Load Pattern-Based Classification of Electricity Customers

Gianfranco Chicco, *Member, IEEE*, Roberto Napoli, *Member, IEEE*, Federico Piglione, Petru Postolache, *Member, IEEE*, Mircea Scutariu, *Member, IEEE*, and Cornel Toader, *Member, IEEE*

***Abstract*—Accurate knowledge of the customers' consumption patterns represents a worthwhile asset for electricity providers in competitive electricity markets. Various approaches can be used for grouping customers that exhibit similar electrical behavior into customer classes. In this paper, we focus on two approaches for customer classification—a modified follow-the-leader algorithm and the self-organizing maps. We include an overview of basic theory for these methods and discuss the performance of the customer classification on the real case of a set of customers supplied by a distribution company. We compare the results obtained from the two approaches by means of two suitably defined adequacy indicators and discuss the potential applications of the surveyed approaches.**

***Index Terms*—Clustering algorithms, customer classes, load patterns, self-organizing maps.**

I. Introduction

With the advent of competitive electricity markets, the distribution side of the electricity industry has to face new challenges in providing satisfactory service to customers. At the top of these challenges, there is the constant pressure for continuously decreasing the distribution service costs, which eventually reflects on the satisfaction of the supplied customers. At the same time, the distributor needs to provide these services with a fair revenue, in order to cover its distribution costs. The electricity markets operation imposes increased competition on distribution and supply companies, tightening the margins available for these companies to score profits.

In the market scenario, electricity providers have been given new degrees of freedom in defining tariff structures and rates under regulatory-imposed revenue caps. This requires a suitable grouping of the electricity customers into customer classes. Various criteria for customer partitioning are in place in several countries, most of which are based on rated electrical values and activity-type parameters. However, these criteria typically are poorly correlated to the actual electrical behavior of the customers [1]. Taking the customers' electrical behavior into account is a key factor for setting up new tariff offers, leading to tariff structures more closely related to the actual costs of electricity provision in different time periods.

Enhancing the knowledge on the way customers demand electricity calls for extended monitoring of the customers' consumption and the adoption of suitable discrimination techniques able to isolate customer subsets which exhibit sufficiently similar electrical behavior. Several options well documented in the literature rely upon methods extracted from a wide range of fields, ranging from data mining-related techniques [2] to neural network-based approaches [3].

Although the researchers' attention has been attracted by the load pattern issues for quite some time, seemingly less effort has been put in the analysis of available tools to produce a fair, sound classification system able to provide distinct and nonoverlapping customer classes. Results of a classification algorithm for the British electricity market in a confined area are discussed in [4]. Other efforts were aimed at investigating the means available to characterize the load patterns for customer classification [5], [6]. The classification process used in [1] and [5] is based on a suitable automatic clustering procedure [7], slightly rearranged to embed a measure of the diversity of discriminating features in the focused population.

In the approach reported in [6], the emphasis went on a newly emerged technique for load pattern characterization. The technique is based on wavelet packet transform applied to the load pattern, which can be easily assumed as a signal, producing a set of characteristic wavelet coefficients. These coefficients allow for load pattern classification, provided Euclidean-based distances between them are computed. A recent different approach [8] adopts hierarchical clustering, consisting of a three-step procedure centered on grouping the normalized measured load patterns into a hierarchical cluster tree. A method based on nonlinear load research estimate approach is presented in [9] for the estimation of substation peak loads by means of monthly load shapes of defined customer classes. The load analysis approach used in [3] is aimed at employing statistical and neural-based methods to perform load pattern clustering.

The focus of this paper is on adequacy assessment of customer classification tools in terms of their accuracy for the classification process. We surveyed the consistency of two important classes of tools—unsupervised clustering algorithms and self-organizing maps (SOM), presented in Section II. Two indicators of clustering adequacy, proposed by the authors in [1] and briefly overviewed in Section III, are used for comparing the results obtained from the selected classification tools. Results of an extensive case study performed on a large set of customers supplied by the Romanian electricity distribution company are

Manuscript received July 4, 2003.

G. Chicco, R. Napoli, and F. Piglione are with the Dipartimento di Ingegneria Elettrica Industriale, Politecnico di Torino, Torino I-10129, Italy (e-mail: gianfranco.chicco@polito.it; roberto.napoli@polito.it; federico.piglione@polito.it).

P. Postolache and C. Toader are with the University Politehnica of Bucharest, Power Engineering Faculty, Bucharest RO-79590, Romania (e-mail: petru.postolache@k.ro).

M. Scutariu is with ELECTRICA Muntenia Sud Distribution and Supply Company, Bucharest RO-792091, Romania (e-mail: mscut@sdb.ro).

Digital Object Identifier 10.1109/TPWRS.2004.826810

presented in Section IV. The clustering adequacy assessment is addressed in Section V. Section VI contains the concluding remarks.

II. Classification Tools and Models

The *load pattern* associated with any customer contains the information needed to judge the affiliation of the customer to a defined group. Load patterns need to be extracted through field measurements and represented by using a similar scale for the purpose of their comparison. Here, one possibility is to rescale the load patterns with respect to an energy-related parameter (e.g., the average power), or to resort to a normalization in the range (0,1) by using as normalizing factor the peak value of the pattern over the time interval of definition. We choose the latter solution, since it allows for introducing a suitable weighting factor which enhances the performance of the clustering algorithm used, as shown in Section II-A. The peak power is assumed as *reference power* associated to each normalized load pattern.

We consider a population of M customers, each customer being represented by a load pattern consisting of a group of H data $\mathbf{l} = \{l_h, h = 1, \ldots, H\}$. We denote by $\boldsymbol{L} = \{\mathbf{l}^{(m)}, m = 1, \ldots, M\}$ the set of load patterns associated to the M customers. The customer classification is performed in two successive phases. In the first phase, a clustering process is used to form K clusters, in such a way to operate the partitioning of the initial $\boldsymbol{L}$ load patterns into the subsets $\boldsymbol{L}^{(k)} \subset \boldsymbol{L}$, for $k = 1, \ldots, K$, each of them containing $n^{(k)}$ load patterns. The customer classes are formed in the successive phase. Depending on the method used, the number of customer classes may be the same as the number of clusters formed, or a postprocessing of the clustering results may be required to form the final number of customer classes. Both cases are examined and discussed in this paper. In the following comments, we refer to K customer classes.

Once the customer classes have been formed, the representative load pattern $\mathbf{r}^{(k)}$ of each customer class is derived from the load patterns in $\boldsymbol{L}^{(k)}$. This is accomplished by computing the weighted average of the $n^{(k)}$ load patterns, using the reference power of each pattern as weight. The load pattern obtained can eventually be normalized to its peak value. The resulting set $\boldsymbol{R} = \{\mathbf{r}^{(k)}, k = 1, \ldots, K\}$ contains the *class representative load profiles*.

The classification process can be performed by using various clustering tools. We present and compare a modified follow-the-leader unsupervised clustering algorithm and the SOM. A brief overview of these methods is presented in the sequel.

A. Modified Follow-The-Leader Algorithm

The clustering procedure we used to group the customers on the basis of their load patterns is based on the original follow-the-leader algorithm proposed in [7], which does not require initialization of the number of clusters and computes the cluster centers automatically. The authors have modified the Euclidean metric used in the original algorithm by introducing for each distinctive feature a weighting factor $\sigma_h^2/\overline{\sigma}^2$, where σ_h^2 is the variance of the hth feature of all the load patterns in the population and $\overline{\sigma}^2$ is the average value of the variance σ_h^2 for $h = 1, \ldots, H$. If all load patterns have been normalized to the range (0,1), the impact of high-variance features is amplified in the computation of the weighted Euclidean distance

$$d\left(\mathbf{l}^{(m)}, \mathbf{q}_i^{(k)}\right) = \sqrt{\sum_{h=1}^{H} \frac{\sigma_h^2}{\overline{\sigma}^2} \left(l_h^{(m)} - q_{ih}^{(k)}\right)^2} \qquad (1)$$

where $\mathbf{q}_i^{(k)}$ designates the kth cluster center for the ith cycle of the clustering process $q_{ih}^{(k)}$ and its hth component.

The clustering procedure (Fig. 1) consists of several cycles $i = 1, \ldots, I$. The first cycle $(i = 1)$ creates K clusters with $n^{(k)}$ load patterns belonging to each cluster $k = 1, \ldots, K$. The number of clusters created depends on the user-defined distance threshold ρ. All load patterns (subjects) are sequentially processed. Once a load pattern is presented, if the distance (1) computed with respect to all of the existing cluster centers exceeds the distance threshold, a new cluster is formed. Otherwise, the load pattern is assigned to the cluster to which center the distance is minimum and the cluster center is correspondingly updated. The successive cycles $(i > 1)$ may change the cluster composition, by reassigning the patterns to the closest cluster and updating the cluster centers until the number of patterns changing clusters in a single cycle is null. As a last-resource stop criterion, the maximum number of cycles I should be large enough to allow for clustering stabilization before the Ith cycle.

B. SOM Approach

The Kohonen SOM [10] is an unsupervised neural network that projects a H-dimensional data set into a reduced dimension space (usually a monodimensional or bidimensional map, Fig. 2). The projection is topological preserving, that is the proximity among objects in the input space is approximately preserved in the output space.

The SOM is composed of a predefined grid containing $N_1 \times N_2$ H-dimensional units $\mathbf{c}_\mathrm{k}$ that form a competitive layer. In the traditional competitive learning, only one unit responds at the presentation of the generic input sample $\mathbf{x}_\mathrm{i}$. Therefore, the activation function is an inverse function of $\|\mathbf{x}_\mathrm{i} - \mathbf{c}_\mathrm{k}\|$, so that the unit closest to $\mathbf{x}_\mathrm{i}$ wins the competition and updates itself moving its weight vector toward the sample $\mathbf{x}_\mathrm{i}$ of a certain amount.

This procedure simply performs a k-means clustering in the reduced dimension space of the map and cannot preserve the topology of the H-dimensional input space (clusters that are distant in the original space could become close in the reduced dimension space). Kohonen, referring to biological considerations such as the *tonotopic map* of the brain, proposed a new competitive learning algorithm that updates not only the weights of the winning unit, but also the weights of its neighbor units in inverse proportion of their distance. Moreover, the neighborhood size of each unit shrinks progressively during the training process, starting with nearly the whole map and ending with the single unit. In this way, the map auto-organizes so that units spatially

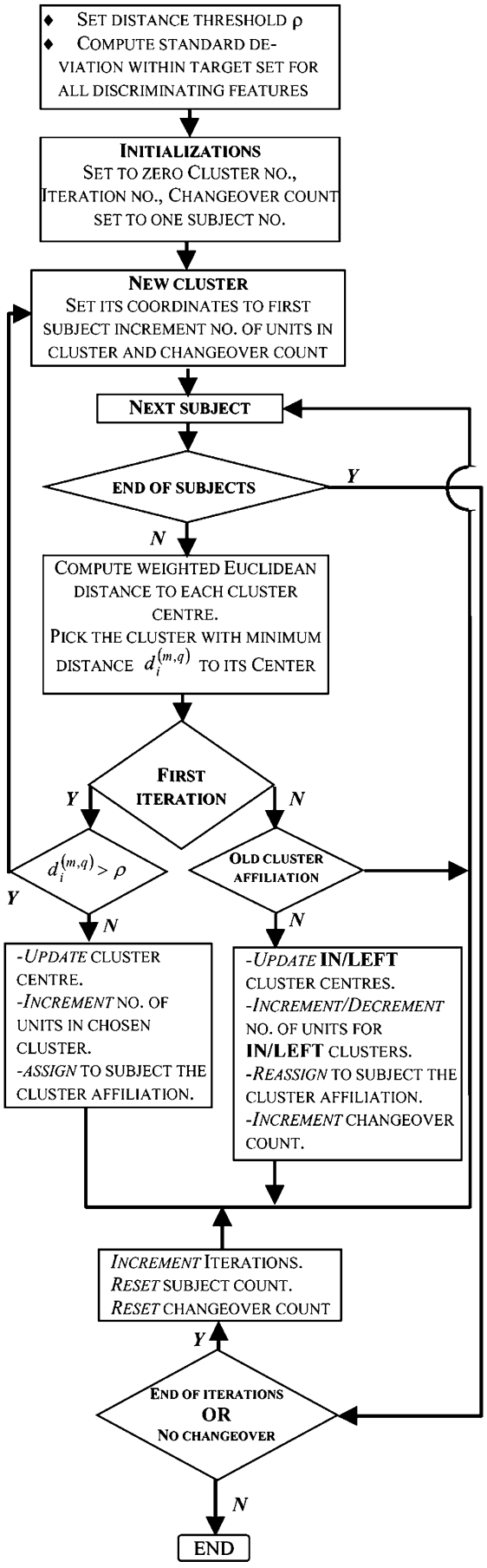

Fig. 1. Conceptual flowchart of the clustering procedure.

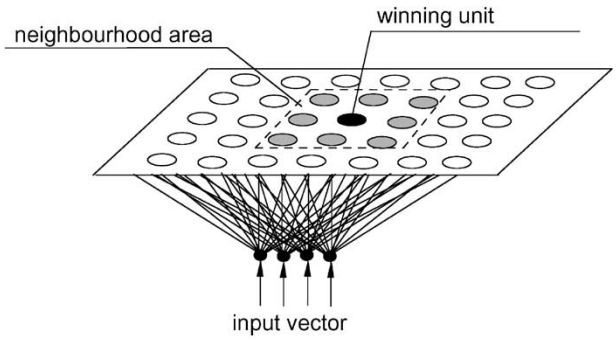

Fig. 2. Structure of the Kohonen SOM.

close correspond to similar patterns in the original space. Therefore, the presentation of each sample updates the generic unit $\mathbf{c}_k$ according to

$$\mathbf{c}_k^{(\text{new})} = \mathbf{c}_k^{(\text{old})} + \eta(t)\lambda_{kw}(t)\left(\mathbf{x}_i - \mathbf{c}_k^{(\text{old})}\right) \tag{2}$$

where $\eta(t)$ is the *learning rate* and $\lambda_{kw}(t)$ is the value of the *neighborhood function* referred to the generic unit k, being w the identifier of the winning unit.

The result of this peculiar training procedure is that all of the SOM units become apt to classify samples belonging to the data set space. The units that classify the training samples form receptive areas, named *activity bubbles*, which follow the probability distribution of the training data set. However, if the map is large enough, the activity bubbles are encircled by other units that never win the competition during the training session. These are named *dead units* and are trained to form connection zones among the activity bubbles and could win for samples not included in the training data set. Therefore, the SOM projection behaves as an "elastic" surface, which adapts spontaneously to the whole data set space. The aim of the SOM is then to perform an understandable projection of the original data set into a reduced dimension space, so that clusters could be directly perceived by inspection. Obviously, the effective cluster formation is not immediate, but requires a further selection work that could be accomplished directly by visual inspection of the map or by means of some automatic clustering procedure, as discussed in Section IV-B.

III. Classification Adequacy Assessment

A. General Outline and Definition of the Distances

In order to assess the consistency of the load patterns inside the resultant clusters/groups, a measure of adequacy of the classification process should be adopted. For this purpose, some indexes have been proposed in [1], able to grasp the load pattern diversity in the clusters obtained. Distances are seen in terms of multidimensional vectors. Therefore, the following definitions assist the formulation of adequacy measures:

1) the distance between two load patterns (e.g., between two members $\mathbf{l}^{(i)}$, $\mathbf{l}^{(j)}$ of the set $\boldsymbol{L}$)

$$d\left(\mathbf{l}^{(i)}, \mathbf{l}^{(j)}\right) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}\left(l_h^{(i)} - l_h^{(j)}\right)^2} \tag{3}$$

2) the distance between a representative load curve $\mathbf{r}^{(k)}$ and the subset $\boldsymbol{L}^{(k)}$, defined as the geometric mean of the distances between $\mathbf{r}^{(k)}$ and each member of $\boldsymbol{L}^{(k)}$

$$d\left(\mathbf{r}^{(k)}, \boldsymbol{L}^{(k)}\right) = \sqrt{\frac{1}{n^{(k)}} \sum_{m=1}^{n^{(k)}} d^2\left(\mathbf{r}^{(k)}, \mathbf{l}^{(m)}\right)}. \quad (4)$$

B. Adequacy Measures

We assume the classification tool which exhibits good adequacy ensures the determination of classes well separated (distinct class representative load profiles) and compact (load patterns of the customers included in each class very similar to their representative ones). Two hypothesis are needed for enabling comparisons among the classification results: each member of the target set to be classified must be represented and supplied to any classification tool in the same format, and the number of customer classes formed must be the same. We assume the classification procedure based on any of the above techniques results into K customer classes and in the subsets $\boldsymbol{L}^{(k)}$ containing the load pattern partition for $k = 1, \ldots, K$.

We use the two adequacy measures proposed in [1]

a) the mean index adequacy (MIA)

$$MIA = \sqrt{\frac{1}{K} \sum_{k=1}^{K} d^2\left(\mathbf{r}^{(k)}, \boldsymbol{L}^{(k)}\right)} \quad (5)$$

b) the clustering dispersion indicator (CDI), depending on the distance between the load patterns in the same cluster and (inversely) on the distance between the class representative load profiles

$$CDI = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^{K} \left[\frac{1}{2 \cdot n^{(k)}} \sum_{n=1}^{n^{(k)}} d^2\left(\mathbf{l}^{(n)}, \boldsymbol{L}^{(k)}\right)\right]}}{\sqrt{\frac{1}{2 \cdot K} \sum_{k=1}^{K} d^2\left(\mathbf{r}^{(k)}, \boldsymbol{R}\right)}}. \quad (6)$$

The classification tool which produces, for a given number of resulting classes, the smaller MIA or CDI values prevails over the others in terms of adequacy.

IV. Application of the Classification Techniques

The performance of the modified "follow-the-leader" clustering and of the SOM for electricity customer classification has been investigated on a target customer set extracted out of the customers of the Romanian national electricity distribution company, Electrica. A measurement campaign that encompassed 234 customers spread all over the country was performed to collect the load patterns over three-week time intervals. All surveyed customers pertain to industrial, services, and small-business activity types. The measured load patterns refer to weekdays. The information has been processed in order to obtain the normalized representative load pattern for each customer member of the target set, which ensures a fair basis for comparison. The customer representative load patterns and the corresponding reference powers are the inputs of the classification tools described in Section II-A and Section II-B. In order to ensure a comparative assessment of the two surveyed classification tools, we employed the adequacy measures described in Section III. Results are presented and discussed in the sequel.

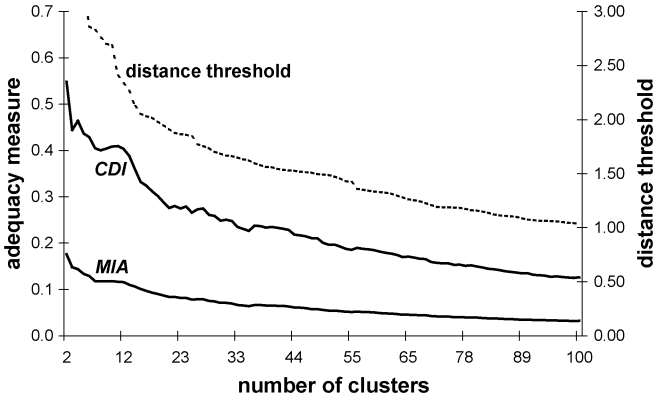

Fig. 3. MIA, CDI and distance threshold variation for different numbers of clusters resulting from the modified follow-the-leader algorithm.

TABLE I
EVOLUTION OF THE MODIFIED FOLLOW-THE-LEADER ALGORITHM

Cycle #	No. clusters	No. of patterns changing cluster
1	16	234
2	16	27
3	16	9
4	16	13
5	16	5
6	16	2
7	16	0

TABLE II
RESULTS OF THE MODIFIED FOLLOW-THE-LEADER ALGORITHM

Cluster #	No. samples	Cluster #	No. samples	Cluster #	No. samples
1	56	7	15	13	1
2	43	8	3	14	3
3	18	9	1	15	1
4	21	10	1	16	1
5	42	11	1		
6	24	12	3		

A. Application of the Modified Follow-The-Leader Algorithm

Several executions were undertaken with various distance thresholds to modify the resulting number of clusters. Fig. 3 shows the distance threshold values corresponding to a given number of clusters and how the MIA and CDI variation depends on the number of resulting clusters. Running the modified follow-the-leader algorithm with a distance threshold $\rho = 2.266$ results in $K = 16$ clusters, which is assumed to be a reasonable value for the number of clusters (the choice of the number of clusters depends on the need of the operators). The performance of the algorithm for $K = 16$ is shown in Table I, while Table II shows the number of samples included in each cluster. The representative load patterns are presented in Fig. 4.

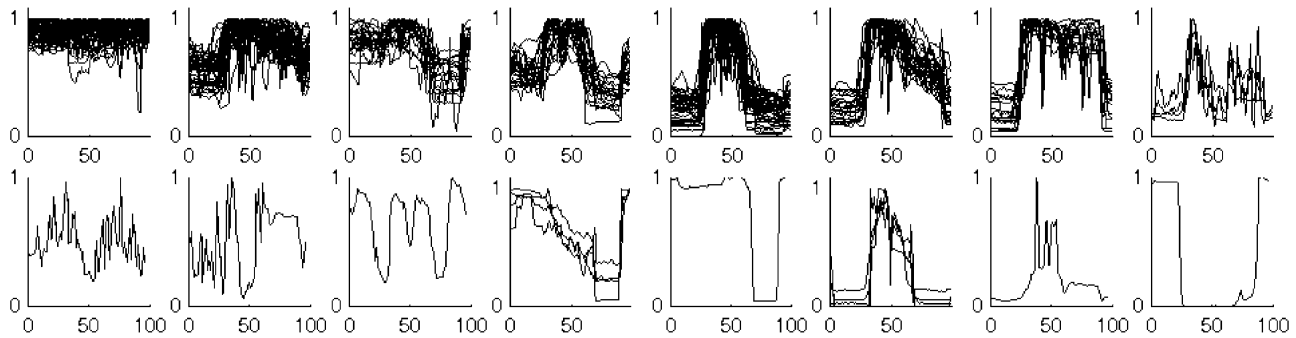

Fig. 4. Clustering results for the modified follow-the-leader algorithm. Horizontal axis: quarters of hour. Vertical axis: per-unit power.

B. Application of the SOM

The SOM performs the projection of the H-dimension space containing the M representative load patterns into a bidimensional space containing $N = N_1 \times N_2$ units. Each unit is represented by a H-dimension vector. In order to obtain a faster stabilization of the map units during the training procedure, the grid edges should roughly correspond to the major dimensions of the probability distribution of the data set. We used rectangular grids with height/width ratio N_1/N_2 calculated as a function of the ratio between the two major eigenvalues of the covariance matrix of the data set [11]. Following the same principle, the learning session becomes faster if the units are initialized as a linear combination of the eigenvector corresponding to the same two major eigenvalues. Another key issue is the choice of the adequate number of units. A heuristic formula [11] suggests using for N a particular number $N^* = 5\sqrt{M}$, so that the number of map units is lower than the number of samples. However, we observed that larger maps help in obtaining a cleaner separation of the receptive areas.

Since the samples are randomly presented during the training, different training sessions produce obviously different maps. However, if the grid parameters (shape and size) and the learning parameters (neighborhood function, shrinking rate, learning rate) are judiciously chosen, the impact of different runs on the final cluster formation is low.

The quality of the obtained map in terms of the training data set is measured by the *quantization error* ε_Q and the *topographic error* ε_T. The first one measures the resolution of the map as the average distance from each sample of the data set to its winning units

$$\varepsilon_Q = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{c}_w - \mathbf{x}_i\| \tag{7}$$

The second one measures the distortion of the map as the percentage of samples for which the winning unit and the second winning unit are not neighboring map units

$$\varepsilon_T = \frac{1}{M}\sum_{i=1}^{M} \text{neighb}(\mathbf{c}_{w1}, \mathbf{c}_{w2}) > 1 \tag{8}$$

where, for each sample, $\mathbf{c}_{w1}$ and $\mathbf{c}_{w2}$ are, respectively, the first and second closest units.

We further define the number of *receptive units* n_R as the total number of units that win the competition in the training set, that is, the complement of the number of dead units. We use the ratio $\psi_M = n_R/M$ as another way to represent the resolution property. A ratio ψ_M lower to unity even for very large numbers of N means that in the map it is not possible to obtain different winning units for every load pattern. Another ratio $\psi_N = n_R/N$ is used to represent the degree of utilization of the map.

Finally, in order to find clusters on the trained map, we employed both visual inspection and an automatic k-means clustering procedure. The k-means algorithm searches a predefined number of clusters on the maps units, performing several runs for finding the clustering with the minimum quantization error. Note that this procedure is feasible just because of the topological preserving properties of the SOM. Actually, if the map were

TABLE III
SOM DIMENSION IMPACT ON ERRORS AND TRAINING TIME

$N_1 \times N_2$	N	ε_Q	ε_T	Training time (s)
13 x 6	78	0.923	0.004	1.2
14 x 7	98	0.879	0.026	2.6
19 x 8	152	0.809	0.043	3.0
20 x 10	200	0.760	0.026	4.1
23 x 11	253	0.719	0.030	5.6
25 x 12	300	0.677	0.026	7.5
27 x 13	351	0.639	0.030	9.7
31 x 13	403	0.614	0.026	13.0
32 x 14	448	0.573	0.009	16.8
33 x 15	495	0.549	0.013	20.8
34 x 16	544	0.532	0.000	26.2
38 x 16	608	0.492	0.004	33.8
38 x 17	646	0.475	0.009	39.1
39 x 18	702	0.453	0.000	48.2
42 x 18	756	0.428	0.000	58.3
42 x 19	798	0.412	0.000	66.7
66 x 30	1980	0.087	0.000	737.8

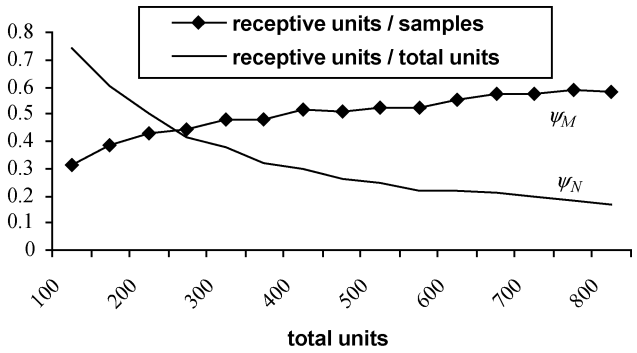

Fig. 5. Number of receptive units versus number of samples and total units.

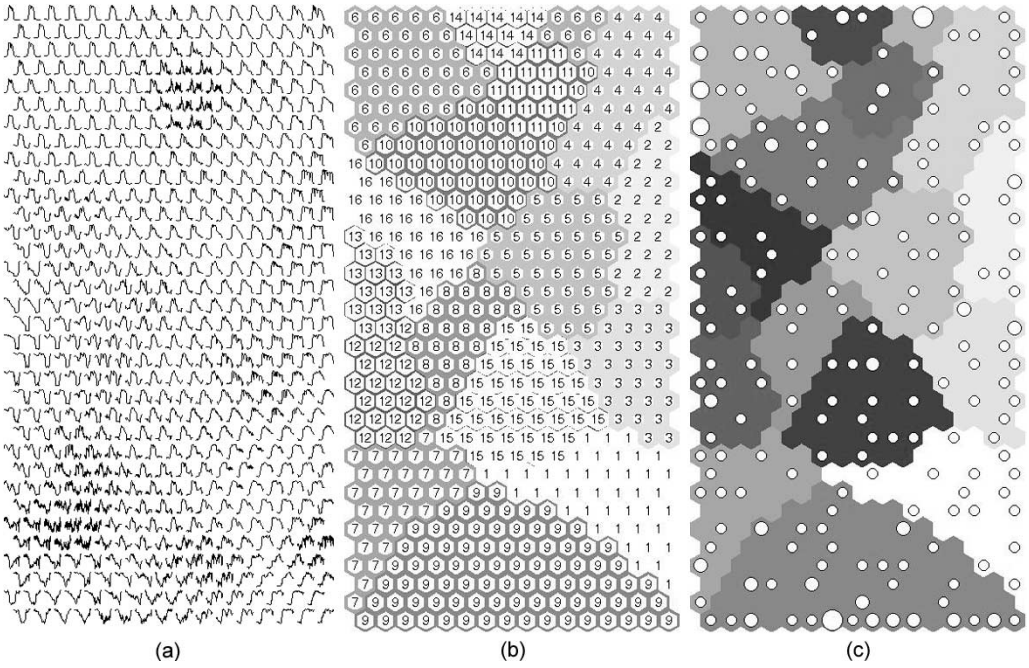

Fig. 6. Characteristics of the SOM application with the 34 × 16 map: a) the "codebook" of the 34 × 16 map; b) cluster numbering for $K = 16$; c) location of the winning units for $K = 16$.

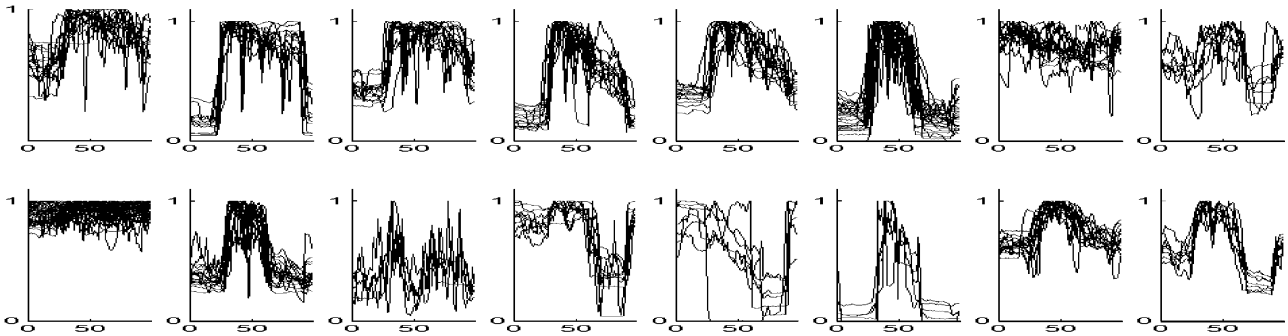

Fig. 7. Clustering results for the SOM with $K = 16$. Horizontal axis: quarters of hour. Vertical axis: per-unit power.

trained by a simple competitive learning algorithm, the clusters found by k-means would be composed by nonadjacent units.

We performed several tests with maps of different size. In all of the tests, the map grid was rectangular, according to the criteria above exposed, composed by hexagonal cells and with a neighborhood function of normal type. The overall results are presented in Table III (computation times are referred to a Pentium III 733-MHz PC). In our case, $N^* = 76.5$. Several maps have been considered, starting from the 13 × 6 map with $N = 78$, up to the large 66 × 30 map with $N = 1980$. While the quantization error decreases by increasing the map size, the topographic error exhibits a nonmonotonic behavior. The values assumed by the ratios ψ_M and ψ_N in function of the total number of units are shown in Fig. 5. The ratio ψ_M does not reach the unity value, and for the large 66 × 30 map $n_R = 230$, leading to $\psi_M = 0.983$, meaning that at least one unit is activated by multiple load patterns. The ratio ψ_N decreases when the total number of units increases, resulting in $\psi_N = 0.116$ for the 66 × 30 map.

In order to detail the SOM solution, we show the results obtained by using the map with 34 × 16 units. Fig. 6(a) shows the patterns learnt by the units—the so-called "codebook" of the map. These patterns are different with respect to the input load patterns, their shape being set and modified during the training process. On the results of the trained map, the k-means algorithm was programmed to find $K = 16$ clusters. The cluster numbering on the partitioned map is shown in Fig. 6(b), while Fig. 6(c) shows the size of the activity bubbles, where the circles drawn on the maps have a diameter proportional to the number of samples classified by each unit. The larger activity bubbles are used to form the main clusters. Small bubbles and single units correspond to particular cases. The cluster numbering shown in Fig. 6(b) corresponds to the clusters of Fig. 7, where the load patterns belonging to each cluster are plotted in the row order.

V. Performance Comparisons

The results of the clustering process performed with the modified follow-the-leader or with the SOM are the number of customer classes and the class composition. Each class is then represented by its class representative load profile. Comparing

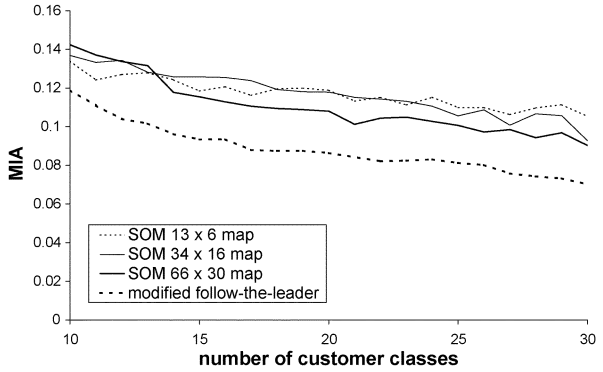

Fig. 8. Performance comparison through MIA calculation.

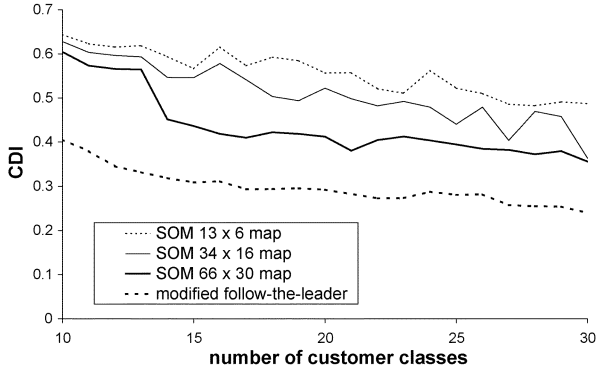

Fig. 9. Performance comparison through CDI calculation.

the performance of the methods requires computing the MIA and CDI indicators. According to the formulation of these indicators, which explicitly contain the number of resulting classes K, the comparisons makes sense only when K is the same. Fig. 8 and Fig. 9 show the results of adequacy comparison by, respectively, using the MIA and CDI indicators for a number of customer classes variable from 10 to 30. The modified follow-the-leader algorithm always exhibits better performance with respect to the SOM, with lower MIA and CDI values. For the SOM, enlarging the number of units in the maps leads to an improvement of the adequacy indicators, but the adequacy level does not reach the one of the modified follow-the-leader algorithm. Note also that the results for a given value of K depend on the map initialization, so that the evolution of the indicators with respect to K is not monotonic.

In the SOM maps, both MIA and CDI decrease with the increase of the number of clusters because the resolution of the obtained clustering obviously increases. However, the modified follow-the-leader procedure always performs better whereas SOM clustering shows higher values of MIA and CDI. The reason is that the first method works in the full M-dimension space of the samples, whereas the SOM units are M-dimension representations of the original samples slightly skewed by the projection in the two-dimensional map. Moreover, MIA values of the SOM clustering tend to be similar for small grids because in these grids there are few dead units. Lower MIA values, for a given number of resulting clusters, are obtained in large grids where the clusters are encircled by dead units; therefore, the centroids are better defined. This does not happen for the CDI indicator, because the inner average distance of the clusters always decreases if the number of available units increases. Moreover, by comparing Fig. 4 and Fig. 7, the result is that the modified follow-the-leader algorithm tends to isolate load patterns with uncommon behavior, while the SOM is less effective in this task due to the input space compression.

VI. Concluding Remarks

Two methods have been presented for performing the classification of the electricity customers on the basis of their electrical behavior. The modified follow-the-leader algorithm works directly with the whole set of data and gives automatic clustering without modifying the search space. The number of clusters cannot be fixed directly, but it depends on the user-defined threshold parameter, so that obtaining the desired number of clusters may require successive executions of the algorithm by adjusting the threshold value.

The SOM modifies the search space, showing its results on a two-dimensional map, thus providing a very simple visual explanation of the clustering procedure. The computation time remains reasonable for the SOM application with today's computers. A drawback of the SOM is that it produces as output the units with their associated samples, requiring postprocessing of the results to form the clusters.

The two methods presented can effectively assist the electricity providers in performing customer classification. The different but to some extent complementary characteristics of the two methods suggest using them in a way depending on the objectives. The high adequacy of the modified follow-the-leader algorithm makes it effective for applications to large customer sets and for easily isolating uncommon load patterns, while the easy visualization of the SOM makes it useful especially for tutorial and training purposes.

The formation of customer classes is the first step to study possible options for tariff diversification. The activity already in progress is related to assign dedicated tariff rates to each customer class in such a way to maximize the profits of the electricity providers under the constraints imposed by regulatory rules in the form of price or revenue caps. Results of this work will be reported by the authors in the near future.

References

[1] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Customer characterization options for improving the tariff offer," *IEEE Trans. Power Syst.*, vol. 18, pp. 381–387, Feb. 2003.

[2] B. D. Pitt and D. S. Kirschen, "Application of data mining techniques to load profiling," in *Proc. IEEE Power Ind. Comput. Applicat.*, Santa Clara, CA, May 16–21, 1999, pp. 131–136.

[3] A. Nazarko and Z. A. Styczynski, "Application of statistical and neural approaches to the daily load profile modeling in power distribution systems," in *Proc. IEEE Transm. Dist. Conf.*, vol. 1, New Orleans, LA, Apr. 11–16, 1999, pp. 320–325.

[4] S. V. Allera and A. G. Horsburgh, "Load profiling for the energy trading and settlements in the UK electricity markets," in *Proc. DistribuTECH Eur. DA/DSM Conf.*, London, U.K., Oct. 27–29, 1998.

[5] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader, "Electric energy customer characterization for developing dedicated market strategies," in *Proc. IEEE Porto PowerTech*, Porto, Portugal, Sept. 10–13, 2001, paper POM5-378.

[6] M. Petrescu and M. Scutariu, "Load diagram characterization by means of wavelet packet transform," in *Proc. 2nd Balkan Power Conf.*, Belgrade, Yugoslavia, June 19–21, 2002, pp. 15–19.

[7] Y.-H. Pao and D. J. Sobajic, "Combined use of unsupervised and supervised learning for dynamic security assessment," *IEEE Trans. Power Syst.*, vol. 7, pp. 878–884, May 1992.
[8] D. Gerbec, S. Gasperic, I. Simon, and F. Gubina, "Hierarchic clustering methods for consumers load profile determination," in *Proc. 2nd Balkan Power Conf.*, Belgrade, Yugoslavia, June 19–21, 2002, pp. 9–15.
[9] R. P. Broadwater, A. Sargent, A. Yarali, H. E. Shaalan, and J. Nazarko, "Estimating substation peaks from load research data," *IEEE Trans. Power Delivery*, vol. 12, pp. 451–456, Jan. 1997.
[10] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. New York: Springer-Verlag, 1989.
[11] *SOM Toolbox for Matlab 5*. Helsinki, Finland: Helsinki Univ. Technology, 2000.

Gianfranco Chicco (M'98) received the Ph.D. degree in electrotechnical engineering from the Politecnico di Torino, Torino, Italy, in 1992.

Currently, he is Associate Professor of distribution systems at the Politecnico di Torino, Torino, Italy. His research activities include power systems and distribution systems analysis, competitive electricity markets, load management, artificial intelligence applications, and power quality.

Dr. Chicco is a member of IREP and AEI.

Roberto Napoli (M'74) graduated in electrotechnical engineering from the Politecnico di Torino, Torino, Italy, in 1969.

Currently, he is Professor of Electric Power Systems at the Politecnico di Torino and Chairman of the Italian Electric Power Systems National Research Group. His research activities include power system analysis, planning and control, artificial intelligence applications, and competitive electricity markets.

Dr. Napoli is a member of CIGRE, IREP, and AEI.

Federico Piglione graduated in electrotechnical engineering from the Politecnico di Torino, Torino, Italy, in 1977.

Currently, he is Associate Professor of industrial electrical systems at the Politecnico di Torino, Torino, Italy. His major research interests include power system analysis, load forecasting, neural networks, and artificial intelligence applications to power systems.

Petru Postolache (M'01) received the M.Sc. degree in electrical engineering and the Ph.D. degree from the University Politechnica Bucharest (UPB), Bucharest, Romania.

Currently, he is Professor with the Electric Power Engineering Department at UPB. His main research interests include electrical energy efficiency, power distribution, electrical disturbances, and power quality analysis.

Dr. Postolache is a Member of the National Romanian Committee of CIRED and Chairman of its SC4.

Mircea Scutariu (M'99) received the M.Sc. and Ph.D. degrees in electric power engineering from University Politehnica Bucharest (UPB), Bucharest, Romania, in 1990 and 1997, respectively.

Currently he is the Head of the Technical Bureau of the Electrica Muntenia Sud Distribution and Supply Company, Bucharest, Romania, and his assignments involve technical support related to all electricity distribution issues. His areas of scientific interest cover the probabilistic methods applied to power systems, power quality, load management, and distribution network optimization.

Dr. Scutariu is a Member of ASTER.

Cornel Toader (M'01) received the M.Sc. and Ph.D. degrees from the University Politehnica Bucharest (UPB), Romania, in 1968 and 1988, respectively.

Currently, he is Full Professor of Electric Energy Utilization at the UPB and Head of the Electric Energy Utilization Group of the Electric Power Engineering Department. His research interests include power quality issues analysis, network modeling, and electrotechnologies.