5

Kohonen Maps Combined to K-means in a Two Level Strategy for Time Series ClusteringApplication to Meteorological and Electricity Load data

Khadir M. Tarek, Khdairia Sofiane and Benabbas Farouk
Laboratoire de Gestion Electronique de Documents (LabGED)
University Badji Mokhtar of Annaba, Algeria
khadir@labged.net, khdairia@labged.net, benabbas@labged.net

1. Introduction

Since the start of the computer era, a substantial amount of information and data are stored on numerical form. Automatic classification becomes therefore a very useful tool in order to reduce data dimension and extract maximum knowledge for such configurations (Jajuga *et al.*, 2002).

Data classification is a very important data analysis operation, consisting in regrouping objects of a similar data set into homogenous classes. Two main types of classifications exist: supervised and unsupervised classification. Supervised classification is based on a set of objects L of known classes, called training set, with the main goal being to identify candidate objects into their belonging classes. Where, unsupervised classification consists in partitioning a set of data D into sub-sets of similar attributes called classes or clusters (Halgamuge, 2005). Unsupervised classification is termed *clustering,* and will be so in the remaining of the chapter.

Many linear approaches such as Principal Component Analysis (PCA) (Jolliffe, 2002) and K-means were extensively used for the classification and clustering purposes, with an application to identification of meteorological scenarios in (Reljin *et al.,* 2003). Although PCA proved to be a very useful knowledge extraction technique it suffers from poor visualisation when dealing with complex structure representations of a data sample (Vesanto, 1999).

Nonlinear classification and clustering approaches stand as a strong alternative in order to treat the complexity and visualisation problem inherited from large multidimensional data sets. Self Organising Maps (SOM) or Kohonen maps qualify as a strong, leading, nonlinear approach. In the remaining of this chapter, they will be combined to K-means in order to solve the meteorological and electricity load day type clustering problems.

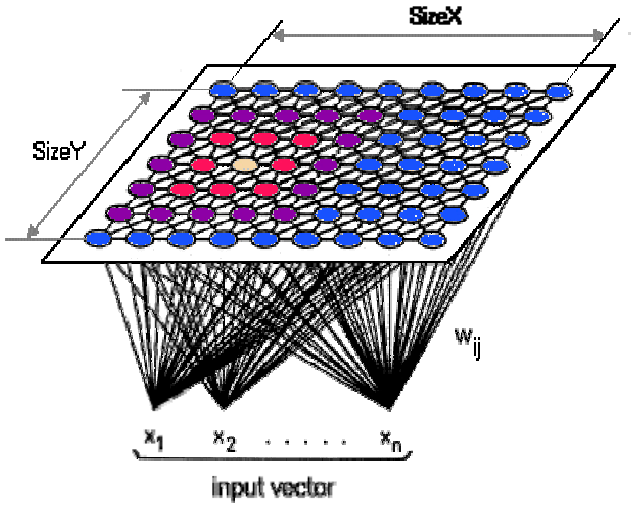


Fig. 1. Two dimensional Kohonen map

2. The Kohonen Self-Organizing Map

The Kohonen self-organizing map (SOM) is an unsupervised classification method, which transforms a set of complex data to one or two dimensional vectors with a simple geometric relationships, and preserving the most important initial data metrics during the display, i.e. the close dataset of the input space will have close representations in the output space and thus will be classified in the same cluster or nearby clusters (Kohonen, 1990, Dreyfus *et al.*, 2004). The self organizing map is suitable for data survey because it has prominent visualization properties; it is also a very effective tool for visualizing and exploring multidimensional data (Himberg, 2000; Vesanto, 1999). SOM has two layers, the input and the Kohonen or output layer, Figure 1.

The network consists in a grid of output nodes connected to the inputs via a set of weights. When presented with the k^{th} input vector P_k ∈R1xn, the network calculates the activation of each node using P_k as:

$$a_{i,j,k} = W_{i,j} P_k \quad (1)$$

where $a_{i,j,k}$ and $W_{i,j}$ are the activation of, and weight (∈ R1×n) connecting P_k to, node i, j respectively. P_k is said to be mapped onto the node with the highest activation. After several inputs have been presented, similar inputs are mapped to the same or adjacent nodes, i.e., within a small neighbourhood. A neighbourhood of size Nc around node i, j is defined as nodes $i \pm Nc$ to $j \pm Nc$. P_k for the current study is formed in two steps.

Each neuron of the topological layer is completely connected to the input layer neurons $Wi = (W_{1i} \ldots W_{ni})$, the weight vectors of these connections form the referent or prototype associated to each neuron, it has the same dimension as the input vectors. In each training step, one sample vector **x** from the input data set is chosen and a similarity measure is

calculated between it and all the weight vectors of the map. The Best-Matching Unit (BMU), is the unit whose weight vector has the greatest similarity with the input sample *P*. The similarity is usually defined by means of a distance measure; typically Euclidian distance. The use of neighbourhood concept introduces the topological constraints in the final SOM geometry.
The weights may or may be not, initialised randomly. In some cases they are initialised around the mean of the inputs as the inputs are all similar and thus restricted to a small portion of the space.
The neurons of competitive networks (Kohonen maps) learn to recognize groups of similar input vectors. Thus, the neuron whose weight vector is closer to the input vector is then updated to be even closer. The result is that the winning neuron is more likely to win the competition next time if similar vector is presented, and less likely to win when a very different input vector is presented. The training stage stops when any of the following conditions are met: the maximum number of epochs is reached, the performance has been minimized to the goal, or maximum amount of time has been exceeded.
During training the inputs are presented one by one and the weights of the triggered node (the node to which the inputs is mapped) and nodes in its neighbourhood are updated as in equation (2).

$$W_{i,j}(m+1) = W_{i,j}(m) + \alpha(m)\left[P_k - W_{i,j}(m)\right] \tag{2}$$

Where a is the adaptation gain, with $0 < a < 1$, and m is the iteration number. This has the effect of increasing the activation of the triggered node and its neighbours. In a single iteration all the inputs are presented and the weights adapted. After several iterations, the neighbourhood size is reduced by one and so on until zero, i.e., the triggered node only is adapted.
The SOM has proved his usefulness for multidimensional dataset clustering treating non-linear problems. The SOM is capable to extracting the statistical properties of time series.

3. The K-Means Clustering Algorithm

The K-means clustering algorithm is a most known vector quantization method; it groups in classes a set of points of the observations space without having any information of particular properties of these groups. Objects are classified as belonging to one of k groups, k chosen a priori. K-means quickly converges to a local minimum of its cost function (Bradley & Fayad, 1998; Kanungo *et al.,* 2002). The aim of K-Means clustering is the optimisation of an objective function that is described by the following equation:

$$l(w,x) = \sum_{x_i} \left\| z_i - w_{x(z_i)} \right\|^2 = \sum_{c} \sum_{x_i \in P_C \cap d} \left\| z_i - w_c \right\|^2 \tag{3}$$

where he expression:

$$I_c = \sum_{z_i \in P_C \bigcap A} \left\| z_i - w_c \right\|^2 \tag{4}$$

represents the local inertia, compared with the referent *Wc* of the training set observations A which are affected to this referent, these observations belong thus to the subset P_C. Inertia I_c is the quantization error obtained when deciding to replace P_C observations by the referent W_c which represents them. The quantity *I (W, X)*, which represents the sum of local inertia I_c is given by:

$$I(W, X) = \sum_c I_c \sum_{\substack{x_i \in A \\ X(x_i)=c}} \left\| z_i - w_c \right\|^2 \tag{5}$$

The K-means algorithm is iterative, where every iteration can be performed in two stages:
Assignment phase: This stage aims to minimize the function *I (W, X)* compared to the assignment function *X* (determining the set of referents).
Minimization phase: The second stage aims optimizing referents in order to representing the best observation points in *p* classes.
The main objective of applying K-means to time series analysis is to identify the different clusters representing the series situation using clustering methods. These methods must provide groups which members are close (have high similarity degree) and well separated. In the clustering process, there are no predefined classes and no examples that would show what kind of desirable relations should be valid among the data, so it is natural to be asked about the validity and quality of the results obtained. Two different sets of validity indices may be used for comparing the results when dealing with K-means: the internal and external criteria. The first indices category quantifies the match between a subjective partition and the idea that there is a good classification (Guérif, 2006) and the most properties commonly searched are compactness and group's separability. Different internal validity indices will be used in the rest of the chapter for meteorological and electricity load clustering, and are summarized in what follows:
***Davies-Bouldin index*:** takes into account both the compactness and the separability between groups, the value of this index is even lower than the clusters are compact and well separated (Davies & Bouldin, 1997). This index favoured hyperspheric groups and it is particularly well adapted for use with the K-means clustering algorithm (Guérif, 2006).
***Silhouette index*:** Kaufman and Rousseeuw (1990) suggest choosing the number of groups k> 2, which gives the greatest value of silhouette.
***Homogeneity and separation*:** homogeneity is calculated as the average distance between each input vector and the centre of the group to which it belongs. The separation is calculated as the average distance between the weighted groups centres (Chen *et al.,* 2002).
***The System Evolution (SE) method*:** Analyzes a dataset as a pseudo thermodynamics system, partition energy $E_p(k)$ denotes the border distance between two closest clusters (called twin-clusters) among the k clusters, while merging energy $E_m(k)$ denotes the average distance between elements in the border region (Wang *et al*., 2007).
***Weighted inter-intra index*:** proceeds with a forward searching and stops at the first mark to the bottom of the index, which indicates the optimal number of groups (Strehl, 2002).
For the external category the validity indices are used:

Rand index and Mirkin metrics. The rand index shows the proportion of pairs object where two partitions are concordant (Guérif, 2006), whereas the Mirkin metrics is defined as the number of edges that exists only in one of two partitions.
Hubert index. Higher values of this index show a large similarity between two groups (Halkidi *et al.,* 2001).

4. A two clustering level approach

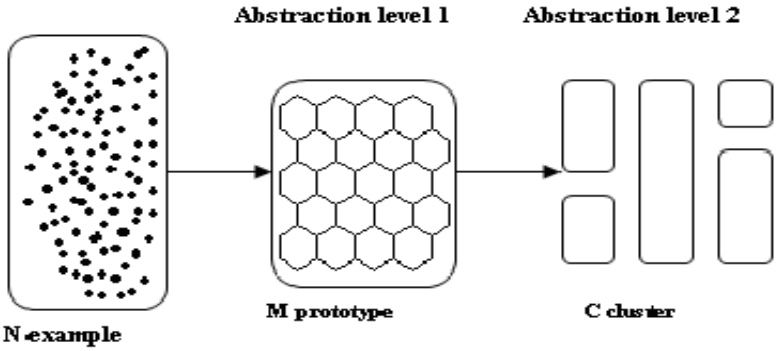


Fig. 2. First abstraction level is obtained by creating a set of prototypes vectors using the SOM. Clustering of the SOM creates the second abstraction level.

The number of prototype vectors resulting from SOM clustering is large especially when dealing with highly multidimensional time series applications. Only one classification level can then be revealing. A high level is interesting because it provides more detailed quality analysis and less compresses the dataset if we summarize all days by representatives of a small class's number (Rousset, 1999). It also can be very difficult to attribute some units of the input vector to a given cluster given by the map. The problem lies in the selection of some clusters border, where a clear distinction between two clusters is impossible. A second clustering stage becomes then useful to remove ambiguity and validate the SOM results.
The approach used in this chapter, is depicted in Fig. 2, the first abstraction level is achieved by creating a set of prototypes using SOM. These prototypes are then clustered in the second abstraction level using K-means clustering algorithm. It was noticed that clustering a large multidimensional time series data using only k-means generates a more computational time than the two-level clustering approach. Another advantage of this approach is the noise reduction (Vesanto & Alhoniemi, 2000), as the prototypes are local averages of the data and therefore less sensitive to random variations than the original data.

5. Application to Meteorological Parameters Clustering

It is extremely important to consider the effect of meteorological conditions on air pollution, because they directly influence the dispersion possibilities of the atmosphere. Severe pollution episodes in the urban environment are not usually attributed to sudden increases in the emission of pollutants, but to certain meteorological conditions which decrease the

capacity of the atmosphere to disperse pollutants (Ziomas et al., 1995). In meteorological studies it is very difficult to represent the data in statistically independent terms because the air pollution depends on all meteorological parameters (Kalkstein, 1991). One of the most interesting applications in the field of meteorological studies is the use of clustering methods in order to extract a representative set of prototypes (clusters) of the meteorological models in an area of interest. This technique has been successfully used in numerous studies such as (Eder et al., 1994). Principal component analysis (PCA) and K-means clustering algorithms have been used in (Reljin et al., 2003) to determine the synoptic weather scenarios. PCA is a powerful linear technique for data reduction (Kwan *et al.,* 2003), but suffers poor visualization as it is not well adapted to represent the datasets complex structure (Laitinen *et al.,* 2002). Recently, and as an alternative tool being used to deal with the complexity of multidimensional data, Kohonen self-organizing maps were used for clustering data in various ecosystems: forest, agriculture, etc. (Recknagel, 2002; Suwardi *et al*., 2007), water quality (Aguilera *et* al., 2001; Tison et al., 2005) and day type identification for electrical load (Khadir et al., 2006). Although the SOM has proved its efficiency in meteorological parameters clustering such as in (Hewitson & Crane, 2002; Cavasos, 2000; Turias et al., 2006), it is difficult to clearly identify the clusters and their borders when the map is very populated.

A two level clustering approach is proposed in this work in order to analyse and identify the meteorological day type for Annaba region in Algeria. In the first stage the SOM was used to reduce the set of prototypes which are then clustered using the K-means clustering algorithm in the second stage. This approach is more powerful than that of a direct clustering in data partitioning and computing time reduction. The correctness of clustering algorithm results is verified using quantitative validation based on two criterions categories (internal and external) and qualitative criteria, these cluster validity indices allowed us to respond to some frequently asked questions such as: "how many clusters are there in the dataset?", "does the resulting clustering scheme fits our data set?", "is there a better partitioning for our dataset?".

5.1. Area of Study and Used Data

Annaba region is located in the Eastern part of Algerian coast (600 km of Algiers), Fig. 3. The town is constituted of a vast plain bordered in the South and West, of a mountainous massive in North, and by the Mediterranean Sea in the East (Mebirouk & Mebirouk –Bendir, 2007). Its basin shaped topography, supports air stagnation and creation of temperature inversions. These situations allow the pollutants accumulation and the rise in concentration rates which results from it. Industry is the main factor causing air quality deterioration; this industrialization has allowed providing the needs of the country and population in iron and steel products, nitrate fertilizers, railway constructions and many other transformation industries. Controversially, it caused a disproportionate urbanization of the town with all its corollaries.

Fig. 3. Location of Annaba region

The dataset used in this study includes 04 meteorological parameters collected for 60 months (1995 to 1999) with a 3 hours expiry, therefore each row of the dataset (unspecified day) is characterized by 32 parameters during the 24 hours. The meteorological parameters which are obtained from the weather station of Annaba are dynamic and thermodynamic air descriptions: the pressure measured in tenth of millibars ; The temperature measured in tenth of °C; The moisture humidity in hundredths and the wind speed measured in nodes. A pre-treatment phase is needed to prepare the data, consisting in noise elimination, error corrections and data standardisation.

5.2. Results of the two Stages Clustering Approach

A. Results of the SOM Map

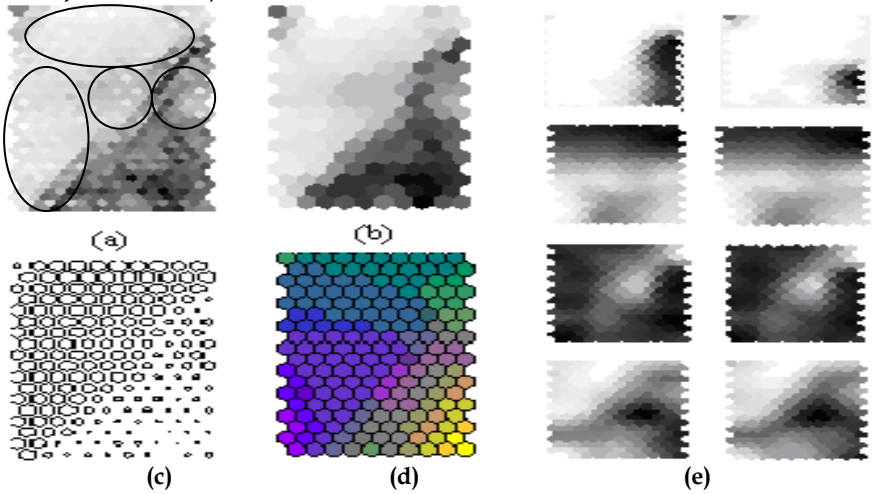


Fig. 4. The U-matrix map, (b and c) are average version of U-matrix, (d) the color coding map, (e) some component plane.

The results obtained from the Kohonen map using the specified dataset are shown in Fig. 4; the maps are connected to adjacent hexagonal nodes with sizes 18×10 by adapting the meteorological situations of the area. There are no explicit rule allowing the choice of Kohonen network node's number, but the principle is that the size should allow easier detection capabilities (Hautaniemi *et al.,* 2003). For this reason different experiments have been done to determine the optimal number of Kohonen units, by changing the number of nodes and checking the performance of each solution. Also different experiments have been done by changing the training parameters in order to determine the appropriate SOM for this dataset. Fig. 4 (a) provides a visualization of the U-matrix which represents a relative measurement of distance between the network coloured units, where the grey colour (shade) of the hexagon indicates the distance measure of the node to its adjacent. More the shade is dark more the distance is large; a cluster which represents similar data vectors can be seen as a clear zone with dark borders.

Fig. 4 (b) and 4 (c) present the average version of the U-matrix, for Fig. 4 (c) the size of each unit of the Kohonen maps is proportional to the average distance to its neighbours. It can be seen, for example, in Fig. 4 (a) that U-matrix provides clustering information of similar units which are presented with some circles on the map. However, the map of U-matrix indicates the situation where the distance measure was not reliable to determine the representative clusters. As reported by (Kiang *et al.,* 2003), it is difficult to visually SOM units when the network is strongly populated. In this case, the decision seems to be difficult and the use only of the Euclidean distance to select the meteorological clusters is not reliable.

To overcome the SOM deficiency in clustering data, a combination of the distance measure and the SOM colour-coding are used. The SOM colour-coding is a method for clustering data, according to their properties (Vesanto, 1999). As shown by Fig. 4 (d), the units which have similar parameters evaluate automatically similar colours of nodes on the grid. Greater distance measures of the network nodes are automatically assigned to different colours and clusters. To select a cluster, we first identify the clusters region based on the discoloration of units. In the situations when colours of nodes are not clear to indicate the differences of the clusters, the distance measures are then used to verify the clusters, Although, it was very difficult to attribute some units to a given group. The problem was the selection of some clusters border, it can be seen that a second clustering stage is useful to remove ambiguity and validate the SOM results.

B. Refining SOM Results by K-means

The K-means clustering algorithm has been applied to group the SOM units with different k-values (the number of clusters in which data are partitioned). Due to the inherent process randomness and because these methods depend on initial centres, the order of the presentation and the geometric properties of the data, a relatively high number of experiments (50 were ran in this study) has to be done and their results checked. The best partitioning for each (k) is selected using the error criteria described by equation (3), also the optimal number of clusters among different values of k is selected according to the validity indices described in section 2. The results of these indices are shown in Fig. 5, .6 and 7.

According to Davies-Bouldin index shown in Fig. 5, a negative peak is noticed at k=6 which indicate the optimal number of clusters proposed by this index. As well as the system evolution method where results values are presented in Fig. 6 indicating that the optimal partition of the dataset is obtained for k=6. The same result is proposed by silhouette and

inter-intra weighted indices shown in Fig. 7. According to the different indices values, the clusters obtained are well separated and homogeneous.

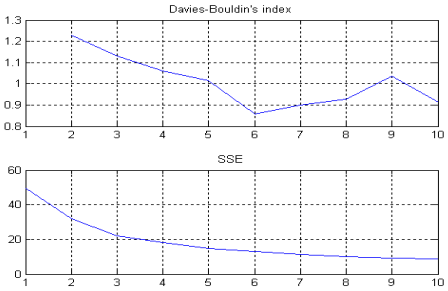


Fig. 5. Davies-Bouldin index and SSE

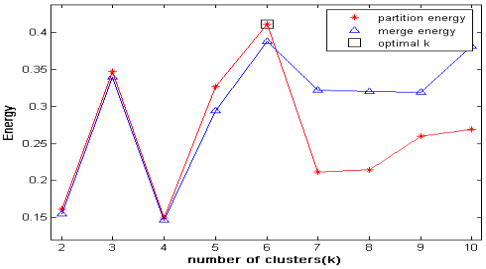


Fig. 6. System Evolution method results

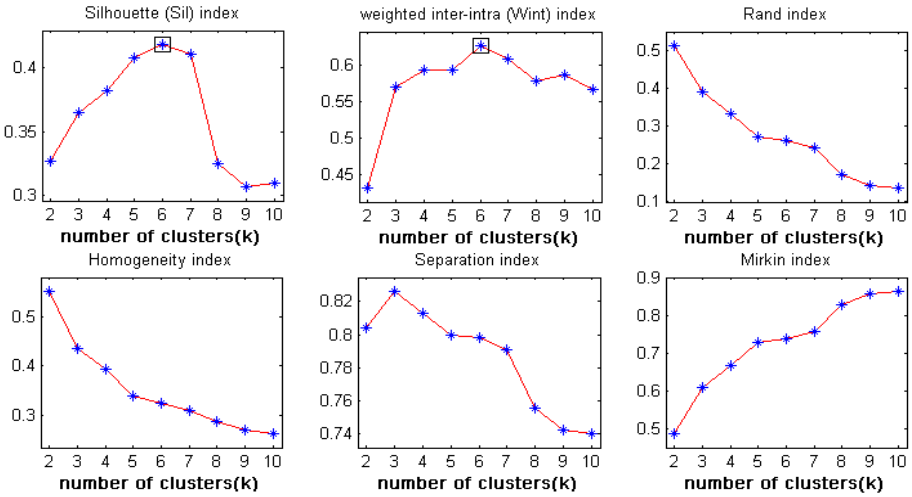


Fig. 7. Internal and external validity indices

The results of the two stage clustering procedure are shown in Fig. 6 and the average meteorological parameters for each cluster are shown in Fig. 9. The cluster C3 is characterized by a steady pressure throughout the 24 day hours, and high temperature which exceeds 25°C during the day and slightly lower in the night, this cluster is also characterized by a high pressure during the night which decrease in the day, the wind speed is very low in the night period and starts increasing during the day, according to the monthly distribution of clusters shown in Fig. 10 this cluster represents the warmer months. The sixth cluster is particularly concentrated in the winter and autumn months and is mainly characterized by a steady pressure and a high wind speed during the day.

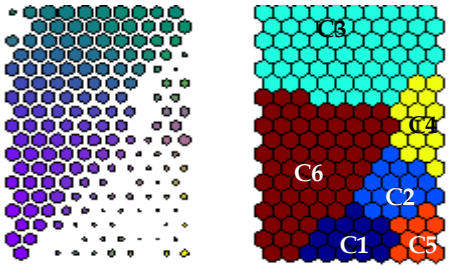


Fig. 8. Second stage clustering results

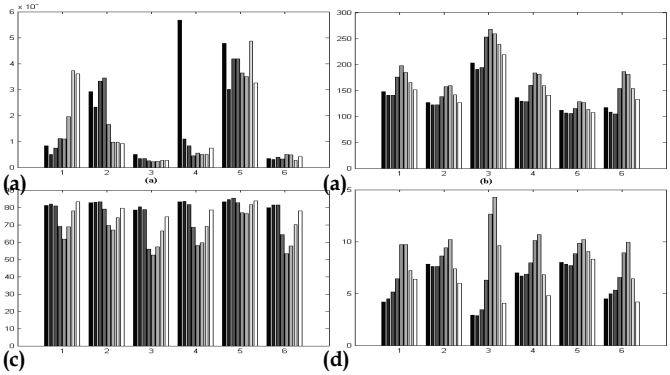


Fig. 9. (a) Pressure mean values, (b) temperature mean values, (c) humidity mean values, (d) wind speed mean values.

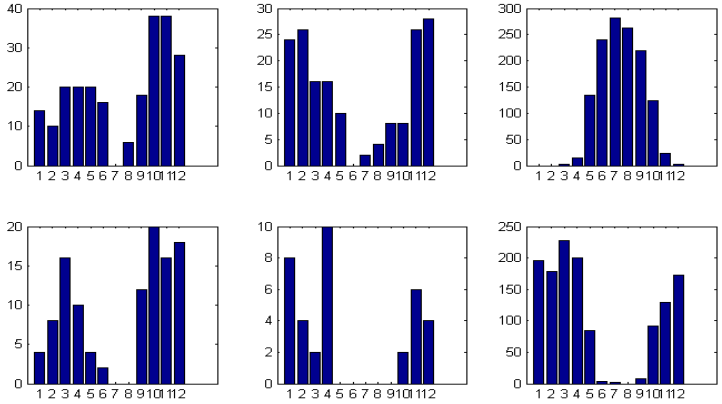


Fig. 10. Monthly distribution of clusters

The fourth cluster parameters are similar to the sixth with a high pressure in the night period and larger wind speed. The fifth cluster is characterized by a high pressure and humidity compared to the other clusters, temperature and wind speed are stable and low all the day hours. The first cluster is almost similar to the fourth with a low pressure at the

beginning of the day, however growing with time, and a lower wind speed values. The second cluster seems to be a sub-cluster of C4 with a small increase in pressure.

6. Day type Identification of Electricity load

Short term electricity load forecasting is nowadays, of paramount importance in order to estimate next day electricity load resulting in energy save and environment protection. Electricity demand is influenced (among other things) by the day of the week, the time of year and special periods and/or days, all of which must be identified prior to modeling. This identification, known as day-type identification, must be included in the modeling stage either by segmenting the data and modeling each day-type separately or by including the day-type as an input. It is proven that the day types or daily consumer's habits for different periods of time, such as working days, weekends, special holidays, etc affect heavily the load shape (Fay, 2004). Different prediction models may then be designed for each day type.

6.1. Overview of Algerian Electricity load

Electrical demand in Algeria from 01/01/2000 to 31/12/2004 is shown in Fig. 11. As can be seen there is an upward trend in the data reflecting increased economic activity over this period.

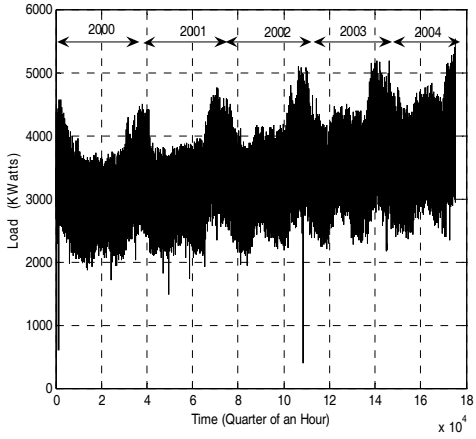


Fig.11. Algerian electricity load 2000-2004.

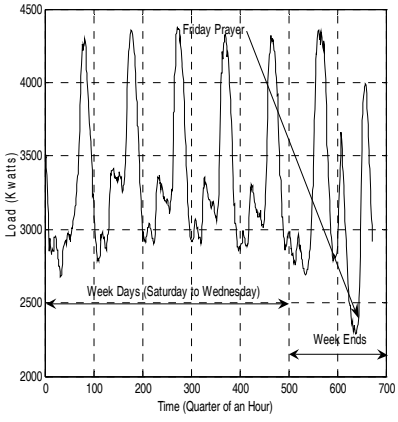


Fig. 12. Weekly load.

Daily load data can be disaggregated into distinct groups (called day-types) each of which has common characteristics. As can be seen in (Fig 12.) there is, for example, an obvious difference between the shapes of the load on a typical weekend day, such as Friday and a working day like Saturday or Sunday due to decreased economic activity and the weekly religious prayer on Friday. Note that in Algeria the weekend is on Thursdays and Fridays. Furthermore, there is a distinct difference between the shape of a typical winter day and summer day.

6.2. Day type identification using Kohonen Map

The existence of several different day-types has been shown by several researchers (Bretschneider *et al.*, 1999; Hsu and yang, 1991; Muller and Petrisch, 1998) However, the level of desegregation in day-type selection is, to a large extent, subjective and dependant on the judgment of the forecaster. As pointed out by (Hubele and Cheng 1990), the application of a separate load forecasting model for different seasons (for example summer, autumn, winter and spring) has the advantage that the models do not need to incorporate seasonal information.

Further desegregation of the load by day of the week (for example Summer Sunday, Winter Sunday, Summer Monday etc.) reduces further the amount of information that the model need incorporate. Such approaches have been implemented successfully by (Srinivasan *et al.*, 1999) and (Mastorocostas *et al.*, 1999), to mention but a few. Where a single model is used for all the data, the day-type information is often incorporated as an additional input (two examples are (Chen *et al.*, 1992) and (Lertpalangsunti and Chan, 1998). In either case the day-types must, however, be identified. The selection of day-types can be guided by analytical techniques. The self-organising feature map or Kohonen map (Kohonen, 1990) would appear ideal for day-type identification as the number and similarity between day-types is not known *a priori*. The Kohonen map can be implemented for day-type identification in several different ways (examples are (Fay and Ringwood, 2003; Hsu and yang, 1991; Muller and Petrisch, 1998) however differences in the results are insignificant in most cases thus the algorithm used by Hsu and Yang (Hsu and yang, 1991) was chosen.

For the present trials, the full years of data from 2003 and 2004 for the region of Algiers (north center and capital of Algeria) were used. The Kohonen map was trained using the following parameters, an initial neighborhood size of $Nc=1$, adaptation gain equal to 0.003, a total number of iteration m=10 and a grid size 18*18 (324) in total.

Initially, the daily load curve is extracted from each day to give a set of load curves that have a minimum value of zero and a maximum value of one (Hsu and yang, 1991).

$$Y'(i)_k = \frac{Y(i)_k - \min Y_k}{\max Y_k - \min Y_k} \quad i = 1,..,24 \tag{6}$$

where $Y'(i)_k$ and $Y(i)_k$ are the i^{th} elements (hour) of the load curve $Y'_k \in R^{1\times 24}$, and actual load $Y_k \in R^{1\times 24}$ of day k respectively. The load curves are then normalised to give them unity length:

$$P(i)_k = \frac{Y^i(i)_k}{\left(\sum_{j=1}^{24} {Y'_k}^2\right)^{1/2}} \quad i = 1,..,24 \tag{7}$$

where $P(i)_k$ is the i^{th} element of P_k. The weights are initialised as:

$$W_{i,j} = \left\| \left[\left(\mu_p(1)\right),, \left(\mu_p(24)\right) \right] + 5\mu \left[\left(\mu_p(1)\right),, \left(\mu_p(24)\right) \right] \right\| \tag{8}$$

where $\mu_p(1)$ and $\rho_p(1)$ are the sample mean and standard deviation of P(i) over all k, u is a uniformly distributed random number in the range -0.5 to 0.5 and Wi,j is normalised to unit length as in (Hsu and yang, 1991). Weight update is then done following Equation (9) repeated below for clarity:

$$W_{i,j}(m+1) = W_{i,j}(m) + \alpha(m)\left[P_k - W_{i,j}(m)\right] \quad (9)$$

Fig 13 shows the triggered nodes identified for the years stating from 2000 until 2004. We notice that the triggered nodes are located in the map (i between 0 and 17) and (j between 10 and 20).

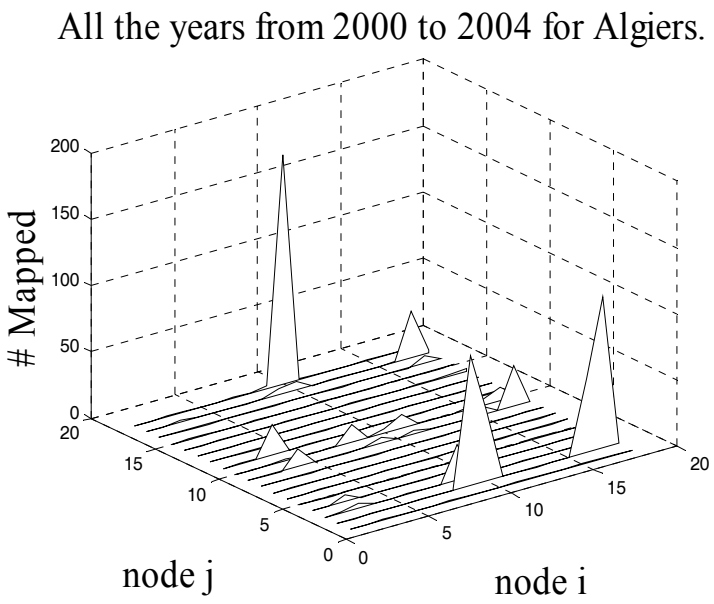


Fig. 13. Kohonen map results for Algiers load

It can be seen, Figure 14, that week days activate roughly the same map nodes where, the weekend activate different nodes for the Algiers load. This is true for Friday which is the weekly prayer occurring from 12 to 2:30. Thursday and Friday are the day of the weekend in Algeria.

Weekdays and weekends however appear differently on the map. The nodes that are triggered from Saturday to Wednesday occupy the same parts of the grid but Thursday and Friday (weekends) loads; trigger different parts of the grid showing the difference between these day types. The figure shows the difference between these days for Algiers load, where the disparate distribution of Fridays appears clearly, and is heavily dependant on seasonal effect.

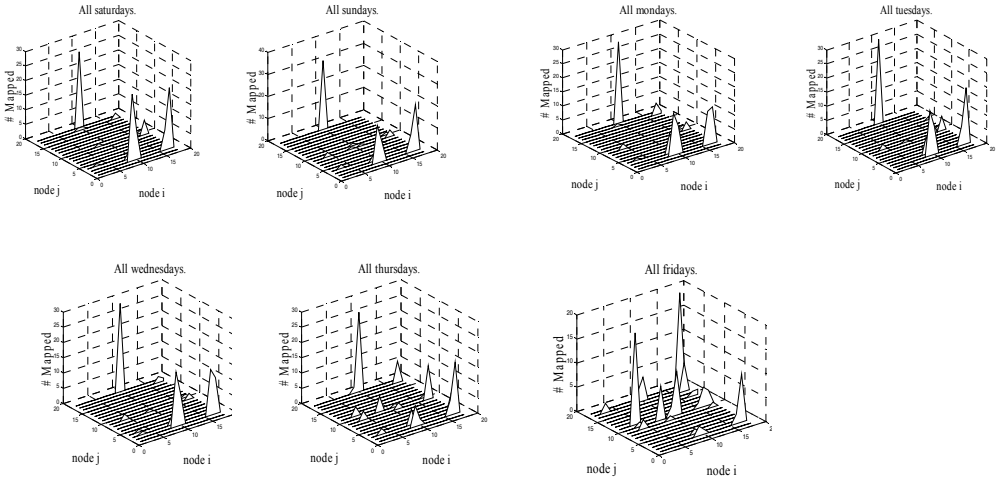


Fig. 14. Nodes triggered for working days(Saturday to Wednesday) and Week days (Thursday and Friday) loads fort he region of Algiers.

The seasonal effect is clearly shown for northern cities, Fig. 15 for Algiers where peaks appear along the longitudinal axe of the SOM with respect to monthly (seasonal) load. As for southern cities, minor seasonal effect is noticed. As can be seen in Fig. 9, the monthly SOM representation shows common peak for all months with a second peak appearing from May to august. The number of visually identified clusters may be numbered as 8 or 9 clusters.

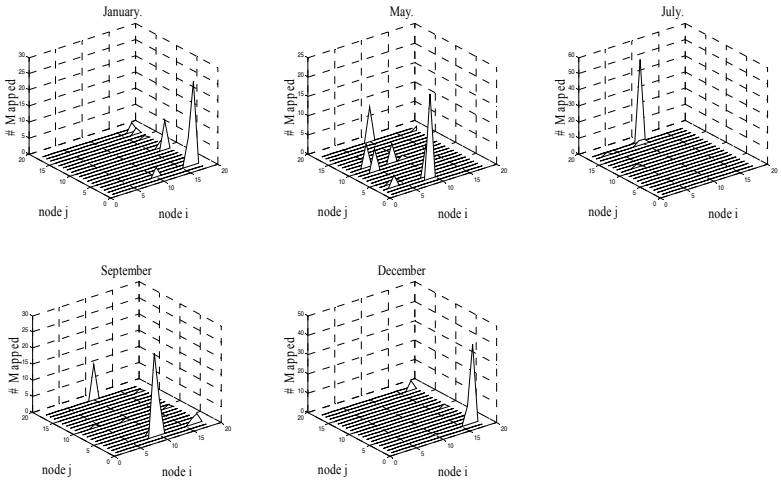


Fig. 15. Seasonal day-type identification for Algiers.

The K-mean algorithm is performed on the output obtained by Kohonen map in order to better define boundaries between clusters, and thus defining clearly the cluster number. The selection of the adequate cluster's number is accomplished using the Davies- Bouldin index defined earlier, and given in the following equations:

$$DB = \frac{1}{k}\sum_{i=1}^{k} R_i \; with \; R_i = max R_{ij} \;\; j = 1, \ldots n \; and \; i \neq j \tag{10}$$

k is the number of clusters and

$$R_{ij} = \frac{\left(s(C_i) + \; s(C_j)\right)}{\sigma(C_i \, , C_j)} \tag{11}$$

where $s : C \rightarrow R$ measures the scatter within a cluster and $\sigma : CxC \rightarrow R$ is a cluster to cluster distance measure.

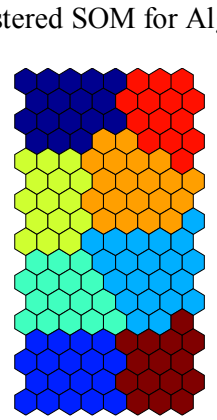


Fig. 16. Definite clusters identified for Algiers load

This clustering procedure aims to find internally compact spherical clusters which are widely separated.
As shown in Fig. 16, the number of clusters is found to be 8 with a value of *DB=0.8788*.

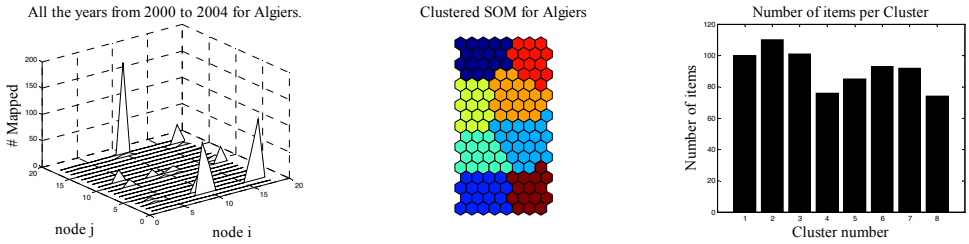


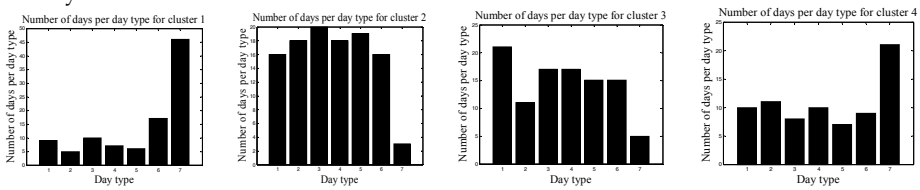
Fig 17. Kohonen map, clustered SOM with k-means and Number of items per cluster

Fig. 17 summarizes the Kohonen map, the clustered SOM and the number of items per cluster for Algiers's load. It can be also deducted that some clusters are dominant in terms of number of days, e.g., C3 .
Table 1 shows the weekly clusters distribution. Where, for example C1 contains a majority of Fridays. This includes that the cluster represent weekends at certain season.

Region	Algiers						
Day type	Sat	Sun	Mon	Tue	Wed	Thu	Frid
C1	9	5	10	7	6	17	46
C2	16	18	20	18	19	16	3
C3	21	11	17	17	15	15	5
C4	10	11	8	10	7	9	21
C5	13	15	12	15	15	12	3
C6	8	12	13	9	11	18	22
C7	18	19	13	14	17	9	2
C8	9	13	11	14	15	9	3

Table 1 Weekly distribution of cluster for Algiers load

Fig. 18 shows the weekly and monthly distribution of clusters. Detailed content of each cluster in terms of day types and number of days is also shown. For example Table 1, shows that C1 contains a majority of Fridays, which makes it a class containing weekends and bank holidays.



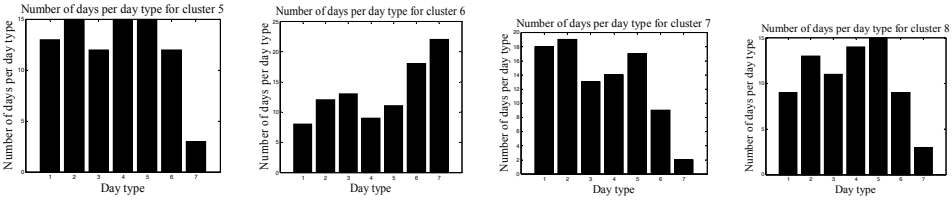


Fig 18. Weekly distribution of cluster

7. Conclusion

Time series analysis using Kohonen maps, allows a rough visual identification of the different existing classes. The K-Means algorithm comes as a complement for better class clustering and a clear frontiers definition when validated using different types of indices. Using a two stage clustering procedure seems to be more efficient than a direct clustering approach involving only SOM or K-means algorithms from an applicative and results view points. The obtained classification is also more compact as it merges neighbouring clusters into one.

Different clusters have been identified with clear borders definition for both day type identification along with a comprehensive analysis for constituents of each cluster in terms of size, day types and seasonal effects for meteorological and electricity load with, respectively, six and eight identified clusters. The results obtained may then be used to design prediction multi-model systems according to the number and the nature of each cluster (data type). Such approach may be more advantageous and can improve the results of a unique global predictor or classifier.

Acknowledgments

This work was supported by the TASSILI program n° 07 MDU 714.

8. References

Aguilera, P.A., Frenich, A.G., Torres, J.A., Castro, H., Vidal, J.L.M. and Canton, M. (2001). Application of the Kohonen neural network in coastal water management: methodological development for the assessment and prediction of water quality. *Water Res. 35, 4053–4062.*

Annas, S., Kanai T. and Koyama, S. (2007). Principal Component Analysis And Self Organizing Map For Visualizing And Classifying Fire Risks In Forest regions, *Agricultural information research, 16(2), 44-51.*

Bradley, P. and Fayyad, U. (1998). Refining initial points for K-means clustering. *International Conference on Machine Learning (ICML-98*), 91-99.

Bretschneider, P., Rauschenbach, T., and Wernstedt, J., (1999). Forecast using an adaptive fuzzy classification algorithm for load," *6th European Congress on Intelligent Techniques and Soft Computing, 3, 1916-1919.*

Cavazos, T. (2000). Using self-organizing maps to investigate extreme climate events: an application to wintertime precipitation in the balkans, *Journal of climate, 13, 1718–1732.*

Chen, G., Jaradat, S.A. and Banerjee, N. (2002). Evaluation and comparison of clustering algorithms in anglyzing es cell gene expression data, *Statistica Sinica* 12, 241-262.

Chen, S.T. Yu, D.C. Moghaddamjo, A.R. (1992). Weather sensitive short-term load forecasting using non-fully connected artificial neural network, *IEEE Transactions on Power Systems, 7 (3), 1098-1104.*

Davies, D.L. and Bouldin, D.W. (1997). A Cluster Separation Measure. *IEEE Transactions On Pattern Analysis And Machine Intelegence, PAMI 1(2):224-227.*

Dreyfus G. , Martinez, J.M., Samuelides, M., Gordon, M.B., Badran, F., Thiria, S. and Hérault, L. (2004). *Reseaux De Neurones : Méthodologie Et Application*, Eyrolles ISBN : 2-212-11464-8, France.

Eder, B.K., Davis, J.M. and Bloomfield, P. (1994). An automated classification scheme designed to better elucidate the dependence of on meteorology. *Journal of Applied Meteorology, 33, 1182–1199.*

Fay, D., Ringwood, J.V., Condon, M., and Kelly, M. (2003). 24-hour electrical load data -a sequential or partitioned time" series? Journal of Neurocomputing, 55(3-4), 469-498.

Fay, D. (2004). *A strategy for short-term load forecasting in Ireland*, Ph.D Thesis, Dept. of Electronic Engineering, Dublin City University, Ireland.

Guérif, S. (2006). *Réduction de dimension en Apprentissage Numérique non supervisé,* Ph.D thesis, Université Paris 13, France.

HALGAMUGE, S.K. (2005*). CLASSIFICATION AND CLUSTERING FOR KNOWLEDGE DISCOVERY (STUDIES IN COMPUTATIONAL INTELLIGENCE),* WANG, L. (EDS), SPRINGER, 3540260730, THE NEDERLANDS.

Halkidi, M., Batistakis, Y. and Vazirgiannis. M. (2001). On Clustering Validation Techniques. *Intelligent Information Systems Journal, 17(2-3): 107-145.*

Hautaniemi, S. Yli-Harja, O., Astola, J., Kauraniemi, P., Kallioniemi, A., Wolf, M., Ruiz, J,. Mousses S. and kallioniemi, O. (2003). Analysis and visualisation of gene expression microarray data in human cancer using self-organizing maps. *Machine Learning, 52, 45-66.*

Hewitson, B. C. and Crane, R. G. (2002). Self-organizing maps: applications to synoptic climatology, *Climate Research, 22, 13–26.*

Himberg, A. (2000). SOM Based Cluster Visualization and Its Application for False Coloring, *Proceedings of International Joint Conference on Neural Networks (IJCNN), 3, 587-592.*

Hsu, Y.Y. and Yang, C.C. (1991). Design of artificial neural networks for short-term load forecasting Part I: Self-organising feature maps for day type identification, *IEEE Proceedings-C, 138(5), pp 407-413.*

Hubele, N.F. and Cheng C.S., (1990). Identification of seasonal short-term forecasting models using statistical decision functions, *IEEE Transactions on Power Systems, 5 (1), 40-45.*

Jajuga, K., Sokolowski, A and Bock, H.H, (2002). *Classification, Clustering, and Data Analysis: Recent Advances and Applications (Studies in Classification, Data Analysis, and Knowledge Organization)*, Springer, ISBN : 354043691X, Berlin.

JOLLIFFE, I.T., (2002). *PRINCIPAL COMPONENT ANALYSIS, SPRINGER*, 0387954422, NEW YORK.

Kalkstein, L.S. (1991). A new approach to evaluate the impact of climate on human mortality. *Environmental Health Perspectives 96, 145–150.*

Kanungo, T., Mount, D. M., Netanyahu, N., Piatko, C., Silverman, R. and Wu, A. (2002). An efficient k-means clustering algorithm: analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7), 881-892*.

Kaufman L. and Rousseeuw P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley.

Khadir, M.T., Fay, D. and Boughrira, A. (2006). Day Type Identification for Algerian Electricity Load using Kohonen Maps, *Transaction on engineering, computing and technology 15, 296-300.*

Kiang, M.Y., Hu, M.Y. and Fisher, D.M. (2004). An extended self organizing map network for market segmentation a telecommunication example. *Decision Support Systems, DECSUP-11061-12.*

Kohonen , T. (1990). The self-organising map, *Proceedings IEEE, 78 (9). 1464-1480.*

Laitinen, N., Ranatanen, J., Laine, S., Antikainen, O., Rasanen, E., Airaksinen, S. and Yliruusi, J. (2002). Visualization of particle size and shape distributions using self-organizing maps. *Chemometrics and intelligent laboratory systems, 62, 47-60.*

Lertpalangsunti N., and Chan C.W., (1998). An architectural framework for the construction of hybrid intelligent forecasting systems: application for electricity demand prediction, *Engineering Applications of Artificial Intelligence, 11, pp 549-565.*

Mastorocotas P.A., Theocharis, J.B. and Bakirtzis, A.G. (1999). Fuzzy modelling for short term load forecasting using the orthogonal least squares method, *IEEE Transactions on Power Systems, 14 (1), 29-35.*

Mebirouk H., and Mebirouk-Bendir F. (2007). Principaux acteurs de la pollution dans l'agglomération de annaba. Effets et développements, *Colloque International sur l'Eau et l'Environnemen*t, Alger.

Muller, H. Petrisch, G. (1998*). Energy and load forecasting by fuzzy neural networks.* In: H. Jurgen, H.J. Zimmermann eds., Proceedings, European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, September. Aachen: Elite foundation, 1925-1929.

Recknagel, F. (Ed.), (2002). *Ecological informatics: understanding ecology by biologically-inspired computation*. Springer, Berlin.

Reljin, I., Reljin, B. and Jovanovi, G. (2003). Clustering and mapping spatial-temporal datasets using som neural networks , *Journal Of Automatic Control, 13(1), 55-60.*

Rousset, P. (1999). *Applications des algorithmes d'auto-organisation à la classification et à la prévision.* Ph.D thesis, University of Paris I, France.

Srinivasan D., Tan S.S., and Chang, E.K. (1999). Parallel neural network-fuzzy expert system for short-term load forecasting: system implementation and performance evaluation, *IEEE Transactions on Power Systems, 14(3), 1100-1106.*

Strehl. A. (2002). *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D thesis, The University of Texas at Austin.

Tison, J., Park, Y.-S., Coste, J., Wasson, M..G., Ector , L., Rimet, F. and Delmas, F. (2005). Typology of diatom communities and the influence of hydro-ecoregions: A study on the French hydro system scale, *Water Research 39- 3177–3188.*

Turias, I.J., Gonzalez, F.J., Martin, M.L. and Galindo, P.L. (2006). A competitive neural network approach for meteorological situation clustering, *Atmospheric Environment, 40, 532–541.*

Vesanto, J. (1999). SOM-Based Data Visualization Methods, *Intelligent Data Analysis, 3(2), 111-126.*

Vesanto J. and Alhoniemi , E. (2000). Clustering of the Self-Organizing Map, *IEEE Transactions on Neural Networks, special issue on data mining, 11(3), 586−600.*

Wang, K., Zhang, J., Zhang H. and Guo, T. (2007). *Estimating the Number of Clusters via System Evolution for Cluster Analysis of Gene Expression Data*, Technical report. Xidian University, P. R. China.

Xu, k.C.R. andHaynes, L. (2001). A new data clustering and its applications .*Proceeding of SPIE-the international society for optical engineering, 4384, 1-5.*

Ziomas, I.C., Melas, D., Zerefos, C.S. and Bais, A.F. (1995). Forecasting peak pollutant levels from meteorological variables. *Atmospheric Environment 29, 3703–3711.*

Self-Organizing Maps

Edited by George K Matsopoulos

ISBN 978-953-307-074-2
Hard cover, 430 pages
Publisher InTech
Published online 01, April, 2010
Published in print edition April, 2010

The Self-Organizing Map (SOM) is a neural network algorithm, which uses a competitive learning technique to train itself in an unsupervised manner. SOMs are different from other artificial neural networks in the sense that they use a neighborhood function to preserve the topological properties of the input space and they have been used to create an ordered representation of multi-dimensional data which simplifies complexity and reveals meaningful relationships. Prof. T. Kohonen in the early 1980s first established the relevant theory and explored possible applications of SOMs. Since then, a number of theoretical and practical applications of SOMs have been reported including clustering, prediction, data representation, classification, visualization, etc. This book was prompted by the desire to bring together some of the more recent theoretical and practical developments on SOMs and to provide the background for future developments in promising directions. The book comprises of 25 Chapters which can be categorized into three broad areas: methodology, visualization and practical applications.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821