Some Other Applications of the SOM algorithm : how to use the Kohonen algorithm for forecasting

Marie Cottrell
SAMOS-MATISSE, CNRS UMR 8595, Université Paris 1
cottrell@univ-paris1.fr

Abstract: The Kohonen algorithm has very interesting properties of self organization, which are widely used for exploratory data analysis and visualization. But the Kohonen maps can also be useful to forecasting tasks, study of temporal evolutions, explanation of complex prediction models.
The examples that are used to present the methods are issued from several papers by Patrice Gaubert, Bernard Girard, Patrick Letrémy, Patrick Rousset, Joseph Rynkiewicz.

1 Introduction

We suppose that the reader is familiar with the Kohonen algorithm. See for example Kohonen (1984, 1993, 1995), Kaski (1997), Cottrell, Rousset (1997) for an introduction to the algorithm and to its applications to data analysis.

It is an original classification algorithm which presents two essential differences with the traditional methods of classification: it is a stochastic algorithm, and an a priori neighbourhood concept between classes is defined. The neighbourhood between classes may be chosen on a wide range of representations: grid, string, cylinder or torus, called Kohonen maps. The classification algorithm is iterative. The initialisation consists of associating a code vector (or representative) to each class, chosen at random in the space of observations. Then, at each stage, an observation is randomly chosen, compared to all code vectors, and a winner class is determined, i.e., the class of which the code vector is nearest in the sense of a given distance. Finally, the code vectors of the winner class and those of the neighboring classes are moved closer to the observation.

Hence, for a given state of code vectors, an application associating to each observation the number of the nearest code vector (the number of the winning class) is defined. Once the algorithm converges, this application respects the topology of the space of entries, in the sense that after classification, similar observations belong to the same class or to neighboring classes. This main property of the Kohonen algorithm is the so-called topology conservation property. This feature allows to represent the proximity between data, as in a projection, along the Kohonen map.

After training the Kohonen map, each class is represented by its code vector, its elements are similar between them, and resemble the elements of neighbor classes.

An inconvenience of the basic algorithm is that the number of classes needs to be determined a priori. In order to palliate for this inconvenience, a hierarchical type of classification of the code vectors is undertaken, so as to define a smaller number of classes called clusters. The classes and clusters can then be represented on the Kohonen map corresponding to the chosen topology. As neighbouring classes contain similar observations, the clusters gather contiguous classes, which gives interesting visual properties. It is then easy to establish a typology of individuals, by describing each of the clusters by means of traditional statistics.

These properties are well known and widely used for exploratory data analysis, visualization of multidimensional data, segmentation of complex data etc.

According to the problem, one may also use each cluster to define a specific model (regression, auto-regression, factor analysis, etc…).

In the following, we present other applications of the Kohonen algorithm for studying temporal processes. Section 2 deals with an application of the Kohonen maps to forecast fixed-size vectors. In section 3, we show how to represent individual trajectories on a Kohonen map. Section 4 presents an example where the Kohonen classification helps to interpret the explanatory variables.

2. Using Kohonen classification to forecast vector data

When the problem is to predict a curve or a vector (for example a consumption curve for the next 24 hours), the usual prediction methods can be deceptive. The vectorial ARMA methods are difficult to use, present some theoretical drawbacks. The cost function associated to Multilayer Perceptron with vectorial outputs is not easy to implement. The problem can also be considered as a long-term prediction problem, where each predicted value is introduced as a new input to predict the next one. The difficulty is that the quality of the prediction decreases with the term, as well in linear model (the prediction squashes) as in non linear model (in MLP models for example, the prediction can become chaotic).

When all the vectors to predict have the same dimension and can be considered as a curve of a fixed size, a Kohonen maps can provide very good results, as shown in Cottrell, Girard, Rousset (1998). The idea is very simple. For each curve, we define its mean, its variance, its normalized profile, where the level effect is cancelled, and the variability is standardized.

Then the problem is divided into four steps :

1) to forecast the mean and the variance of the next vector; any classical method (linear, non linear as MLP) can be used for both cases, since it is only a one-step prediction,

2) to make a classification of the normalized profiles, on a Kohonen map, and identify on the map what are the past days which belong to each Kohonen class,

3) to estimate the profile of the next day, by looking for the similar days in the past (Tuesdays in October, Sundays in June, etc.) on the map. The weighted mean of the code vectors of the similar days is taken as expected profile for the day to predict.

4) to redress the profile, by multiplying by the standard deviation and adding the mean), to compute the predicted curve.

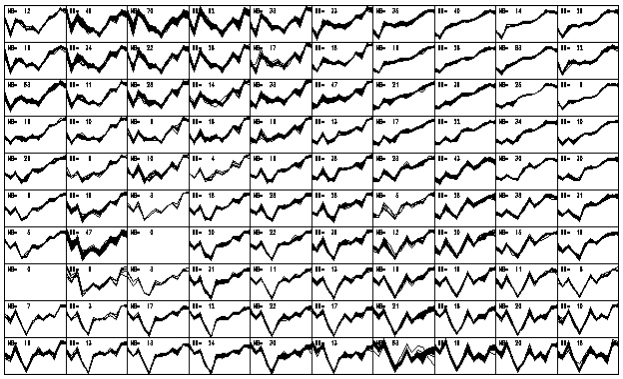

Figure 1 : classification of all the profiles on a cylindrical map.

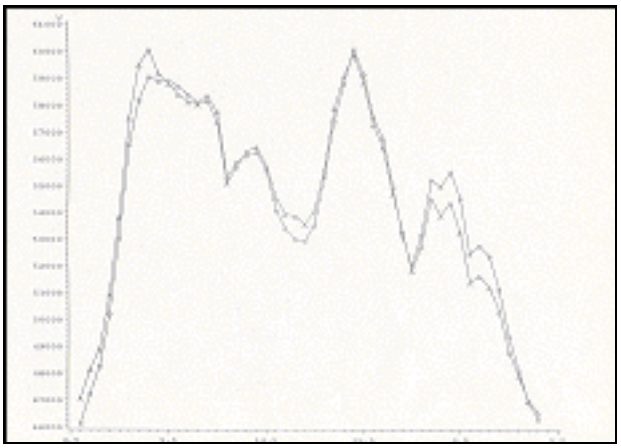

Figure 2: Real curve and its prediction for Tuesday 15 January 1991.

This method is low consuming time and easy to implement. It has been used in various real world problem, Cottrell et al. (1995), Cottrell, Rousset, Girard (1998).

3. Using Kohonen classification to study individual trajectories

This kind of application of the Kohonen was suggested by Serrano-Cinca in Deboeck & Kohonen (1998), and is widely used in Gaubert, Cottrell (1999), and Akarçay-Gürbüz, Perraudin (2002) for example.

The idea is also very simple. Let us suppose that we have several observations for each individual, corresponding to measurements made for several dates (or years). If all the observations are classified on a Kohonen map (as if they were different data), it is possible to study the change of "state" of a given individual along time.

Let us consider for example the data that are studied in Gaubert, Cottrell (1999). They are 2507 heads of households, present in their family during the period 1984-1992. Each of them is described by 15 variables (age, family size, characteristics of the main job), measured in 1984, 1988, 1992. A Kohonen classification is achieved on 7521 observations on a 8 by 8 bi-dimensional map.

On the resulting map, each part can be identified with some particular situation (the lower left corner contains individuals with no job most of the year, the central region contains people exerting more than one job at the same time, in the upper right corner there are the best job situations, with stability and high pay, etc.).

So it is possible to draw for each individual its trajectory along the period that was observed. In this case, the Kohonen map helps us to define a trajectory and to visualize it. Further, by grouping the 64 Kohonen classes into some super classes (the authors consider 7 and 4 super classes), it is possible to quantify the transition probability to stay in the same class all the time, to change from a class to another, to study the stability of the labor market, and so on.

See for example below two examples of such trajectories.

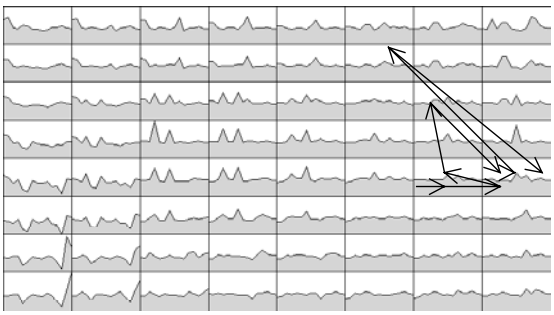

Figure 3 : Trajectory of an individual staying in good job situation during the whole period

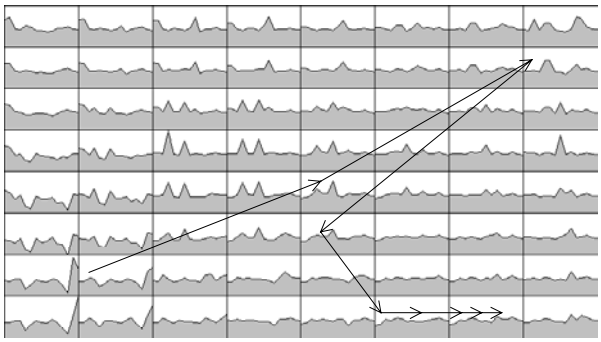

Figure 4 : Trajectory of an individual leaving the more precarious situation to reach, after one year in a good situation, an intermediate position

4. Using Kohonen classification to interpret the prediction

In this example, the Kohonen algorithm is used after a first modelization to explain the results. Let us suppose that to predict some time series, the best model seems to correspond to several different regimes (or hidden states). The data seems to be piecewise stationary, and several models are necessary to model them. As the changes of regimes occur at unknown epochs, it is usual to use an Hidden Markov Model (HMM) combined with several prediction models, which can be linear or non linear. If the HMM model seems to identify several well separated regimes, a Kohonen classification can be used to achieve the identification of these regimes.

Let us take as example a study by Rynkiewicz, Letrémy (2002), which deals with the forecasting of the ozone pollution level in Paris. The classical statistical models give good prediction, but their results are not quite good for the pollution peaks. The authors use an autoregressive model, with Markovian regime changes, that they model by using a HMM. The quality of the prediction is increased, two regimes are necessary, and a clear segmentation of the pollution time series does appear.

Here the Kohonen map is a tool which give a clear interpretation of the nature of these two régimes.

For the prediction, the inputs are: the maximum of the pollution rate on the day before, the global radiation, the mean speed of the wind, the maximal temperature of and the temperature gradient of the day. Two states for the hidden Markov chain are sufficient, two different autoregressive models are defined, one is linear and seems to be associated to the low or medium values, while the second is a Multilayer Perceptron, specialized in the high values.

To better understand the nature of both hidden states, the authors classify all the observations (that are 5-dimension vectors) in a 7 by 7 Kohonen map. These 49 classes are grouped into 5 super classes, easy to interpret.

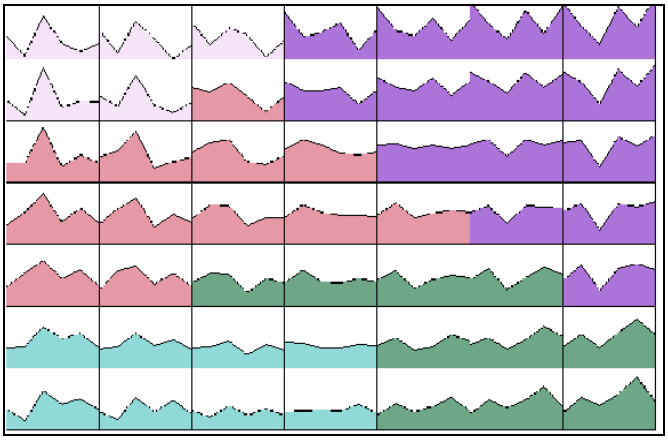

Figure 5 :The code vectors on the Kohonen map

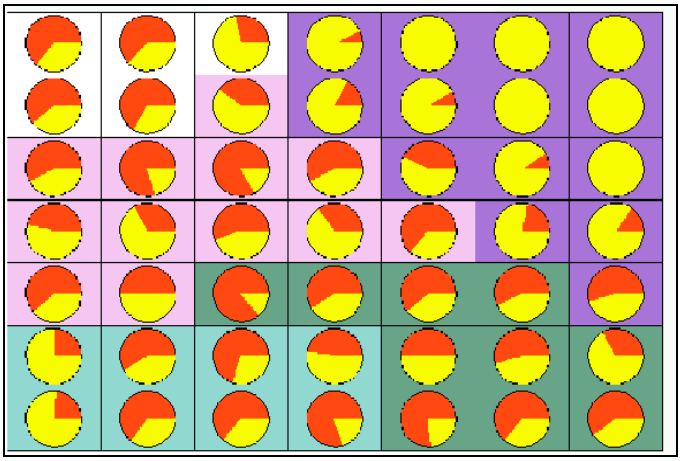

Figure 6 : Distribution of the two states on the Kohonen map, light gray is high pollution (2), obscure gray is low pollution (1).

The upper right corner contains the situations with high pollution levels, low wind, high temperature and gradient. Almost all the observations in this zone were identified by the non linear model, that is the state 2 of the HMM. Below, there are classes with observations whose values are near the means (except the temperature). The upper left corner contains the observations with low speed of wind and low gradient, etc. We can observe that the meteorological variables are not very discriminant to separate the hidden state 1 from the hidden state 2, which occurs in

almost all the regions on the map, except the upper right corner which is specialized in the state 2.

In this case, the use of the Kohonen classification shows that similar meteorological conditions observed on one day can produce both situations (high or low pollution). However the performances of the prediction are quite good, which means that the persistence in time of these meteorological conditions is essential, and are not shown in the Kohonen map. It would be necessary to use a vector with more components, by taking not only the variables of the day, but also the previous ones.

5. Conclusion

The Kohonen techniques show their very large capabilities to study other aspects of data analysis. In particular, the combination of the Kohonen algorithm with temporal data and times series allows to put in evidence some characteristics of the data which could be obscure or little visible.

All these properties contribute to the very large popularity of this important and so useful algorithm.

References

Akarçay-Gürbüz A., Perraudin C.(2002), Comment situer l'économie de la Turquie parmi les économies de l'UE? Une analyse exploratoire. *Proc. ACSEG 2002*, Boulogne sur Mer.

Cottrell M., Girard B., Girard Y., Muller C.and Rousset P. (1995), Daily Electrical Power Curves : Classification and Forecasting Using a Kohonen Map, *From Natural to Artificial Neural Computation, Proc. IWANN'95*, J.Mira, F.Sandoval eds., Lecture Notes in Computer Science, Vol.930, Springer, p.1107-1113.

Cottrell, M. & Rousset, P. (1997) : The Kohonen algorithm : a powerful tool for analysing and representing multidimensional quantitative et qualitative data, *Proc. IWANN'97*, Lanzarote, *Juin 1997*, J.Mira, R.Moreno-Diaz, J.Cabestany, Eds., Lecture Notes in Computer Science, n° 1240, Springer, p. 861-871.

Cottrell, M., Fort, J.C. & Pagès, G. (1998) : Theoretical aspects of the SOM Algorithm, *Neurocomputing,* 21, p. 119-138.

Gaubert P., Cottrell M. (1999), A dynamic analysis of segmented labor market, *Fuzzy Economic Review*, Vol. IV, N° 2, p.63-82.

Kaski, S. (1997) : Data Exploration Using Self-Organizing Maps, *Acta Polytechnica Scandinavia*, 82.

Kohonen, T. (1984, 1993) : *Self-organization and Associative Memory*, 3°ed.,

Springer.

Kohonen, T. (1995) : *Self-Organizing Maps*, Springer Series in Information Sciences Vol 30, Springer.

Rynkiewicz J., Letrémy P.(2002), Etude de la segmentation d'une série de pollution en niveau d'ozone, Communication to *Journées MAS 2002*, Grenoble.