CLASSIFICATION TEMP DATA WITH SELF-ORGANIZING MAPS

D. Lahoz and M. San Miguel

Abstract. Given the complexity of the data TEMP, meteorological measurements in height realized with weather balloon, is not habitual its directs use in the models of meteorological prediction. The majority of the models use some type of group of the such data as Analysis of Principal Components, Analysis Cluster and others. In this paper the classification is done with Self- Organizing Maps (SOM)(cf. [1]). Before there has been obtained the number of cluster in the data(cf. [2]). They are calculated the references classes in some data TEMP series measured in weather station of the INM (Instituto Nacional de Meteorología).

Keywords: Self- Organizing Map, TEMP, Classification, Cluster, Meteorological data

AMS classification: 62-07, 68T10, 91C20, 92B20

§1. Introduction

This paper is a part of a general model for the study of the wind and their proprieties in wind farms of Ebro Valley. For the implementation of a Physical- Mathematical Model ([3]) for the prediction of the wind is necessary to study the different meteorologic variables. The meteorologic data of the Instituto Nacional Metoerología, INM, and others official meteorologic institutions) are collected basically in two data types: the SYNOP data and the TEMP data. The first one are measures taken at floor level in the meteorologic stations. Between others recollect the wind speed, the wind direction, temperature,... The TEMP data are another measurements performed with weather balloon in the same weather stations. The balloon perform measurements of the wind speed, wind direction, temperature, pressure,... to different heights. The main problem of TEMP data is his complexity, too many measurements in too many different levels. It is for this reason that the TEMP data cannot to be included directly in the prediction model. In this work a study was carry out for the classification of this data type. For this classification cluster techniques and Self- Organizing Maps tools have been used.

In the second section, the principal ideas of Self- Organizing Maps (SOM) are showed. The next section shows the tools for the obtaining of number of cluster.

The fourth section realizes the descriptive study of the TEMP data. The variables measures, the number of them in a TEMP data, the highs of reference and the variables selected for the study of classification are explained.

The next section shows the results obtained in the application of the techniques and tools of the second and thirst sections in the TEMP data studied in the fourth section.

The last part says the conclusions of the article and possible lines of study to continue the work.

§2. Self- Organizing Maps

The Self- Organizing Maps (SOM), or Kohonen Maps, were used for first time by T. Kohonen in 1981. In 1984 T. Kohonen publishes his book *Self- Organizing Maps* ([1]) where he collected the ideas and tools of SOM techniques. From that moment the technique has been applied to different fields. The SOM is a technique that is within the Artificial Neural Networks. More concretely, it is unsupervised, feedforward and competitive network.
Usually, the SOM is used for the classification of a big data collection. As Neural Network can to work without problems of dimensionality or number of data. The SOM is a classification tool and in some applications, such as in this paper, it is combined with tools of classical statistic cluster. The SOM searches patterns within of the data, without it has got output pattern (unsupervised).
The SOM have the advantage of reproduce the space of data series in a space of dimension two. Logically is more easy to study the proprieties in a two-dimensional space that in a space of higher dimension. In a two-dimensional space exits graphical tools to extract the cluster existed in the data and to calculate the number of those. The graphical and numerical tools used for the study of proprieties of two-dimensional space generated for the SOM are the same that, usually, use the classical cluster.
The basic SOM algorithm and his neural structure (figure 1) are the next:

1. the input of SOM neural network are alls data (pattern in the SOM notation), x_p in the figure 1
2. the input x_p is compared with all neurons of the reference space, m_i in the figure 1
3. to calculate the winner, the neuron $m_g = min_i dist(x_p, m_i)$
4. to modify the winner, m_g, and his neighbourhood, $N(m_g)$, for m_i close to x_p.

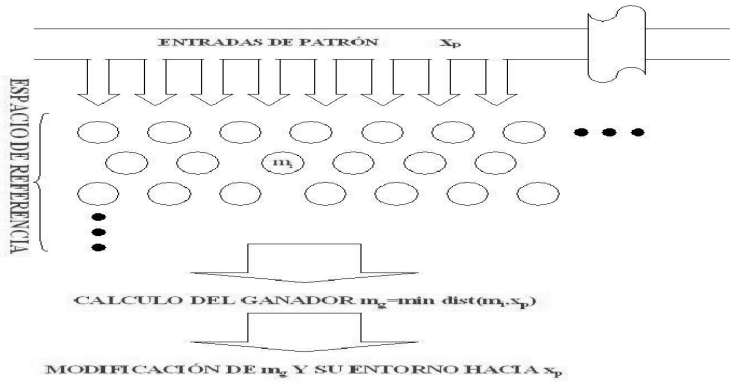

Figure 1: SOM Neural Network basic

The SOM is a Neural Network (NN) where all the inputs connects with all the neurons of the reference space. Also, it is a feedforward NN, the information only goes from the input layer to the output layer, and a competitive NN, there is a winner.
There are many parameters that need to be fixed to begin to appy the SOM algorithmic. Possible changes in the parameters of the basic SOM are:

- They are different reference spaces, of dimension 2 (Sheet or Grid), of dimension 2 and closed (Cylinder) or of dimension 3 and closed (Toroid)([4]).
- To modify the close to winner is necessary to select a neighbourhood topology used. The most utilize are rectangular, hexangular or random neighbourhoods.
- Also, it must select the neighbourhood function that permit to calculate the nodes "nearest" to the winner. Some neighbourhood functions are the Gaussian, the Bubble, the EP,...
- For the calculus of distances is necessary to define a metric. The different metrics and distances more utilize are: Euclidean, Manhattan, dot product, special alphabetic metrics,...
- Usually the modification of valour of m_g and his neighbourhood is decrement with the time, in the first iterations the change is bigger than in the last iterations. The selected learning rate function will modify the speed of convergence of the algorithmic. Different learning rate functions are showed in figure 2

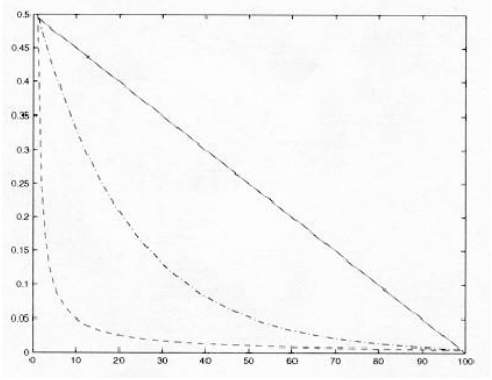

Figure 2: Different learning rate: $\alpha(t) = \alpha_0(1 - t/T)$, $\alpha(t) = \alpha_0(0.0005/\alpha_0)^{t/T}$ and $\alpha(t) = \alpha_0(1 + 100t/T)$ with T maximum number of iterations and α_0 initial valour

Another versions of SOM are: LVQ, ASSOM, FASSOM,...([1]):

§3. Tools for the estimation of the number of cluster

One of the main difficulties in the classification is to find the number of cluster in what divide the data collection. It exits multiples methods and coefficients for this calculus. In this article there has been used some coefficients of the cluster techniques and some graphical tools of the SOM.

In the classical cluster exists different coefficients to study the goodness of division. This measurements, normally, work with the distances denominated between-cluster and within-cluster, this is, they compare the union within a cluster with the relation of this with the others. The different between the coefficients is in the form of to compare those distances. The coefficients

used here are: the gap coefficient ([2]), the CH coefficient([5]), the KL coefficient([6]) and the H coefficient([7]). The expression of this coefficients is showed to continuation.
The gap coefficient of Tibshirani, Walther and Hastie ([2]) compare the within-cluster dispersion of any classification of the data with an appropriate reference null distribution. The expression of coefficient is:

$$Gap(\text{cluster k}) = E\{W_k^*\} - log\{W_k\}$$

with W_k the mean distance within cluster of data, and W_k^* the mean distance within cluster in a reference function (usually the Uniform function in the data range). Finally choose the number of cluster via:

$$\widehat{k} = \text{smallest k such that } Gap(k) \geq Gap(k+1) - s_{k+1}$$

with s_k a measure of the deviation of the reference function.
The second coefficient used is the Calinski and Harabasz (CH) coefficient ([5]). The coefficient expression is :

$$CH(k) = \frac{B_k/(k-1)}{W_k/(n-k)}$$

where B_k and W_k are the between and the within cluster sums of squares. The idea is to maximize CH(k), with CH(1) not defined.
The next coefficient was proposed for Krzanoswki and Lai (KL), ([6]),and it expression is:

$$KL(k) = |\frac{(k-1)^{2/m}W_{k-1} - k^{2/m}W_k}{k^{2/m}W_k - (k+1)^{2/m}W_{k+1}}|$$

with W_k the within cluster sums of squares. It is similar a to maximize $W_k k^{2/p}$, but the authors argued that it may have better properties.
The last coefficient is the Hartigan (H) coefficient ([7]). The expression the coefficient is:

$$H(k) = \left\{\frac{W_k}{W_{k+1}} - 1\right\}/(n-k-1)$$

with W_k the within cluster sums of squares. The idea is to start with $k = 1$ and to add a cluster as long as is sufficiently large. One can use an approximate F-distribution cut-off; instead Hartigan suggested that a cluster be added if $H(k) > 10$. Then the number of cluster selected is the smallest such that $H(k) \leq 10$.
More coefficients, and a comparative study between its, can to find in Cuevas et al. ([8]).
The SOM reproduce the data distribution in a reference two-dimensional space. In this the graphical tools can to be used to choose the number of cluster. One plot too usual is the gray scale plot ([9]) for the number of cases and the distance mean in each node of reference space. In section 5 can be seen the application to the TEMP data.

§4. The TEMP data

The TEMP data are measurements performed for the INM in different weather stations. Each six hours (0, 6, 12 and 18 ot'clock) a weather balloon is sent for the obtention of meteorological

measurements. The weather balloon realizes recollected of pressure, geopotential, temperature, point dew and speed and direction wind. The measurements are obtained in prefixed highs and in highs where a change is produced. Then, some TEMP data recollect many measurements and others one too little. In the figure 3 it can be seen the plot of TEMP data, that shows the the measurement of temperature (žC), point dew (žC), wind speed (knots) and the wind direction (ž grades) for the different geopotential, the pressure is not represented.

The principal problem of TEMP data is that each one haves got different number of measures

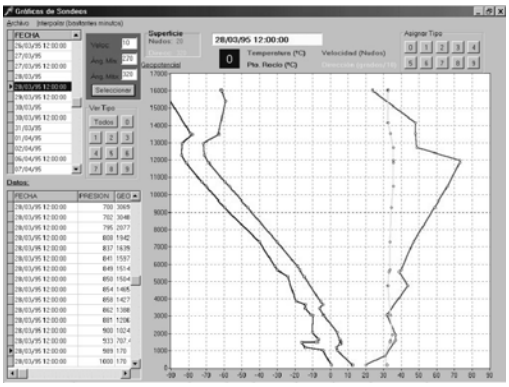

Figure 3: Plot of a TEMP data

and to different highs. For a first classification there has been used five pressures: 300, 500, 700, 850 and 100 mb. In each one of this pressures the other variables has been measured: geopotential, temperature, wind speed and wind directional. Each data have got twenty variables.

The measurements of INM begin the 14-10-1990 and it finish the 31-12-1999. Of this series there have been extracted only the data with alls measurements, the twenty variables. In total the studied series is formed for 33698 TEMP data. Logically has not missing data.

§5. The classification of TEMP data

In this section the TEMP data series of section 4 are classified with the techniques shows in the sections 3 and 2.

Before the calculus of the division is necessary to know the number of groups. For this, first, the coefficients show in section 3 are applied to the TEMP data series. The values obtain in the different coefficients and the number de cluster elected in each one of those is showed in figure 4. The number of cluster selected for this study is seven, the results of gap and KL coefficients.

The SOM method reproduce the data space in a reference space. If this reference space is, for example, two-dimensional, the graphics tools for the calculus of number of cluster can be used. Before the application of SOM is necessary to fix some parameters of method. In this work the SOM method has been applied with the next parameters:

- the reference space is a grid hexagonal two-dimensional with 20 x 15 nodes
- the training rate is lineal, this is, it goes decreasing lineally with each new step
- the neighborhoods are gaussian and the distance selected is the Euclidean distance

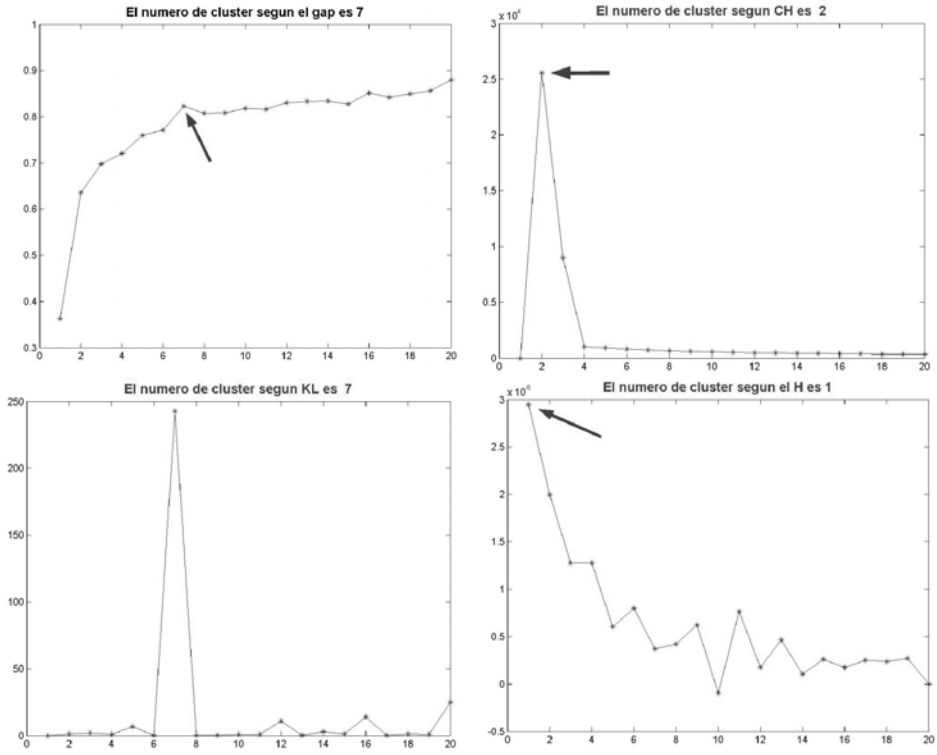

Figure 4: The results of coefficients in the TEMP data. The number of cluster elected for the gap, CH, KL and H coefficients are showed with a arrow

- to avoid the problem of different scales and units all variables used has been standardizing to [-1,1].

For last, the SOM algorithmic has been applied with a mean of twenty sort iterations (where the reference space reproduce the sort of data) and two hundred tune iterations (where the structure of reference space is fitness to the structure of data).
With this parameters the SOM obtain a reference space where to study the number of cluster with different graphical tools. In figure 5 are showed the Gray Scale plot for the number of data associates to each node (node more close), in the left figure, and the Gray Scale plot for mean distance in the cluster formed for each node, in the right figure. In figures 5 and 6 can be seen the division elected for the division in seven part.

Another graphic tools is the U-matrix (it calculate the distance of each node with their neighbouring) and it shows in the figure 7 with a colour scale, left figure. Also, it shows the colour scale for all nodes of reference space for the twenty variables studied, right plot.
The coefficients and the graphic tools in the reference space indicate that the studied data could be partied in seven cluster. For the calculus of clusters and their reference points, it is used the space reference with the same structure of the studied data and less number of data. The seven cluster are obtained with the method propose for Vesanto- Alhonimei ([10]), and the results of partition in the reference space is showed in figure. The difference between the seven cluster can to see in the figure 8, where the valour in each one of twenty variables is showed.

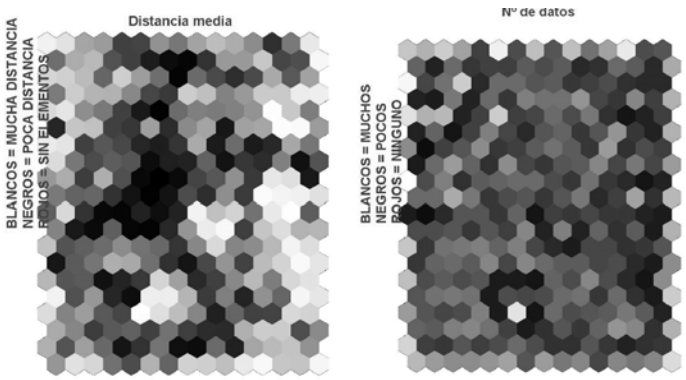

Figure 5: The left plot shows the Gray Scale plot for the mean distance of cluster formed for each node. The right plot shows the Gray Scale plot for the number of data in the cluster formed for each node. The gray scale goes since the black colour, more number of data or more mean distance, to black, less number of data or mean distance. The red colour is for the nodes without data in his cluster.

Figure 6: Division of reference space in the seven cluster

§6. Conclusions

In this paper a classification of TEMP data is showed. Besides, it has been used different coefficients for the calculus of the number of cluster. The partition in seven groups is proposed according to the results obtained for the coefficients and the SOM method.
The SOM tools are used for the calculus of reference point of each cluster. The classification of the data has been realize with the SOM. The study of the cluster is done only with the reference node of each group.
To confirm the goodness of the division is necessary to test it with another TEMP data, of different localizations. Also, some coefficients obtain a different number of clusters and this division can be more generals for all TEMP series.
It is necessary the point of view of a expert in meteorology for the contrast of the results of this study. The meteorologist would be able to say if the clusters obtained are a valid results in the TEMP data in this localization and in the TEMP data general.

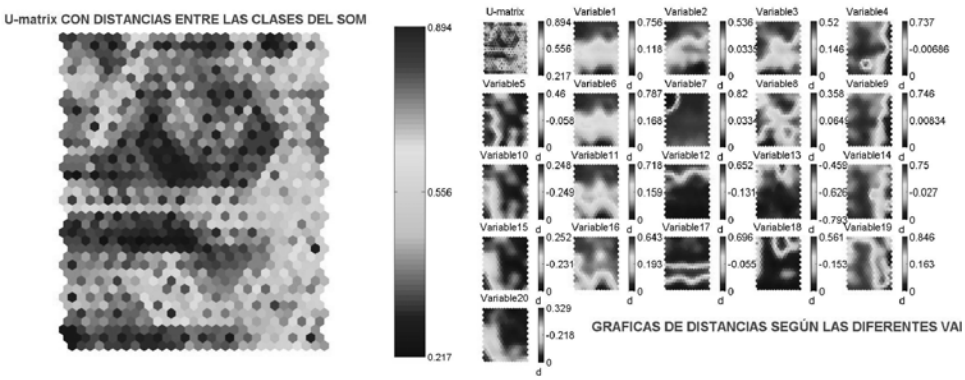

Figure 7: U- matrix of reference space, left figure, and colour scale of reference space and the twenty variables studied, right figure

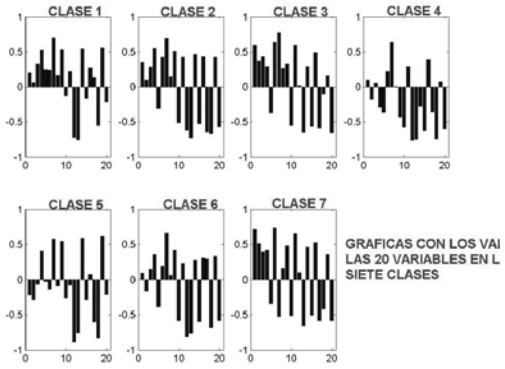

Figure 8: Plot with the valour of twenty variables in the seven reference cluster

Acknowledgements

References

[1] T. Kohonen *Self- Organizing Maps*. Springer, (1997, Second Edition).

[2] R. Tibshirani, G. Walther and T. Hastie. *Estimating the number of cluster in a data set via the gap statistic*. J.R. Statist. Soc.B 63(Part 2), pages 411-423. (2001).

[3] J. A. Escudero Lázaro, D. Lahoz Arnedo. *Physical- Mathematical wind prediction model*. VII Jornadas Zaragoza- Pau de Matemática Aplicada y Estadística.2001

[4] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas. *SOM Toolbox for Matlab 5*. http://www.cis.hut.fi/proyects/somtoolbox/ (2000).

[5] R. B. Calinski and J. Harabasz. *A dentrite method for culster analysis*. Communs statist., 3, 1-27(1974).

[6] W. J. Krzanowski and Y. T. Lai. *A criterion for determining the number of groups in a data set using sum of squres clustering*. Biometrics. 44, 23-34. 1985. (1985)

[7] J. Hartigan. *Clustering Algorithmics*. New- York Wiley. 1975.

[8] A. Cuevas, M. Febrero and R. Fraiman. *Estimating the number of cluster*. Can. J. Statist., 28, 367-382.(2000)

[9] M. A. Kraaijveld, J. Mao and A. K. Jain. *Porc.* 11^{th} *Int. Conf. on Pattern Recognition*. IEEE Comput. Soc. Press, Los alamitos, CA, p. 41. (1992).

[10] J. Vesanto and E. Alhoniemi. *Clustering of the Self-Organizing Map*. IEEE Transactions on Neural Netowrks, Vol. 11. Nž 3. (May 2000)

Lahoz D., San Miguel M.
Department: Métodos Estadísticos. Facultad de Ciencias
University: Universidad de Zaragoza
davidla@unizar.es and msm@unizar.es