ESANN'2000 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2000, D-Facto public., ISBN 2-930307-00-5, pp. 245-254

Bootstrapping Self-Organising Maps to Assess the Statistical Significance of Local Proximity

Eric de Bodt[1], Marie Cottrell[2]

[1] 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

[2] Université Paris I, SAMOS-MATISSE, UMR CNRS 8595
90 rue de Tolbiac,
F-75634 Paris Cedex 13, France

Abstract. One of the attractive features of Self-Organising Maps (SOM) is the so-called "topological preservation property": observations that are close to each other in the input space (at least locally) remain close to each other in the SOM. In this work, we propose the use of a bootstrap scheme to construct a statistical significance test of the observed proximity among individuals in the SOM. While computer intensive at this stage, this work represents a first step in the exploration of the sampling distribution of proximities in the framework of the SOM algorithm.

1. Introduction

The SOM algorithm was introduced by Kohonen in 1981 and has been the focus of a sizeable amount of attention in the scientific community since then. Numerous applications have been proposed (see Kohonen [1995] for a representative list of them) and the theoretical properties have been carefully studied (see Cottrell, Fort & Pagès [1998] for a review of the established results up to now). Henceforth, we will consider here that the SOM algorithm is familiar to the reader.

One of the most attractive features of SOM, (in particular for applications in the field of data analysis), is the so-called "topological preservation property": after organisation through the training algorithm, observations that are close to each other in the input space (at least locally) belong to units that are neighbours (or are actually within the same unit). A question that has not received a lot of attention to date is the statistical significance of the observed neighbourhood in the SOM obtained after learning. Having observed that two individuals from the analysed sample belong to neighbour units, what is the probability that they are actual neighbours in the population? In other words, what is the sampling distribution of the observed proximity and is it possible to propose a statistical test to assess their significance?

To answer this question, we will first recall the central ideas of the bootstrap, as introduced by Efron [1979] and address/solve specific difficulties encountered when applying the bootstrap in the field of neural-networks (for references on bootstrap procedures and their applications, see e.g. Efron, Tibshirani [1986, 1993], Freedman [1981,1984], LePage, Billard [1992], Noreen [1989], …). We will clearly define the concept of proximity, propose a bootstrap procedure adapted to the SOM algorithm and introduce a Binomial test to assess the statistical significance of observed neighbourhoods. Before concluding, we will apply our propositions to three simulated data sets and to a real database.

2. A Bootstrap Scheme adapted to the SOM Algorithm

In real applications, the SOM algorithm is used on a finite data set, which can be seen as a sample from some unknown distribution. One important question that arises about the resulting map is: "Is it reliable?". We propose the use of the bootstrap approach to evaluate the reliability of the map on both the point of view of *quantification* (evaluated by the sum of squares intra-classes, cf. eq. 1) and the *neighbourhood significance* (evaluated by the stability of the observed proximity on the map).

The quality of the quantification is evaluated by the sum of all the distances between the observations and their winning code vector (the weight vector of the closest unit which is the representative vector of the class they belong to). This sum is called *distortion* in the quantification theory, and *sum of squares intra-classes* by the statisticians. It can be expressed by:

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

where U is the number of classes (or units), C_i is the i-th class, G_i is the code vector of class C_i, and d is the classical Euclidean distance in the data space.

Let us recall that the decreasing function associated with the SOM algorithm for a constant size of neighbourhood and finite data set is *the sum of squares intra-classes extended to the neighbour classes*. But actually, in the last part of the iterations no neighbour is considered. And at the end, the SOM algorithm is equivalent to Simple Competitive Learning and minimises exactly the *SSIntra* value.

The bootstrapped samples will help us to study the stability of the distortion by estimating it and its standard deviation regardless of the learning (which depends on the initialisation, order of data presentation, decrease of the neighbourhood size, and the adaptation parameter, etc.).

In regards to the stability of the neighbourhood relation, it is simply evaluated by the number of cases where, during the bootstrap process, two observations are neighbours or not neighbours. The stability of neighbourhood therefore has to be evaluated for a couple of observations and, classically, we have to define the radius of neighbourhood at which the proximity is taken into account (see equation 2). For any pair of data x_i and x_j,

$$STAB_{i,j}(r) = \frac{\sum_{b=1}^{B} NEIGH_{i,j}^{b}(r)}{B} \qquad \text{eq.2}$$

where $NEIGH^{b}_{i,j}(r)$ is an indicator function that returns 1 if the observations x_i and x_j are neighbours at the radius, *r,* for the bootstrap sample, *b,* and *B* is the total number of bootstrapped samples. A perfect stability would lead $STAB_{i,j}$ to always be 0 (never neighboured) or 1 (always neighboured).

The application of the bootstrap procedure to the SOM algorithm raises two specific problems:

- for MLP, the minimised function has a sizeable amount of local minima. Part of the variability of the estimated statistics (*SSIntra,* $STAB_{i,j}$) can be due to this convergence problem. As in Zapranis and Refenes [1999] (cf. supra), we will analyse the impact of the "convergence difficulty" on the stability of the estimations (see section 4 of the paper).
- to evaluate $NEIGH^{b}_{i,j}(r)$, it is necessary to say that x_i and x_j must be part of the bootstrap sample, *b*, which is in no way guaranteed. To solve this problem, we use the same approach as in Efron and Tibshirani [1993]: the $STAB_{i,\,j}(r)$ is evaluated only on the parts of the bootstrap samples that contain the observations x_i and x_j.

The proposed bootstrap procedure is resumed in figure 1. The terminology we will use to present our results is the following:

- if no re-sampling is done (in order to analyse the variability of the results due only to convergence problems), we will talk of Monte-Carlo (**MC**) simulation,
- if re-sampling is done, we will talk of Bootstrap (**B**) simulation,
- if, for each bootstrap iteration, the SOM Map is initialised at random (in the input data space), we will talk of Common Monte Carlo (**CMC**) or Common Bootstrap (**CB**) (depending on the activation of re-sampling or not),
- if, for each bootstrap iteration, the SOM Map is initialised with the weight vectors obtained after the convergence of the initial learning, we will talk of Local Monte Carlo (**LMC**) or Local Bootstrap (**LB**),
- if we do the same computations as in the previous point, but we add a small random perturbation to the weight vectors, we will talk of Local Perturbed Monte Carlo (**LPMC**) or Local Perturbed Bootstrap (**LPB**).

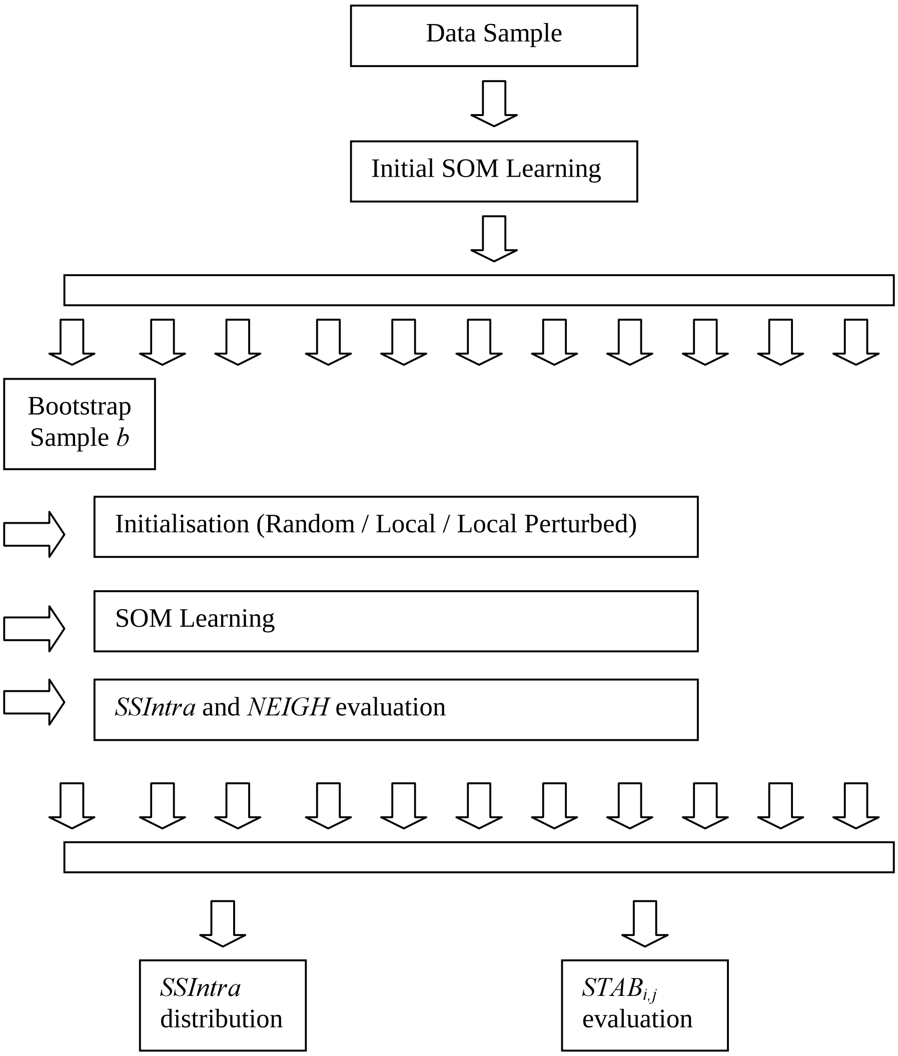

Figure 1: Bootstrap procedure of the SOM algorithm

We can study the significance of the statistics $STAB_{ij}(r)$, by comparing it to the value it would have had if the observations fell in the same class (or in two classes distance of less than r) in a completely random way.

Let U be the total number of classes and v the size of the considered neighbourhood. The size, v, of the neighbourhood 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). For a fixed pair of observations, x_i and x_j , with random repetition, the probability of neighbouring 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 neighbours"), the number, Y, of successes on B trials is distributed as a Binomial distribution, with parameters B and v/U. So, it is possible to build a test of the hypothesis H_0 "x_i and x_j are only random neighbours" against the hypothesis H_1 "the fact that whether x_i and x_j are neighbours, 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 and, for example, with a test level of 5%, we conclude to H_1 if Y is less than $B\dfrac{v}{U}-1.96\sqrt{B\dfrac{v}{U}\left(1-\dfrac{v}{U}\right)}$, or greater than $B\dfrac{v}{U}+1.96\sqrt{B\dfrac{v}{U}\left(1-\dfrac{v}{U}\right)}$.

This gives a level of significance to the presence/absence of the neighbourhood relations.

3. Applications

3.1. Data set and SOM Map

The results that we present and analyse here have been obtained on three simulated data sets[1]; each one representing a specific situation. We will call them: Gaussian_1, Gaussian_2 and Gaussian_3. In each case, they are two-dimensional data sets, obtained by random drawing in an uncorrelated Gaussian distribution. They are represented respectively in figures 2, 3, and 4. The first data set shows a situation where there is only one cluster of observations. The second contains three clusters of equal variance and some overlap. The third is also composed of three clusters, but of different variance and no overlap. Each data set is composed of 500 individuals. And, for data sets Gaussian_2 and Gaussian_ 3; observations 1-166, 167-333 and 334-500 are in the same cluster.

[1] Complementary results have been obtained with several real data sets but the simulated ones allow us to clearly illustrate the results of particular interest.

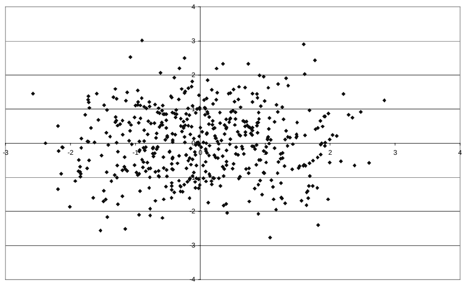

Figure 2: Gaussian_1 data set

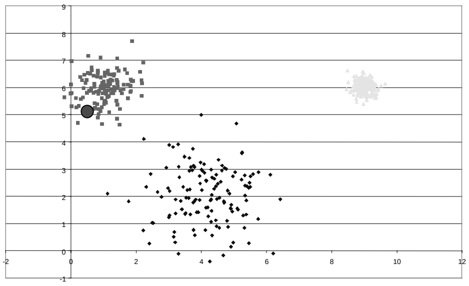

Figure 4: Gaussian_3 data set

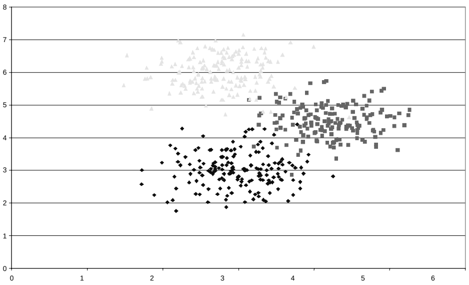

Figure 3: Gaussian_2 data set

For the sake of conciseness, the results presented here are limited to a one-dimensional SOM Map (or string), composed of either 3 or 6 units. Classically, the neighbourhood and the learning rate are decreasing during the learning.

3.2 Variability of *SSIntra* due to convergence of the algorithm

The first point we present, with attendant results, is the variability of *SSIntra* due to convergence of the SOM algorithm. The point here is to see if the existence of local minima can introduce variability in the estimation of *SSIntra*. Table 1 summarises the coefficients of variation, (CV)[2] , for the distribution of *SSIntra* obtained by CMC (no re-sampling and random initialisation at each iteration); Table 2, the CV obtained by LMC (no re-sampling, fixed initialisation at each iteration); and, Table 3, the CV obtained by LPMC (no re-sampling, small random perturbation of the fixed initialisation). Each result presented here has been established with 5000 independent samples[3]. For the sake of conciseness, the results for $STAB_{i,j}$ are not presented here.

[2] The coefficient of variation CV is equal to 100 σ/μ, where σ is the standard deviation, μ, is the mean value.

[3] Such a large number of samples, in practice, is not necessary (100 being enough); but, we wish to be certain of the numerical stability of the results.

The comparison shows quite clearly that the stability of the *SSIntra* estimation does not rely on the mode of initialisation of the bootstrap procedure. By switching from CMC to LMC (or PLMC), i.e. by fixing the initialisation of the weight vectors, the obtained coefficients of variation are almost the same. This result is very different from those obtained by Zapranis and Refenes [1999] when applying bootstrap to MLP and emphasise the great robustness of the SOM algorithm. The most interesting result that appears in tables 1 through 3 is the important impact of the number of units on the CV in Gaussian_3 cases. As can be seen in figures 2 and 3, Gaussian_3 is the only case with well-separated asymmetric clusters. It is clear that the "natural" number of units should be 3 and, in some sense, a SOM Map with 6 units is over parameterised. The stability of *SSIntra* seems, at first sight, to indicate the wrong choice of number of units. It is this point in particular that we will explore in the next section of this paper.

	3 units	6 units
Gaussian_1	0.052	0.045
Gaussian_2	0.051	0.046
Gaussian_3	0.076	0.101

Table 1: Coefficients of variation of *SSIntra* for Common Monte-Carlo (CMC)

	3 units	6 units
Gaussian_1	0.052	0.045
Gaussian_2	0.051	0.046
Gaussian_3	0.067	0.101

Table 3: Coefficients of variation of *SSIntra* for Local Perturbed Monte-Carlo (LPMC)

	3 units	6 units
Gaussian_1	0.053	0.044
Gaussian_2	0.049	0.045
Gaussian_3	0.064	0.103

Table 2: Coefficients of variation of *SSIntra* for Local Monte-Carlo (LMC)

3.3 Assessing the right number of units in a SOM Map

Table 4 shows the CV's of *SSIntra* obtained from the three simulated data sets presented in section 3.1; as well as on a real data set called POP [4]. The results have been obtained using 100 bootstrap samples. They confirm those highlighted in the previous section. For Gaussian_1, where there is only one natural cluster, the CV of *SSIntra* exhibits oscillations around 0.45. For Gaussian_3, as expected, the addition of a fourth unit generates a large increase in the CV. As shown in table 4, for the POP data set, the increase of the CV of *SSIntra* is situated near the addition of the seventh

[4] This actual data (extracted from official public statistics for 1984) was used in Blayo, F. & Demartines, P. (1991): *Data Analysis : How to Compare Kohonen Neural Networks to Other Techniques ?* in *Proceedings of IWANN'91*, Ed. A.Prieto, Lecture Notes in Computer Science, Springer-Verlag, 469-476. It contains 6 variables (annual population growth, mortality rate, analphabetism rate, population proportion in high school, GDP per head, GDP growth rate) for 53 countries.

or eighth unit. The result seems to be surprising for the Gaussian_2 data set; where there is no increase of the CV of *SSIntra* when adding a fourth unit. The explanation lies in the strictly symmetrical form of the three clusters and in their overlapping positions (the instability of the location of the fourth unit does not change the level of *SSIntra* obtained from one bootstrap sample to another bootstrap sample).

Number of units	Gaussian_1	Gaussian_2	Gaussian_3	POP
1	0.052	0.043	0.055	0.046
2	0.045	0.060	0.089	0,079
3	0.059	0.054	0.065	0.073
4	0.055	0.049	0.144	0.068
5	0.044	0.066	0.152	0.085
6	0.051	0.047	0.120	0.088
9	0.054	0.047	0.109	0.147
12	0.037	0.049	0.092	0.180
15	0.040	0.040	0.080	0.187

Table 4 : Coefficients of variation of *SSIntra* obtained after Local Bootstrap

Pair of obs.	Gauss_2 3 units	Pair of obs.	Gauss_3 3 units	Pair of countries	POP (*r*=0) 6 units	POP (*r*=1) 6 units
137/43 Cl1/Cl1	1	137/43 Cl1/Cl1	0	49/21 Turkey/Upper Volta	0.04**	0.65**
137/255 Cl1/Cl2	0	137/255 Cl1/Cl2	1	49/13 Turkey/Cuba	0***	0.22***
137/437 Cl1/Cl3	0	137/437 Cl1/Cl3	0	49/47 Turkey/Sweden	0***	0.05***
137/70 Cl1/Cl1	1	137/70 Cl1/Cl1	0	49/19 Turkey/France	0***	0***
137/278 Cl1/Cl2	0	137/278 Cl1/Cl2	0	49/20 Turkey/Greece	0***	0.25***
43/255 Cl1/Cl2	0	43/255 Cl1/Cl2	0	21/13 Upper Volta/Cuba	0***	0***
43/437 Cl1/Cl3	0	43/437 Cl1/Cl3	0	21/47 Upper Volta / Sweden	0***	0***
43/70 Cl1/Cl1	1	43/70 Cl1/Cl1	1	21/19 Upper Volta / France	0***	0***
43/378 Cl1/Cl1	0	43/378 Cl1/Cl1	0	47/19 Sweden/France	1***	1***
255/437 Cl2/Cl3	0	255/437 Cl2/Cl3	0	13/47 Cuba / Sweden	0.02**	0.81***
255/70 Cl2/Cl1	0	255/70 Cl2/Cl1	0	13/19 Cuba / France	0.02**	0.78***
255/378 Cl2/Cl3	0	255/378 Cl2/Cl3	0	13/20 Cuba / Greece	0.69***	0.97***

Table 5: Frequencies of neighbourhood obtained by Local Bootstrap

**significant at 5 % ***significant at 1%

3.4 A statistical test of the proximity relations among observations in the SOM Map

In this section, we present results concerning the stability of the neighbourhood relations that appears in the SOM maps. The first three "pair" columns concern the neighbourhood with radius r=0, (i.e. the observations are considered as neighbours only if they belong to the same class). The last column shows the results for the POP data set with a radius neighbourhood of 1 (i.e. the observations are neighbours if they belong to the same class or to two adjacent classes).

Table 5 shows the results concerning $STAB_{i,j}$. In the columns "Pair of obs", the cluster ownership are mentioned for the observations two in number and for data sets Gaus_2 and Gaus_3 (e.g. the first pair of observations of Gaus_2 data set is 137/43; Cl1/Cl1 means that observation 137 is a member of cluster 1 and observation 43 is a member of cluster 1). For the POP data set, we mention the country names. The number of units is mentioned in the title of the columns. All estimations have been computed with 100 bootstrap samples. The levels of significance have been calculated from a Binomial distribution with p=1/6 (cf section 2). The main results are as follows:

– For the Gaus_2 data set, we obtain strictly what was expected: if two observations are in the same cluster, the probability they belong to the same unit is 1 (and vice-versa). We have to remember that the SOM algorithm is a stochastic one…
– For the Gaus_3 data set, the conclusions are the same as those obtained for the Gaus_2 data set; except for observation 137, which is wrongly associated with observations of the second cluster. In figure 4, we mark this observation with a red point. As we can see, it is located in the second cluster (while issued from the first one). This corresponds with an error of classification since its location and the results obtained by bootstrap are fully coherent.
– For the POP data set, the observed similarities between the countries agree with the economic situation in the year 1984, as far as we know. It is necessary to study the map in a more detailed way to fully interpret the results, but it is out of the scope of this paper. However, it is evident that France is completely different from Upper Volta (presently Burkina-Faso), and that France and Sweden are very similar with respect to the considered variables (see appendix).

4. Conclusion

These are preliminary results, but are nonetheless very promising. We intend to pursue these tracks by:

- systematically studying how to determine the correct number of units using the coefficients of variation of the *SSIntra* for the bootstrapped samples, according to the number of units;

- analysing the stability of the neighbourhoods according to the number of units more deeply (as we saw, the stability disappears when the number of units is over-dimensioned);
- applying these methods to numerous real data and applying, in this context, well-known numerical optimisations to the Monte-Carlo procedure.

We think that this kind of work can supply the innumerable users of the SOM maps with a new tool that can make them increasingly confident in the power and effectiveness of the Kohonen algorithm.

References

[1] Cottrell M., Fort J.C. & Pagès, *Theoretical Aspects of the SOM Algorithm*, Neurocomputing, 21, 1998, p. 119-138.

[2] Efron B., *Bootstrap Methods : Another Look at the Jackknife, The 1977 Rietz Lecture*, The Annals of Statistics, vol. 7, n°1, 1979, p. 1-26

[3] Efron B. & Tibshirani R., *Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy*, Statistical Science, vol. 1, n°1, 1986, p. 54-77

[4] Efron B. & Tibshirani R., *An Introduction to the Bootstrap*, Chapman and Hall, 1993

[5] Freedman D.A., *Bootstrapping Regression Models*, Annals of Statistics, vol. 9, 1981, p. 1218-1228

[6] Freedman D.A., *On Bootstrapping Two-Stage Least-Squares Estimates in Stationary Linear Models*, vol. 12, n°3, 1984, p. 827-842

[7] Kohonen T., *Self-Organising Maps*, Springer, Berlin, 1995.

[8] LePage R. & Billard L., *Exploring the Limits of Bootstrap*, Wiley, 1992

[9] Noreen, E.W., *Computer Intensive Methods for Testing Hypotheses - An Introduction*, Wiley, 1989

[10] Zapranis A. & Refenes A.P., *Principles of Neural Model Identification, Selection and Adequacy*, Springer, 1999.