i.e. the region of the space nearer to G_i than to any other centroid. Note that the objective function associated with the SOM algorithm for a constant size of neighborhood and finite data set is *the sum of squares intra-classes extended to the neighbor classes,* (see [3]). But actually, one usually ends with no neighbor for the last iterations of the SOM algorithm; at the end of its convergence, the SOM algorithm thus exactly minimizes the *SSIntra* function.

The Monte-Carlo and/or bootstrap methods will allow us to estimate the variability of *SSIntra*, in other words to assess if one may be confident in the stability of the quantization obtained by the SOM. Note that we do not speak about the value (location) of the centroids themselves, but on how they quantify the space in average. If the SOM is computed several times according to the Monte-Carlo or the Boostrap principle detailed in the previous section, one can calculate the mean $\mu_{SSIntra}$ and the standard deviation $\sigma_{SSIntra}$ of the distortion. The variability of *SSIntra* is the evaluated by its coefficient of variation *CV* defined as follows:

$$CV(\theta) = 100 \frac{\sigma_\theta}{\mu_\theta} \quad (2)$$

where θ is the parameter to examine, here *SSIntra*.

4 Stability of the Neighborhood Relations in the SOM

The second main goal of the SOM is the so-called *topology preservation*, which means that close data in the input space will be quantized by either the same centroid, either two centroids that are close one from another on a predefined string or grid. Often, for example when the SOM is used as a visualization tool, it is desirable to have an objective measure of this *neighborhood* property. We then define for any pair of data x_i and x_j,

WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

$$STAB_{i,j}(r) = \frac{\sum_{b=1}^{B} NEIGH_{i,j}^{b}(r)}{B} \quad (3)$$

where $NEIGH^{b}_{i,j}(r)$ is an indicator function that returns 1 if the observations x_i and x_j are neighbor at the radius r for the bootstrap sample b, and B is the total number of bootstrapped samples. If the radius r is 0, it means that we evaluate if the two data are projected on the same centroid; if $r = 1$, it means that we evaluate if the two data are projected on the same centroid *or* on the immediate neighboring centroids on the string or grid (2 on the string, 8 on the grid), etc.

A perfect stability would lead $STAB_{i,j}$ to always be 0 (never neighbor) or 1 (always neighbor). We can study the significance of the statistics $STAB_{ij}(r)$, by comparing it to the value it would have if the observations fell in the same class (or in two classes distant of less than r) in a completely random way. Let U be the total number of classes and v the size of the considered neighborhood. The size v of the neighborhood can be computed from the radius r by $v = (2r + 1)$ for a one-dimensional SOM map (a string); and $v = (2r + 1)^2$ for a two-dimensional SOM map (a grid), if edge effects are not taken into account. For a fixed pair of observations x_i and x_j, with random drawings, the probability of neighboring would be v/U. If we define a Bernoulli random variable with probability of success v/U, (where success means: "x_i and x_j are neighbor"), the number Y of successes on B trials is distributed as a Binomial distribution, with parameters B and v/U. Therefore, it is possible to build a test of the hypothesis H_0 "x_i and x_j are only randomly neighbors" against the hypothesis H_1 "the fact that whether x_i and x_j are neighbors or not is meaningful".

If B is large enough (i.e. greater than 50), the binomial random variable can be approximated by a Gaussian variable, making the hypothesis test easier. For example, with a test level of 5%, we conclude to H_1 if Y is less than $B\frac{v}{U} - 1.96\sqrt{B\frac{v}{U}\left(1-\frac{v}{U}\right)}$ or greater than $B\frac{v}{U} + 1.96\sqrt{B\frac{v}{U}\left(1-\frac{v}{U}\right)}$. Note that in the case of the bootstrap, B depends on the pair (x_i, x_j), since the bootstrapped samples have to contain both data: we follow the same approach as in [4] which consists in evaluating $STAB_{i,j}(r)$ only on the samples that contain observations x_i and x_j.

5 Experiments

The above described indicators have been evaluated on artificial and real databases; a selection of results follow.

WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

5.1 Databases

Three artificial databases have been used: Gauss_1, Gauss_2 and Gauss_3. All three are two-dimensional data sets, obtained by random drawings on uncorrelated Gaussian distributions. They are respectively represented in figures 1, 2, and 3. Gauss_1 contains only one cluster of observations. Gauss_2 contains three clusters of equal variance and some overlap. Gauss_3 is also composed of three clusters, but of different variances and without overlap. Each data set has 500 observations. For data sets Gauss_2 and Gauss_ 3, observations 1-166, 167-333 and 334-500 are in the same cluster.

A real database, POP_84, was also used. It contains six ratios measured in 1984 on the macroeconomic situation of countries: annual population growth, mortality rate, analphabetism rate, population proportion in high school, GDP per head and GDP growth rate. This dataset has been already used in [5], and is available through [6].

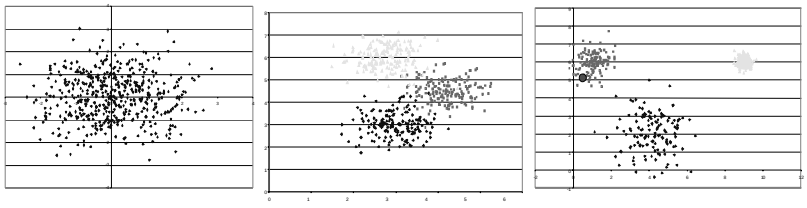

Figure 1: Gauss_1, Gauss_2 and Gauss_3 databases

5.2 Stability of the Distortion Error

Table 1 summarizes the results on the coefficient of variation (2) of the distortion (1), measured on the three artificial databases, and obtained by the CMC, LMC and LPMC methods on 5000 independent samples (note that such a large number of samples is not necessary in practice to obtain reliable results; 100 samples is already a good choice). The Kohonen map used in these simulations is a 3- or 6-units 1-dimensional string.

method	CMC		LMC		LPMC	
# units	3	6	3	6	3	6
Gauss_1	5.2	4.5	5.3	4.4	5.2	4.5
Gauss_2	5.1	4.6	4.9	4.5	5.1	4.6
Gauss_3	7.6	10.1	6.4	10.3	6.7	10.1

Table 1: Coefficients of variation of *SSIntra*, obtained with 1-dimensional 3- and 6-units SOM, and with CMC, LMC and LPMC Monte-Carlo simulations.

Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

An expected result is that the CV is low in each case, and does not seem to be influenced by the initialization method used (CMC, LMC or LPMC). This enforces the idea that the SOM is a reliable method, not falling too much into the traps of local minima encountered with other neural network models (see for example [7] for opposite results obtained with MLP).

A more surprising result at the first sight is the increase of the CV when switching from a 3- to a 6-units map on the Gauss_3 distribution. This can however be explained by the fact that the Gauss_3 distribution contains 3 well separated clusters. When using a 3-units map on this distribution, one centroid will obviously converge around the center of each cluster. Adding units will however lead to situation where the supplementary ones may be "captured" by one cluster or by another, increasing the variability of the situation after convergence. This is of course not (or less) true with the Gauss-1 and Gauss_2 distributions where there is overlap between clusters.

5.3 Assessing the Right Number of Units in the Map

This last comment makes think that it would be possible to estimate the number of units to include in a SOM on a particular database, by examining the evolution of the CV when increasing the number of units. This is an important question when using SOM, and may be considered as a by-product of the proposed measures. Table 2 shows the coefficient of variation of *SSIntra* obtained after Local Boostrap, on the three Gauss_x databases. Table 2 confirms the step between 3 and 4 units on the Gauss_3 databases, while the results on the other databases are less conclusive, as expected.

# units	Gauss_1	Gauss_2	Gauss_3	# units	Gauss_1	Gauss_2	Gauss_3
1	5.2	4.3	5.5	6	5.1	4.7	12.0
2	4.5	6.0	8.9	9	5.4	4.7	10.9
3	5.9	5.4	6.5	12	3.7	4.9	9.2
4	5.5	4.9	14.4	15	4.0	4.0	8.0
5	4.4	6.6	15.2				

Table 2: Coefficients of variations of *SSIntra* obtained after Local Bootstrap.

5.4 Stability of the Neighborhood Relations

Table 3 shows a few results obtained on the POP_84 data set, on selected pairs of countries. The values in the table are those of the *STAB* indicator defined in equation (3). The two first columns refer to a 1-dimensional SOM with 6 units, and show respectively $STAB_{i,j}(0)$ and $STAB_{i,j}(1)$, where i,j identifies the countries, while

the third column refers to $STAB_{i,j}(2)$ on a 7x7 2-dimensional SOM. The results marked with * are those for which the hypothesis test (with 1% significance) defined in section 4 lead to conclude hypothesis H1, i.e. the neighborhood relation (presence or absence) between two countries is meaningful. As expected, countries with similar situations give high values of the *STAB* indicator. Countries with very opposite situations give very low values.

Countries	$STAB_{i,j}(0)$ 6 units	$STAB_{i,j}(1)$ 6 units	Countries	$STAB_{i,j}(2)$ 7x7 units
Turkey/Upper Volta	0.04*	0.65*	Greece/France	0.18*
Turkey/Cuba	0*	0.22*	Australia-France	0.82*
Turkey/Sweden	0*	0.05*	Greece/Belgium	0.21*
Turkey/France	0*	0*	Turkey/France	0.02*
Turkey/Greece	0*	0.25*	Singapore/USA	0.49
Upper Volta/Cuba	0*	0*	Sweden/Japan	0.73*
Upper Volta / Sweden	0*	0*	Greece/Malta	1*
Upper Volta / France	0*	0*	Canada/France	0.84*
Sweden/France	1*	1*	Sweden/France	0.97*
Cuba / Sweden	0.02*	0.81*	USA/Zimbabwe	0*
Cuba / France	0.02*	0.78*	USA/Finland	0.85*
Cuba / Greece	0.69*	0.97*	USA/Australia	0.68*

Table 3: *STAB* indicator on the POP_84 dataset, for selected pairs of countries..

6 Discussion

This paper introduces measures to assess the stability of the SOM convergence, under the distortion and topology preservation point of views. Having objective measures of the stability may enforce the idea that SOM are less sensitive to problems related to convergence and local minima than other neural network paradigms. A by-product to the proposed measure is a way to estimate the number of units in a Kohonen map.

The measure of the stability of the neighborhood relations concerns a particular pair of data. It would be interesting to obtain a global measure of the topology preservation over the whole map. As "reliable" neighborhood relations mean values of the *STAB* indicator near to 0 or 1, a possibility is to draw an histogram of the *STAB* values, and to measure how this histogram is close or not to a "U"-shape (peaks at 0 and 1, and smooth curve with low values in between).

WSOM 2001 proceedings - Workshop on Self-Organizing Maps, Lincoln (United Kingdom), 13-15 June 2001
Advances in Self-Organising Maps, N. Allinson, H. Yin, L. Allinson, J. Slack eds., Springer Verlag, 2001, ISBN 1-85233-511-4, pp. 7-14

References

1. Kohonen T., Self-Organizing Maps, Springer, Berlin, 1995
2. Efron B. & Tibshirani R., Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy, Statistical Science, 1986, 1:54-77
3. Ritter H., Martinetz T. and Schulten K., Neural Computation and Self-Organizing Maps, Addison-Wesley, Reading, 1992
4. Efron B. & Tibshirani R., An Introduction to the Bootstrap, Chapman and Hall, 1993
5. Blayo F., Demartines P., Data Analysis: How to Compare Kohonen Neural Networks to Other Techniques? In A. Prieto ed., Artificial Neural Networks, Lecture Notes in Computer Science, 1991, 469-476
6. Available from http://panoramix.univ-paris1.fr/SAMOS/
7. Zapranis A. & Refenes A.P., Principles of Neural Model Identification, Selection and Adequacy, Springer, 1999