UNDERSTANDING AND REDUCING VARIABILITY OF SOM' NEIGHBOURHOOD STRUCTURE

Patrick Rousset[1], Christiane Guinot[2], Bertrand Maillet[3]
[1]CEREQ, 10 place de la Joliette
F-13567 Marseille, France
rousset@cereq.fr
[2]C.E.R.I.E.S, 20 rue Victor Noir
92521 Neuilly sur Seine, France
christiane.guinot@ceries-lab.com
[3]A.A.Advisors-QCG (ABN Amro Group), Variances and Paris-1 (TEAM/CNRS),
106 bv de l'hôpital F-75647 Paris cedex 13, France
bmaillet@univ-paris1.fr

Abstract *– One of the most interesting features of self-organizing maps is the neighbourhood structure between classes that is highlighted by this technique. The aim of this paper is the presentation of two complementary methods dealing with the variability of the induced neighboured structure. The first method connects this variability to the complexity of the data intrinsic structure. A visualizing tool, called Map of Distances between Classes (MDC), is presented; it basically allows to extract the main information from the very large matrix of distance between Self-Organizing Maps' classes. This matrix is a very accurate description of both the data structure and the SOM algorithm interpretation. In a presence of a complex structure, it enlarges the information set, linking the variability of acceptable representations to the data structure complexity. The second one is a stochastic method based on a bootstrap process aiming to increase the reliability of the induced neighbourhood structure. The resulting (robust) map, called R-Map, is more robust relative to the sensitivities of the outputs to the sampling method and to some of the learning options of the SOM' algorithm (initialisation and order of data presentation). This method consists in selecting one map between a group of several solutions resulting from the same self-organizing map algorithm, but obtained with various inputs. The R-map can be perceived as the map, among the group of solutions, and corresponds to the most common interpretation of the data set structure.*

Key words – Self-Organizing Maps, Robustness, Reliability, Bootstrap, Neighbourhood, Variability, R-Map.

1 Introduction

In the context of classification and data analysis, Self-Organizing Maps (SOM) focus on the neighbourhood structure between classes. Understanding how the complexity of the data can give a rise to several interpretations as increasing the stability of the neighbourhood structure can make SOM more attractive for some users who are confused due to possible various interpretations. The

aim of this paper is to present two complementary approaches (see [6] and [7]) to understand and reduce SOM's neighbourhood structure variability. Among the many causes of such variability, the complexity of the data structure and the learning options of the SOM algorithm are the main ones. Numbers of articles dedicated to the Kohonen algorithm theory specifically focus on convergence (see [1] and [4]) and sensitivity to parameters (initialisation, the order of data presentation, rate of decrease of neighbourhood function, adaptation parameter...). In the same vain, we propose on one hand a visualizing tool to diagnose a link between variability and data complexity, and on the other hand a two-step procedure aiming to increase the reliability of SOM neighbourhood structure.

The first approach is based on the idea that, as usual in data analysis, the simplicity of the representation comes into conflict with the data complexity. In the case of SOM, several neighbourhood organisations of classes may be acceptable. The technique presented here matches the analysis of the data intrinsic structure with the way SOM learns it in order to detect an eventual alternative organisation. For example, when the data structure is complex, the map can adjust it with a fold. In this case, during the learning of the algorithm may choose among a variety of folds that are all satisfying candidates. This situation can thus produce different "equivalent" maps when SOM is executed several times in a row. To connect the resulting neighbourhood structure variability to the complexity of the data intrinsic structure, a visualizing tool, called "Map of Distance between Classes" (MDC), is presented hereafter. This tool allows the extract the main information from a matrix of distance between classes that is rather large (for example, in the case of a map 10 by 10, the matrix has 10,000 values). The MDC represents both large and local distances relying them to the neighbourhood structure. That way, other proximities in the input space than those described by the map are then revealed. To conclude, the MDC is an easy tool for analysing or warning for some variability coming from complexity of the data.

The second approach provides a two-step stochastic method based on a bootstrap process (see [3] and [2]) to increase the reliability of the underlying neighbourhood structure. The increase in robustness is relative to the sensitivities of the output to the sampling method and to some of the learning options (initialisation and order of data presentation). At the first step, a bootstrap process is used to build a table of probability for any pair of individuals to be compared. At the second step, we choose between several maps the one - called R-Map - which exhibits the greatest similarity with this table. Finally, the R-map gives a summary of the data and a neighbourhood structure between classes that is less sensitive to the sampling (due to the first step treatment), to the initialisation and the order of the data presentation (thanks to the second step treatment). The R-map can also be considered as the most common interpretation of the structure among several SOM' solutions. We do not consider that the R-map is the "best" map concerning the interpretation. On the contrary, the variability of interpretations is probably rich in information, especially when one can compare various interpretations with the "common" one. As this second approach second method is generally very time-consuming, it is recommended to first use the Map of Distance between Classes to find eventual structural reasons for variability (in the data structure).

The two approaches are complementary but results from different studies whose contexts are different, respectively biometrics typology and financial strategy. In one case, variability entails curiosity when, in the other case, it means a financial cost. We decided to present both approaches separately, in contexts where their relevance are evident. In the first part, SOM is used to deduce differences causes of variability from the data description (see [7] for the complete description). The MDC is used in the case of the human facial skin sample, that is proven to be complex by a comparison between several previous analyses. In the second part, SOM is used in order to classify financial fund evolutions. The result is satisfying as it is coherent with further complementary analyses, but a crucial point is linked to the variability of such analysis. In particular, the evolution in the time of the map – not linked with a problem of reliability of one specific representation of the structure – can provide further information regarding the composition of a financial market. An

illustration of the R-map is provided below, using a financial database containing hedge fund Net Asset Values (see [5] for a presentation of the database and a financial application).

2 Understanding Variability of SOM' Neighbourhood Structure represented by the Map of Distances between all Classes

SOM' algorithm, completed with its own map representation, is a very flexible method in data analysis and representation. Nevertheless, a high-complex data structure might be difficult to adapt. In particular, a simple representation can be proposed and provides a satisfying interpretation of result, but at the price of a non-robust result. This section aims to link the variability of the output with the intrinsic data structure in order to analyse or warn for it. When this link is established, the variability of potential acceptable representation as an extra information instead of the sign of a lack of robustness. To this end, we propose the Map of Distance between Classes, called MDC, that visualizes the data intrinsic structure overlying to the SOM interpretation. This map is the projection of the matrix of distances between classes on the SOM network. This matrix is a very accurate description of the data structure but is rather large (in the case of a map 7x7, 2401 values are computed). The MDC realises two objectives. From one side, it allows the use of the large matrix of distances for compressing redundancy and summarizing the data structure into terms such as proximities. From the other side, it connects this summary of the input to the SOM interpretation. Finally, the MDC makes the reasons of variability emerge when they refer to the data structure. In order to illustrate the building and the properties of the MDC, we used an example in the context of biometric: the typology of human facial skin proposed by the C.E.R.I.E.S, from a data set of 212 women, that contains the measures of the intensity of seventeen visual or tactile criteria (for the complete study, see [7]). Several clustering methods have shown the complexity of this data intrinsic structure, leading partially to different interpretations. A previous study proposed a new typology with SOM. In a second time, the MDC and the projection of the different classifications on the resulting map revealed that the typologies remoteness coincide with a folder on the map. So, such as the survey of skin characteristics, a data intrinsic structure is able to induce several interpretations and as a result several different neighbourhood structures when using the SOM algorithm. In this case, when it is controlled by MDC, the variability is positive and increase the information set needed for a quality analysis.

The MDC uses any unit of the SOM network as a graphical display to represent a line of the distance matrix. Figure 1 represents the first line of the matrix of distance between classes (referring to centroïds distances). In each unit u, the MDC represents the distance between class C_1 and class C_u. The level of grey defines the distance (the darker, the larger). This representation groups distances to neighboured classes. When grouping this way, the MDC treats a part of the matrix redundancy as distance to neighboured centroïds are closed. Instead of reading the line of the distance matrix value per value, one may also consider it area *per* area (for example 49 values can be compressed in 3 area). Figure 1 shows that class 1 is close to its own neighbours, but also to classes 42 and 49 that was unexpected. This shows that the map makes a fold. Figure 2 visualizes the MDC: the box u displays the u line of the matrix such as the colour of its own unit u' indicates the value of distance between classes u and u' (the first box displays figure 1). That way, the MDC treats another part of redundancy (two neighboured boxes must be similar as much as distance to neighboured centroïds are closed) such as one can consider area of boxes. In the facial skin typology, Figure 2 confirms the coherence with the neighboured structure (neighboured classes coincide with small distances). Nevertheless, boxes 1, 8, 42 and 49 indicates a fold in the map.

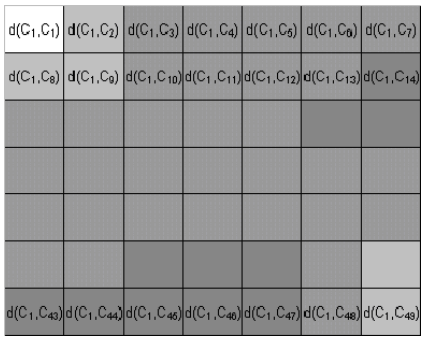

Figure 1: Distances between class 1 and others classes. Unit *u* grey level indicates the distance between class 1 and class *u* (the darker, the higher). Since the distance between class 1 and class 49 is small, the map makes a fold.

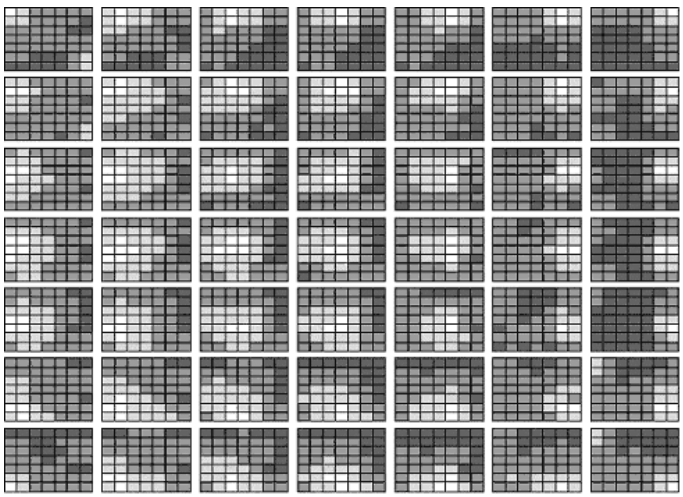

Figure 2: Map of Distances between all Classes. The grey level of unit *u* from box *u'* indicates distance between class *u* and class *u'*. Box 1 corresponds to figure 14.

This fold is confirmed when projecting the map centroïds on the first principal plane. In figure 3, centroïds are projected on the first principal plane, and are closely in connection with four of their eight neighbours. This way, the map network is represented in the input space as a surface that adjusts at best the data. Once the border of the surface is drawn, one can see once again that the folder is confirmed. By projecting on the map several other classifications (resulting from hierarchical clustering methods or segmentation), the folder is revealed: it corresponds to the area where are the main differences between the various typologies (see [7]). As a conclusion, when variability is due to data structure complexity, a solution can come from changing the SOM network structure (tree dimensions or a cylindrical structure…). Otherwise, the following method – leading to the so-called R-Map – could be a satisfying alternative way for increasing the robustness of the induced analysis.

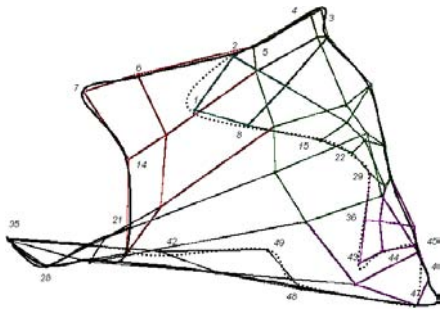

Figure 3: when joining class centroïds with four of them eight neighbours, self-organizing map interprets the data set structure from a flexible surface. The border of the surface is drawn. In the example of the skin typology the surface makes a folder that may cause variability.

3 A Bootstrap Scheme for Building the Table of Probabilities for Individuals to be Neighbours in a Map

While section 2 aims to distinguish the sources of variability coming from database complexity from such resulting form the methodology, this section is dedicated to reducing the second one. When SOM' are used in classification, the algorithm is applied to the complete database that is generally a sample of some unknown stationary distribution. A first concern refers to the question of the stability of the SOM' solution (specifically the neighbourhood organisation) to changes in the sample. A second concern regards the stability to the data presentation order and the initialisation. For limiting the dependence of the outputs to the original data sample and to the arbitrary choices within an algorithm, it is common to use a bootstrap process with a resampling technique. Here, this idea is applied to the SOM algorithm, when estimating an empirical probability for any pair of individuals to be neighbours in a map. This probability is estimated by the number of times the individuals have been neighbours at ray 1 when running several times the same SOM algorithm using re-sampled data series (see Figure 4). In the following, we call $NEIGHT_{boot}$ the table containing empirical probabilities for two individuals to be considered as neighbours at the end of the classification. The algorithm uses only individuals in the given resampled set of individuals (representing around 60% of the original population). At the end, the individuals left are classified using computed distances to centroids. Thus, at each step, the table of empirical probabilities concerns all individuals in the original dataset even if only a part of them have been used within the algorithm.

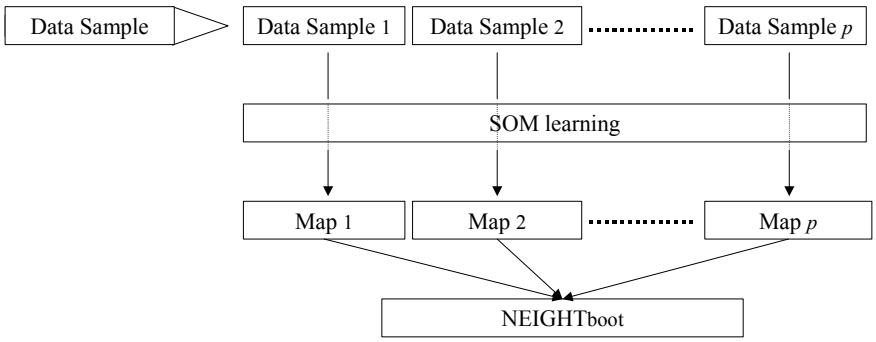

Figure 4: Step1, bootstrap process in order to build the table $NEIGHT_{boot}$ of individual's empirical probability to be neighbours one-to-one.

3.2 Choosing the R-map from the Table of Individuals' Probability to be Neighbours

When the table $NEIGHT_{boot}$ is built, the first step is over. In the second step (see Figure 5.), the SOM algorithm is also executed several times, but without resampling. For any map, we can build the table $NEIGHT_{map}$, similar to previous one, in which values are 1 for a pair of neighbours and 0 for others. Then, using the Frobenius norm, we can compute the distance between both neighbourhood structures, defined respectively at the end of step 1 (resampling the data) and step 2 (computing several maps with the original data), as follows:

$$D = \frac{1}{N^2}\sqrt{\sum_{(i,j)\in P}\left(NEIGHT_{boot}(i,j) - NEIGHT_{map}(i,j)\right)^2} \qquad \textbf{(1)}$$

where P is the set of N^2 individuals pairs (i,j).

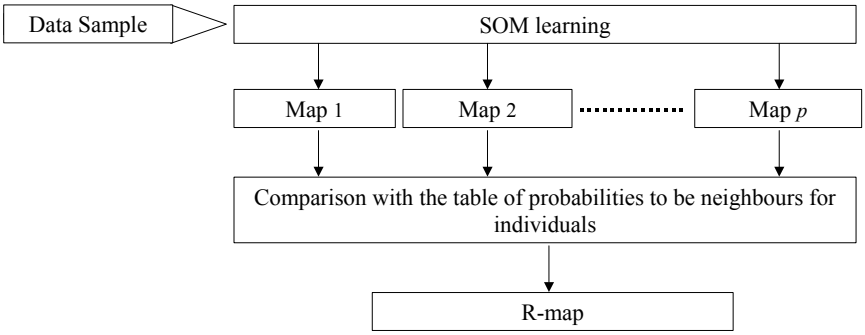

Figure 5: Step2, selection of the R-map between p solutions of the SOM' algorithm

The selected R-Map is the one among all SOM' solutions that minimises the distance *D*. The R-Map indeed gives a summary of the data and of a neighbourhood structure that are less sensitive to the sampling (after to the first step), and to the initialisation and order of the data presentation (after the second step). The R-map can then be considered as the most common interpretation of the structure.

3.1 Choosing the R-Map from the Table of Individuals' Probability to be Neighbours

We compare hereafter the classical 'simple' SOM method with the two-step algorithm presented in previous sections considering the aspect of reliability. More precisely, we first build several maps with the SOM algorithm as in [6] and, second, several maps leading to the definition of the R-map. We then compute both related tables of individual's one-to-one probability to be neighbours: the $NEIGHT_{map}$ (set of 'simple' SOM with no two-step bootstrap process) and the $NEIGHT_{R\text{-}map}$ (corresponding to the two-step algorithm presented above). The theoretical probabilities 1, 0 and U_N/U correspond respectively to individuals that are respectively (surely) neighbours, (surely) not neighbours and neighbours (almost) by chance. The ratio U_N/U corresponds to the uniform probability for a pair of individuals to be affected by mere chance in the U_N classes belonging to U classes (here, U_N equals 9 at ray 1).

As an illustration, we use a study already published, in which was proposed a robust typology of hedge funds based on the NAV time-evolutions, leading to a sound financial characterisation of the typology with external risk measures (see [6] for more details). The typology of reference (given by $StandandPoors^{TM}$) based on the portfolio manager declared strategies is known to be unsatisfying for many reasons, from the lack of transparency of proprietary well-protected strategies, data error, changes in situation or complex mixture of 'pure' strategies. In this context, SOM is used in order to clean the classification. The confidence in the neighbourhood is obviously crucial when real financial applications (fund of funds asset allocation and risk management) are at stake. In this practical example, the data set is composed with 294 funds and 67 observations of monthly NAV from January 1995 to September 2000. The chosen structure of the map is a six-by-six grid.

Thirty R-maps are build each time in step one (using each time thirty samples of randomly chosen 194 individuals amongst the possible 294 hedge funds)**.** From these 30 R-maps, we can build a table of individuals' probability to be neighboured one-to- one called as before $NEIGHT_{R\text{-}map}$. From 30 new self-organizing maps, we can build the equivalent $NEIGHT_{map}$. As an example, we present below Table 1, which is an abstract of three tables glued together: two tables called $NEIGHT_{map}$ (resulting from two repeated SOM on the same original data) and one table called $NEIGHT_{R\text{-}map}$

(resulting from the two-step procedure). Figures in Table 1 concerns empirical probabilities of pairs composed with individual number 1 and individuals numbered from 25 to 33. We can see that, for any pair, the empirical probabilities in the column R-map are closer to 1 (respectively to 0) when they are higher (lower) than 25.00% (*i.e.* U_N/U = 9/36 here). This property indicates that the neighbourhood structure with R-maps is more reliable than with classical maps. In table 2, columns indicate, for any empirical probability to be neighbours, the number of pairs concerned. When we take into consideration the R-map, 42.14%, 16.22 %, 12.41% and 8.58% of the set of pairs of funds have an empirical probability to be neighbours, respectively, lower than 10.00%, greater than 80%, greater than 90%, equal to 100.00%. In comparison, in the case of classical maps, the respective values are 27.85%, 12.03%, 6.07%, 1.4% in a first case (Map 1) and 19.98%, 11.8%, 5.5% et 1.35% in a second case (Map 2). Thus, table 2 shows as well the greater reliability of the neighbourhood structure in the case of R-maps when using the two-step algorithm procedure.

Table 1. Empirical Probability for Funds to be Neighbours

Couples of Funds		Probability of being Neighbours (in %)		
Fund #1	Fund #2	in Map 1	in Map 2	in R-map
...	...	...	...	...
1	25	1.00	0.97	1.00
1	26	0.93	0.80	0.93
1	27	0.97	0.87	1.00
1	28	0.97	0.87	1.00
1	29	0.93	0.80	0.97
1	30	0.17	0.03	0.00
1	31	0.13	0.03	0.00
1	32	0.47	0.50	0.50
1	33	0.23	0.20	0.00
...	...	...	...	...

Table 2 Frequency and Cumulated Frequency of the Probability for Funds to be Neighbours

	Frequency (in pairs)			Cumulative Frequency (in %)		
Probability	in Map 1	in Map 2	in R-map	in Map 1	in Map 2	in R-map
[0.00;0.10]	24 076	17 271	36 420	27.85	19.99	42.14
[0.10;0.20[	10 270	14 568	8 196	39.74	36.84	51.62
[0.20;0.30[	9 664	10 184	6 698	50.92	48.63	59.37
[0.30;0.60[	23 640	8 648	4 792	78.27	76.34	75.59
[0.60;0.70[	5 175	6 106	3 374	84.25	83.40	79.49
[0.70;0.80[	3 412	3 964	3 708	88.20	87.99	83.78
[0.80;0.90[	5 448	5 150	3 288	94.50	93.95	87.59
[0.90;1.00]	4 751	5 229	10 728	100.00	100.00	100.00

3.3 Remarks

As partially shown in the previous illustration, R-map method reduces SOM sensitivity to three parameters (the sample, the data presentation order and the initialisation). A similar technique can include others (the adaptation parameter), but not parameters linked to the neighboured structure (the size of the map). As a second remark, we can indicate that the distance *D* used here is not symmetrical for neighbours and non-neighbours, as a random distribution into the *U* units would create the probability 9/*U* to find neighbours by chance. Such is SOM itself, as individuals defined as neighbours are closed in the input space, but closed individuals can belong to un-neighboured classes (for example when the map is folded). The use of such distance reduces more the possibility to find individuals that are "neighbours by chance" than "non-neighbours by chance". As a remedy,

we can think about using another (*quasi*-symmetric) distance[1]. As a third remark, the R-map does not need more capacity to be computed than the usual Kohonen map, except that the $NEIGHT_{boot}$ table must be kept into memory. This table can be very large (N^2 pairs of N individuals) but can be reduced to a list of pairs such as the distance D would still be significant.

4 Conclusion

The two complementary methods presented here treat the variability of the SOM results. The first one separates the structural variability due to the data, and the second reduces the variability due to the sampling and some SOM parameters. The effects of the data structure are revealed by a map, called MDC, that allows to interpret the matrix of distance between all classes. The method to increase robustness consists in selecting one map between a group of several solutions of the same self-organizing map algorithm. The selected map, called R-map, can be perceived as the map, among the group, that corresponds to the most common interpretation of the data set structure (interpretation means, here, the classification and the neighbourhood structure between classes). The neighbourhood structure is generally more robust with R-maps than one of a randomly selected map among the group. This reliability concerns both sensitivities to the sampling and to some algorithm parameters, in particular the initialisation and the data presentation order. Finally, above aiming to recover robust classification, R-map selection could be a practical way to deliver to self-organizing map users that gives the same result when they are executed several times in a row conditionally to the assessment from the data structure with the MDC.

References

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

[2] De Bodt, E., Cottrell M. (2000), Bootstrapping Self-organising Maps to Asses the Statistical Significance of Local Proximity, *European Symposium on Artificial Neural Networks*, p. 245-254.

[3] Efron, B.,Tibshirani, R. (1993), *An Introduction to the Bootstrap*, Chapman and Hall.

[4] Kohonen, T. (1995), *Self-Organising Maps*, Springer, Berlin.

[5] Maillet, B., Rousset, P. (2003) Classifying Hedge Funds with Kohonen Maps: A first Attempt, *in Connectionist Approaches in Economics and Management Sciences*, Lesage C., Cottrell M. (Eds).

[6] Rousset P., Maillet B. (2005) Increasing Reliability of SOMs' Neighbourhood Structure with a Bootstrap Process, *mimeo* Paris-1, March 2005, 6 pages.

[7] Rousset, P., Guinot, C. (2001) Distance between Kohonen Classes: Visualization Tool to Use SOM in Data Set Analysis and representation, *International Work-Conference on Artificial Neural Networks* 2, p. 119-126.

[1] Such as, for instance: $D_1 = \frac{1}{N}\sqrt{\sum_{(i,j)\in P} \frac{\left(NEIGHTB_{boot}(i,j) - NEIGHTB_{map}(i,j)\right)^2}{\left(NEIGHTB_{boot}(i,j) - 9/U\right)^2}}$.