Comparisons among clustering techniques for electricity customer classification

ARTICLE *in* IEEE TRANSACTIONS ON POWER SYSTEMS · JUNE 2006

Impact Factor: 3.53 · DOI: 10.1109/TPWRS.2006.873122 · Source: IEEE Xplore

3 AUTHORS:

Gianfranco Chicco
Politecnico di Torino
203 PUBLICATIONS **2,158** CITATIONS
SEE PROFILE

R. Napoli
Politecnico di Torino
87 PUBLICATIONS **1,003** CITATIONS
SEE PROFILE

Federico Piglione
Politecnico di Torino
18 PUBLICATIONS **298** CITATIONS
SEE PROFILE

Comparisons Among Clustering Techniques for Electricity Customer Classification

Gianfranco Chicco, *Member, IEEE*, Roberto Napoli, *Member, IEEE*, and Federico Piglione

***Abstract*—The recent evolution of the electricity business regulation has given new possibilities to the electricity providers for formulating dedicated tariff offers. A key aspect for building specific tariff structures is the identification of the consumption patterns of the customers, in order to form specific customer classes containing customers exhibiting similar patterns. This paper illustrates and compares the results obtained by using various unsupervised clustering algorithms (modified follow-the-leader, hierarchical clustering, K-means, fuzzy K-means) and the self-organizing maps to group together customers with similar electrical behavior. Furthermore, this paper discusses and compares various techniques—Sammon map, principal component analysis (PCA), and curvilinear component analysis (CCA)—able to reduce the size of the clustering input data set, in order to allow for storing a relatively small amount of data in the database of the distribution service provider for customer classification purposes. The effectiveness of the classifications obtained with the algorithms tested is compared in terms of a set of clustering validity indicators. Results obtained on a set of nonresidential customers are presented.**

***Index Terms*—Clustering, curvilinear component analysis, customer classification, follow-the-leader, fuzzy K-means, hierarchical clustering, K-means, load pattern, principal component analysis, Sammon map, self-organizing map (SOM).**

I. Introduction

THE evolution of the electricity distribution regulation has given electricity providers new possibilities of formulating dedicated tariffs, to be applied to different classes of electricity customers under a set of regulatory-imposed constraints (e.g., price or revenue caps). For the purpose of defining suitable tariff structures, the existing customer classifications based on the type of activity are scarcely correlated to the actual evolution of the electrical consumption and, as such, give poor information to the distribution providers [1]. Identifying the consumption patterns of the customers and grouping together customers exhibiting similar patterns may be significantly more helpful [2], [3]. In order to form the customer classes, the whole set of customers can be preliminarily partitioned into macro-categories defined on the basis of global criteria (e.g., residential, nonresidential, lighting, etc.). Inside each macro-category, a refined classification can be identified by taking into account the actual electrical behavior of the consumers. For a given customer, the daily load pattern corresponding to a specified *loading condition* (e.g., working days during a given seasonal period) can be built by averaging the load data monitored during a period of observation in that loading condition. By normalizing the daily load pattern data with respect to a reference power, it is possible to form the *representative load pattern* (RLP) of each customer. By assuming the maximum value of the RLP as reference power, all values of the RLP fall into the (0,1) range.

The set of RLPs of the selected customers can be used to define a number of customer classes according to a specified shape-based criterion. For this purpose, suitable clustering techniques are required to form the customer classes. Different types of clustering techniques have been proposed in the literature for assisting customer classification and load profiling, including applications of classical clustering and statistical techniques [1], [4]–[7], neural networks [6], [8]–[10], and fuzzy logic [10], [11]. The clustering results lead to the formation of the *class representative load patterns* (CRLPs) representing the customer classes in the specified loading condition. Each CRLP is built on the basis of the load patterns aggregated in the same customer class and may represent the load profile of the class, adopted for tariff formulation purposes. However, since the RLPs used in the clustering procedure are normalized, the cluster centroids cannot give the CRLPs directly. It is then necessary to compute each CRLP as a weighted average of the initial load pattern data belonging to the corresponding cluster, assuming as weights the reference powers of the individual load patterns.

This paper presents the results of a detailed investigation on the performance of various clustering algorithms, some of which have been adapted by the authors to fit the customer classification needs. The scope of the paper is twofold:

1) comparing the performance of various clustering algorithms, including the modified follow-the-leader, hierarchical clustering of different types, K-means, fuzzy K-means, and the Kohonen self-organizing map (SOM); the results presented in [12] are extended in this paper, particularly for what concerns hierarchical clustering and SOM;
2) comparing a set of techniques aimed at reducing the size of the data set forming the input to the clustering procedure; original developments are presented concerning the application of Sammon maps, principal component analysis (PCA), and curvilinear component analysis (CCA); the data set size reduction allows for storing a smaller number of data for characterizing each customer in the distribution service provider database, thus saving memory space and speeding up the computational procedures.

The effectiveness of the clustering algorithms is compared by using properly defined metrics able to rank the *clustering validity* [1], [9], [12].

Manuscript received January 28, 2005; revised October 11, 2005. This paper is based on a contribution presented by the authors at the 2003 IEEE Bologna Power Tech, Bologna, Italy, June 23–26, 2003. Paper no. TPWRS-00052-2005.

The authors are with the Dipartimento di Ingegneria Elettrica, Politecnico di Torino, I-10129 Torino, Italy (e-mail gianfranco.chicco@polito.it; roberto.napoli@polito.it; federico.piglione@polito.it).

Digital Object Identifier 10.1109/TPWRS.2006.873122

This paper is structured as follows. Section II reviews the clustering techniques used. Section III presents the techniques for data set size reduction. Section IV introduces the indicators for clustering validity assessment. Section V shows the results of applying the clustering and data size reduction techniques to a real set of nonresidential load patterns, comparing the results obtained in terms of clustering validity. The last section contains the concluding remarks.

II. Clustering Techniques

Starting from an initial set of M RLPs, the performance of different procedures for grouping the load patterns is compared. The procedures considered are four unsupervised clustering algorithms (hierarchical clustering, K-means, fuzzy K-means, and modified follow-the-leader) and the SOM. The analysis is developed in a Euclidean distance framework. After illustrating the definitions of distance used in this paper, the characteristics of the clustering procedures are briefly illustrated.

A. Euclidean Distance Framework

For a selected group of M customers, each RLP is characterized by a vector $\mathbf{x}^{(m)} = \{x_h^{(m)}, h = 1,\ldots,H\}$, whose H components can be time-domain data corresponding to time intervals of 15 min, 30 min, or 1 h or any other selected set of features representing the RLP [1]. The whole set of data is $\boldsymbol{X} = \{\mathbf{x}^{(m)}, m = 1,\ldots,M\}$. The clustering process associates the initial M RLPs to K customer classes, so that each cluster $\boldsymbol{X}^{(k)} \subset \boldsymbol{X}$ contains $n^{(k)}$ RLPs, for $k = 1,\ldots,K$. Let us further call $\mathbf{c}^{(k)}$ the H-component vector containing the centroid of the cluster $\boldsymbol{X}^{(k)}$.

In the Euclidean distance framework used, various types of distances are defined as follows:

- *vector-to-vector* distance, e.g., for two vectors $\mathbf{x}$ and $\mathbf{y}$, each of them with H components

$$d(\mathbf{y},\mathbf{x}) = \sqrt{\frac{1}{H}\sum_{h=1}^{H}(y_h - x_h)^2} \tag{1}$$

- *vector-to-set* distance, computed by using the distances between the vector $\mathbf{y}$ and each of the M members of the set $\boldsymbol{X}$

$$d(\mathbf{y},\boldsymbol{X}) = \sqrt{\frac{1}{M}\sum_{\mathbf{x}\in\boldsymbol{X}} d^2(\mathbf{y},\mathbf{x})} \tag{2}$$

- *average set-to-set* distance, given by the mean distance between all pairs of members $\mathbf{x}_q$ of the set $\boldsymbol{X}$ (with Q members) and $\mathbf{y}_j$ of the set $\boldsymbol{Y}$ (with J members)

$$d(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{QJ}\sum_{q=1}^{Q}\sum_{j=1}^{J} d(\mathbf{x}_q,\mathbf{y}_j) \tag{3}$$

- *intraset* distance, computed by using the vector-to-set distances for the M members of the set $\boldsymbol{X}$

$$\hat{d}(\boldsymbol{X}) = \sqrt{\frac{1}{2M}\sum_{m=1}^{M} d^2\left(\mathbf{x}^{(m)},\boldsymbol{X}\right)}. \tag{4}$$

B. Hierarchical Clustering

In hierarchical clustering [13], there are initially M singleton clusters, as much as the number of RLPs. At first, a $M \times M$ similarity matrix is built using the Euclidean norm distance criterion. Let us call $\gamma^{(q,s)}$ the value expressing the similarity between the clusters $\boldsymbol{X}^{(q)}$ and $\boldsymbol{X}^{(s)}$. Afterwards, the M RLPs are grouped into binary clusters by using a linkage criterion based on the similarity matrix. The process is iteratively repeated by merging the clusters of each level into bigger ones at the upper level, until all RLPs are grouped in a single cluster. The history of the process is kept in order to form a binary tree structure, whose root is the cluster that contains the whole data set.

The *linkage criterion* measures the similarity between clusters at each level and determines the cluster formation at the upper level. The extreme cases for these criteria include the *single* linkage, for which the similarity between two clusters depends on the closest pair of members in the two clusters, and the *complete* linkage, for which the similarity between two clusters depends on the farthest pair of members in the two clusters [14]. As such, the single linkage criterion may lead to the formation of few large clusters, whereas the complete linkage criterion may form too many clusters. In order to prevent these effects, other linkage criteria, such as *average distance* and *Ward* [15], have been defined.

With the *average distance* criterion, grouping two clusters $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(t)}$ depends on the average distance

$$\gamma_A^{(s,t)} = d\left(\boldsymbol{X}^{(s)},\boldsymbol{X}^{(t)}\right). \tag{5}$$

Once two clusters $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(t)}$ have been merged to form $\boldsymbol{X}^{(w)}$, the similarity between the new cluster $\boldsymbol{X}^{(w)}$ and another cluster $\boldsymbol{X}^{(g)}$ becomes

$$\gamma_A^{(w,g)} = \frac{1}{2}\left(\gamma_A^{(s,g)} + \gamma_A^{(t,g)}\right). \tag{6}$$

The hierarchical tree (or *dendrogram*) of Fig. 1 is obtained by grouping the RLPs of the data set by this method.[1] The horizontal axis contains the RLP identifiers,[2] whereas the height of each vertical branch represents the similarity between each pair of merged clusters. The final clusters are then constructed by choosing in the binary tree the maximum distance admissible or by directly selecting the distance corresponding to the desired number of clusters.

In the *Ward* linkage criterion, the clusters are formed in order to minimize the increase of the within-cluster sums of squares. The similarity between two clusters $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(t)}$ is then measured as the increase of these sums of squares if the two clusters were merged

$$\gamma_W^{(s,t)} = \frac{n^{(s)}n^{(t)}}{n^{(s)}+n^{(t)}} d^2\left(\mathbf{c}^{(s)},\mathbf{c}^{(t)}\right) \tag{7}$$

where $\mathbf{c}^{(s)}$ and $\mathbf{c}^{(t)}$ are the centroids of the two clusters. Once two clusters $\boldsymbol{X}^{(s)}$ and $\boldsymbol{X}^{(t)}$ have been merged to form $\boldsymbol{X}^{(w)}$,

[1]The dendrograms of Figs. 1 and 2 are built by using the data of the case with $M = 234$ load patterns presented in Section V.

[2]For the sake of clarity of the representation, the RLP identifiers are not shown in Figs. 1 and 2.

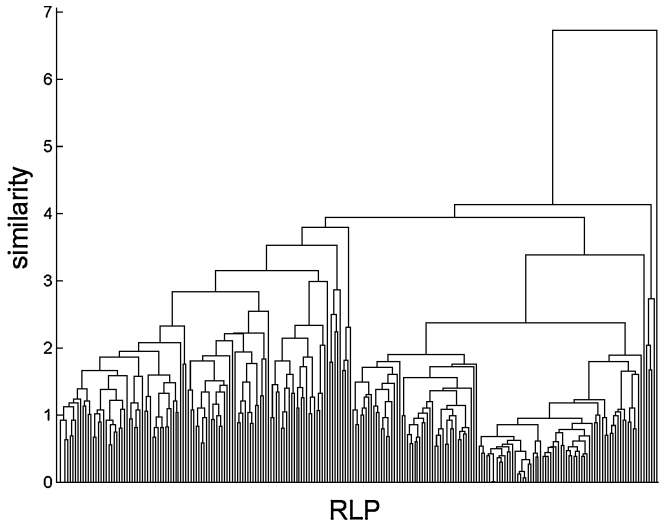

Fig. 1. Dendrogram of the hierarchical clustering with *average distance* criterion. Horizontal axis: RLP identifier. Vertical axis: similarity measure (5) between clusters.

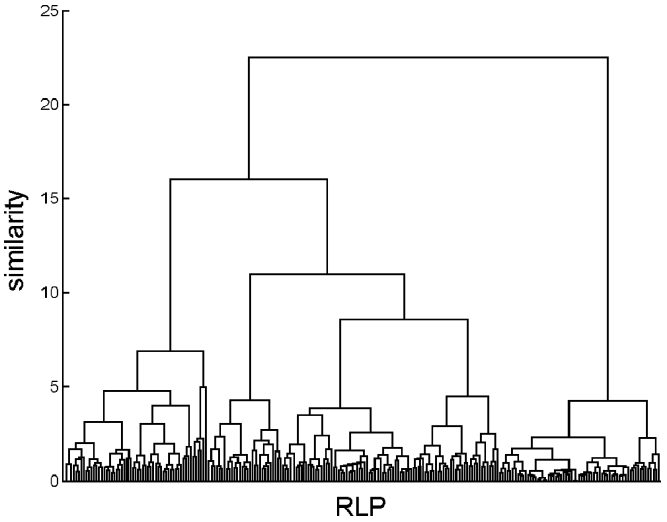

Fig. 2. Dendrogram of the hierarchical clustering with *Ward linkage* criterion. Horizontal axis: RLP identifier. Vertical axis: similarity measure (7) between clusters.

the similarity between the new cluster $\boldsymbol{X}^{(w)}$ and another cluster $\boldsymbol{X}^{(g)}$ becomes

$$\gamma_W^{(w,g)} = \frac{\left(n^{(s)}+n^{(g)}\right)\gamma_W^{(s,g)} + \left(n^{(t)}+n^{(g)}\right)\gamma_W^{(t,g)} - n^{(g)}\gamma_W^{(s,t)}}{n^{(s)}+n^{(t)}+n^{(g)}}. \tag{8}$$

Fig. 2 shows the dendrogram obtained by using the Ward linkage criterion. The comparison between the two hierarchical trees shows that the average distance criterion forms large clusters of similar RLPs and rejects the very dissimilar ones in small or singleton clusters, whereas the Ward criterion prevents the formation of large clusters.

C. K-Means

The classical K-means clustering [16] groups a data set of $\mathbf{x}^{(m)}(m = 1, \ldots, M)$ RLPs in $k = 1, \ldots, K$ clusters by means of an iterative procedure. A first guess is made for the cluster centroids $\mathbf{c}^{(k)}$, for $k = 1, \ldots, K$ (usually chosen at random among the RLPs of the data set). The K centroids classify the RLPs, in the sense that the RLP $\mathbf{x}^{(m)}$ belongs to the cluster $\boldsymbol{X}^{(k)}$ if the distance $d(\mathbf{x}^{(m)}, \mathbf{c}^{(k)})$ is the minimum of all the K distances between $\mathbf{x}^{(m)}$ and the cluster centroids. The estimated centroids are used to classify the RLPs into clusters (usually by Euclidean norm), and their values $\mathbf{c}^{(k)}$ are recalculated in such a way that each hth element of the centroid $\mathbf{c}^{(k)}$ is the average of the hth elements of the load patterns belonging to the cluster $\boldsymbol{X}^{(k)}$. The procedure is repeated until stabilization of the cluster centroids. The optimal number of clusters is not known *a priori*, and the clustering quality depends on the value of K specified by the user.

D. Fuzzy K-Means

Fuzzy K-means clustering [17] is rather similar to K-means clustering, but each RLP $\mathbf{x}^{(m)}$ has a membership degree $a^{(m,k)}$ with respect to each cluster $\boldsymbol{X}^{(k)}$. The procedure is initialized by choosing K RLPs $\mathbf{c}^{(k)}$ as cluster centroids and assigning to the M RLPs of the data set the membership degree with the K centroids

$$a^{(m,k)} = \left(\sum_{\nu=1}^{K} \left(\frac{d\left(\mathbf{x}^{(m)}, \mathbf{c}^{(k)}\right)}{d\left(\mathbf{x}^{(m)}, \mathbf{c}^{(\nu)}\right)} \right)^{\frac{1}{\beta-1}} \right)^{-1} \tag{9}$$

where $\beta > 1$ is a control parameter that specifies the level of fuzziness.[3] Each cluster centroid $\boldsymbol{c}^{(k)}$ is then updated by replacing its value with the fuzzy mean of all the RLPs belonging to the cluster k

$$\mathbf{c}^{(k)} = \left(\sum_{m=1}^{M} \left(a^{(m,k)}\right)^{\beta} \mathbf{x}^{(m)} \right) \left(\sum_{m=1}^{M} \left(a^{(m,k)}\right)^{\beta} \right)^{-1}. \tag{10}$$

The procedure is repeated until stabilization of the cluster centroids. Again, the number of clusters and membership criteria are user-defined parameters, which have to be tuned by trial-and-error.

E. Modified Follow-the-Leader

The follow-the-leader procedure introduced in [18] does not require initialization of the number of clusters and uses an iterative process to compute the cluster centroids. A first cycle of the algorithm sets the number K of clusters and the number of patterns $n^{(k)}$ belonging to each cluster $k = 1, \ldots, K$ by using a follow-the-leader approach, depending on a distance threshold ρ. The subsequent cycles refine the clusters, by possibly reassigning the patterns to closest clusters. The procedure stops when the number of patterns changing clusters in a single cycle is zero. The process is essentially controlled by the distance threshold ρ, which has to be chosen by a trial-and-error approach. This procedure has been modified by the authors to fit the needs of the proposed classification, by taking into account the data dispersion in the input vector [1], [9]. For this purpose, the Euclidean metric used in the original algorithm has been modified by introducing for each index a weighting factor $\sigma_h^2/\overline{\sigma}^2$, where σ_h^2 is the variance of the hth feature computed from all the load patterns in the initial population, and $\overline{\sigma}^2$ is the

[3] In the fuzzy K-means procedure run in this paper, the control parameter was set to $\beta = 2$.

average value of the variance σ_h^2 for $h = 1, \ldots, H$. As such, the impact of the indexes having a high variance is amplified in the computation of the weighted Euclidean distance.

F. SOMs

The SOM [19] is an unsupervised neural network that projects a data set of M vectors $\mathbf{x}^{(m)}$ embedded in a H-dimension space onto a set of C prototype vectors $\mathbf{u}^{(c)}$. The prototype vectors have the same dimension of the original data vectors but are arranged in a lower S-dimension space (usually a two-dimension rectangular grid). The prototype vectors $\mathbf{u}^{(c)}$ are the units of a neural competitive layer, where only one unit responds at the presentation of each input $\mathbf{x}^{(m)}$. The activation function is an inverse function of $d(\mathbf{x}^{(m)}, \mathbf{u}^{(c)})$, so that the unit that is closest to $\mathbf{x}^{(m)}$ wins the competition. The winning unit is then updated according to the relationship

$$\mathbf{u}_{new}^{(c)} = \mathbf{u}_{old}^{(c)} + \eta \left(\mathbf{x}^{(m)} - \mathbf{u}_{old}^{(c)} \right) \tag{11}$$

where η is the *learning rate*. Up to this point, the SOM appears as a conventional competitive learning algorithm. The projection capability arises from the fact that the learning algorithm considers also the position of each unit $\mathbf{u}^{(c)}$ into the S-dimension grid and updates not only the weights of the winning unit but also the weights of its neighbor units in inverse proportion of their distance. 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 the units that are spatially close correspond to similar patterns in the original space. The projection is then *topology preserving*, because the proximity among objects in the input space is about preserved in the output space. A characteristic feature is that during the learning process, the units arrange themselves in receptive areas, named *activity bubbles*, encircled by units (*dead units*) that never win the competition for the samples of the training data set.

The SOM produces a visually understandable projection of the original data in the reduced dimension space of the map grid, but, like other projection algorithms, it is not a direct clustering method. Clusters could be obtained by visual inspection of the map hits (a hit is the winning unit for a given sample of data set) onto the activity bubbles [9]. However, as suggested in [20], it is a better way to cluster first the map prototype vectors by a conventional algorithm and to assign then the samples to the corresponding clusters. If the map is set up with about $C = 5\sqrt{M}$ units, as suggested in [21], this procedure reduces the computational burden of the clustering algorithm. The application of the SOM to electricity customer classification has been illustrated in detail in [9].

III. Techniques for Data Size Reduction

A generalization of the concepts introduced for handling the data structures is required to assess the effectiveness of using different sets of features. In particular, this paper investigates the possibility of adopting a set of features built in a reduced vector space but able to satisfactorily explain the characteristics of the input data set by using a small number of features. For this purpose, the original set of M RLP data contained in the set $\boldsymbol{X}$ may be subject to a vector space transformation that maps it into a new set $\boldsymbol{Z}$. Then, running a selected clustering procedure results in forming a specified number K of clusters. Each cluster $\boldsymbol{Z}^{(k)} \subset \boldsymbol{Z}$ contains $n^{(k)}$ RLPs, for $k = 1, \ldots, K$. The results of clustering are the customer class composition and the cluster centroids built in the transformed vector space. However, it is possible to retain only the information about the customer class composition and to re-compute the centroids by using the data in the original vector space. This would allow easy comparison among the clustering results, by using properly defined clustering validity indicators, such the ones presented in Section IV.

The techniques analyzed in this paper, aimed at reducing the number of dimensions in which the clustering is performed include the PCA, the Sammon map, and the CCA.

A. PCA

The PCA [22] is an optimal linear reduction method that projects a data set embedded in an H-dimension space in a R-dimension subspace, with $R \ll H$. The projection is generated by a set $\boldsymbol{V}_{\mathrm{R}}$ containing RH-dimensional vectors that maximize the explained variance of the original data set. The overall idea is to capture the maximum amount of covariance amongst the data features and eliminate by the projection those features that are highly correlated among themselves. Mathematically, the procedure requires computing the eigenvectors of the covariance matrix of the data set

$$\mathbf{V} = \mathrm{eigvec}\left(\mathrm{cov}(\mathbf{X}^{\mathrm{T}})\right) \tag{12}$$

where $\mathbf{X} = [\mathbf{x}^{(m)} - \overline{\mathbf{x}}, \quad m = 1, \ldots, M]$ is the matrix of the data set vectors, with H features (rows) and M samples (columns), $\overline{\mathbf{x}}$ is the vector containing the mean value of each feature, and $\mathbf{V}$ is the square eigenvector matrix with H eigenvectors (columns). The columns are ordered for decreasing values of the corresponding eigenvalues. The matrix $\mathbf{P}$ representing the projected $H \times R$ data set $\boldsymbol{P}$ is then given by

$$\mathbf{P} = \mathbf{X}^{\mathrm{T}} \mathbf{V}_R \tag{13}$$

where $\mathbf{V}_R$ is the reduced eigenvectors matrix, with H features (rows) and the first R eigenvectors (columns). The only drawback of the PCA is that it is a linear technique, and therefore, it cannot capture any nonlinear correlation among the variables. The major benefit of the PCA is the data compression, but, if the projection is made in a two-dimension subspace, the PCA could be also useful for the visual inspection of the data set. However, the map quality is lower than the one produced by the SOM.

B. Sammon Map

The Sammon map [23] overcomes the PCA limits by performing a nonlinear projection that preserves the topology of the original space. In order to achieve this goal, an error function is defined as follows:

$$E = \frac{1}{\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} d(\mathbf{x}_i, \mathbf{x}_j)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{\left[d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j) \right]^2}{d(\mathbf{x}_i, \mathbf{x}_j)} \tag{14}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ and $d(\mathbf{y}_i, \mathbf{y}_j)$ are, respectively, the distances between the pairs of data vectors in the H-dimension original space and in the R-dimension reduced subspace. Since in (14) the error function is divided by $d(\mathbf{x}_i, \mathbf{x}_j)$, the algorithm enhances the small distances in the original space. In this way, the Sammon map can "unfold" nonlinear structures inherent to the data set in the original space. The procedure initializes the projected vectors $\mathbf{y}_i$ in random fashion and uses then a gradient descent algorithm to minimize the Sammon error function. Like the PCA, the Sammon map performs an effective data compression, but, for two-dimension projections, it is also useful for data visualization.

C. CCA

The CCA [24] is a self-organizing neural network that projects an H-dimension data set in a R-dimension map (with $R \ll N$) preserving the original topology and unfolding nonlinear data structures in the original space. It represents a nonlinear extension of the PCA and an improvement of the Sammon map. The CCA is composed by a set of units, each having two weight vectors $\mathbf{x}_i$ and $\mathbf{y}_i$. The two vectors have, respectively, the dimension of the input and the output space. This is another way of representing an SOM structure, where the vectors $\mathbf{x}_i$ are the prototypes, and the vectors $\mathbf{y}_i$ are fixed and carry the position of the unit onto the grid. The learning process starts with the vector quantization on the input vectors and the random initialization of output vectors. The solution procedure minimizes iteratively the error function

$$E = \frac{1}{2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \left[d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{y}_i, \mathbf{y}_j) \right]^2 F\left(d(\mathbf{y}_i, \mathbf{y}_j), \lambda \right) \tag{15}$$

where $d(\mathbf{x}_i, \mathbf{x}_j)$ and $d(\mathbf{y}_i, \mathbf{y}_j)$ are the distances in the input and output space, whereas $F(.)$ is a decreasing function, with decay rate λ, of the distances in the output space. This iterative approach, similar to that one of the SOM, allows local topology preservation and enhances the small distances in the input space. Like the SOM, the main purpose of the CCA is to give an understandable representation of the data projected in a low-dimension map.

IV. Clustering Validity Assessment

A general scheme for comparing the clustering results obtained from different clustering techniques and different feature sets is presented. A key point for the sake of comparison is that the set $\boldsymbol{X}$ of the data used for computing the centroids must be always the same, even though the clustering has been run by using a set of data $\boldsymbol{Z}$ built in a transformed vector space, as discussed in Section III.

Let us consider the results of a clustering algorithm run by using a set of features $\boldsymbol{Z}$ to form K customer classes. Let us then consider the set $\boldsymbol{C} = \{\mathbf{c}^{(k)}, k = 1, \ldots, K\}$ of the centroids computed, with reference to the data contained in the set $\boldsymbol{X}$, on the basis of the customer partitioning into classes resulted from clustering. The clustering validity indicators considered in this paper have been defined according to a metric that merges the information on the compactness of the load patterns belonging to the same class and (inversely) on the inter-distance among the CRLPs [1], [9], [12]. These indicators depend on the set of data $\boldsymbol{Z}$ used for running the clustering algorithm, the set of data $\boldsymbol{X}$ used for computing the centroids $\boldsymbol{C}$, and the specified number of clusters K. The indicators include the *clustering dispersion indicator (CDI)* [1], [12]

$$\mathrm{CDI}(\boldsymbol{Z}, \boldsymbol{X}, K) = \frac{1}{\hat{d}(\boldsymbol{C})} \sqrt{\frac{1}{K} \sum_{k=1}^{K} \hat{d}^2 \left(\boldsymbol{X}^{(k)} \right)} \tag{16}$$

the *modified Dunn index (MDI)*, adapted from the original Dunn index [25] by using the Euclidean distances (1) and (4), for $i, j = 1, \ldots, K$

$$\mathrm{MDI}(\boldsymbol{Z}, \boldsymbol{X}, K) = \max_{1 \leq q \leq K} \left\{ \hat{d} \left(\boldsymbol{X}^{(q)} \right) \right\} \times \left(\min_{i \neq j} \left\{ d \left(\mathbf{c}^{(i)}, \mathbf{c}^{(j)} \right) \right\} \right)^{-1} \tag{17}$$

the *scatter index (SI)*, derived from the *proportion of scatter* accounted for by clustering, introduced in [5]

$$\mathrm{SI}(\boldsymbol{Z}, \boldsymbol{X}, K) = \left(\sum_{m=1}^{M} d^2 \left(\mathbf{x}^{(m)}, \mathbf{p} \right) \right) \left(\sum_{k=1}^{K} d^2 \left(\mathbf{c}^{(k)}, \mathbf{p} \right) \right)^{-1} \tag{18}$$

where $\mathbf{p}$ is the *pooled scatter*

$$\mathbf{p} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{x}^{(m)} \tag{19}$$

and a Euclidean form of the *Davies–Bouldin index (DBI)* [26], representing the system-wide average of the similarity measures of each cluster with its most similar cluster, for $i, j = 1, \ldots, K$

$$\mathrm{DBI}(\boldsymbol{Z}, \boldsymbol{X}, K) = \frac{1}{K} \sum_{k=1}^{K} \max_{i \neq j} \left\{ \frac{\hat{d} \left(\boldsymbol{X}^{(i)} \right) + \hat{d} \left(\boldsymbol{X}^{(j)} \right)}{d \left(\mathbf{c}^{(i)}, \mathbf{c}^{(j)} \right)} \right\}. \tag{20}$$

A common characteristic of these indicators is the fact that *lower* values of the indicator correspond to *better* clustering validity. In addition, since the indicators are explicitly defined in function of the number K of customer classes, comparing the clustering results is significant only when the number of customer classes formed by the various algorithms is the same.

V. Applications to Electricity Customer Clustering

A set of $M = 234$ nonresidential customers connected to the MV distribution system has been considered [9]. The RLP of each customer is obtained by averaging the data measured with a 15-min cadence in a day with a given loading condition (spring weekdays). Then, each RLP contains $H = 96$ values. This set of data $\boldsymbol{X}$ has been formed in the original vector space. Each clustering algorithm assigns the RLP of each customer to a specific cluster, providing a complete and nonoverlapping positioning of all customers.

Most of the clustering algorithms used require a preventive assignment of the number of clusters to be formed. The only exception is the modified follow-the-leader algorithm, in which the number of clusters decreases by increasing the distance threshold [9], so that its distance threshold has been adjusted

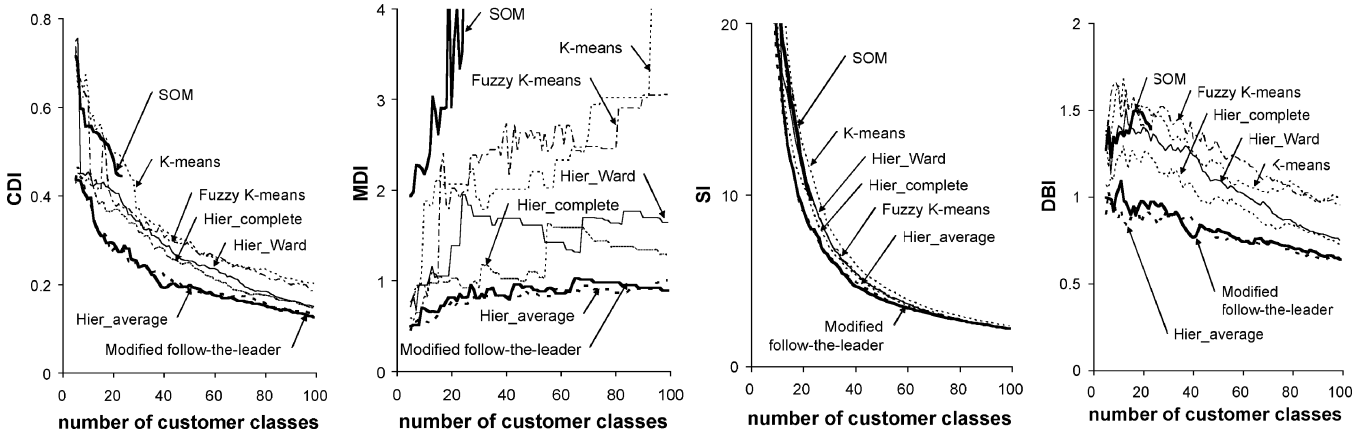

Fig. 3. Performance of the clustering validity indicators with the clustering techniques under test for $K = 5$ to 100.

during the analysis in order to obtain the same numbers of clusters imposed to the other algorithms.

Some indications on the "best" number of clusters from the classification point of view may be obtained from the clustering validity indicators shaped as decreasing functions of the number of customer classes, such as the CDI or the SI. In this case, the best number of clusters could correspond to the knee of the curve [27]. A more theoretical approach could be applied by considering the Bayesian information criterion (BIC) [28] or the informational complexity criterion (ICOMP) [29]. From Fig. 3, adopting a criterion based on the knee of the curve would result in a relatively high number of clusters (e.g, about 40). However, for our purposes, the choice of the number of customer classes mainly depends on practical aspects, as the willingness of the distribution service provider to create a specific set of tariffs, each of which is then associated to a customer class. As such, the number of customer classes for tariff diversification purposes cannot be too high, in order to allow for easy management of the commercial data and to provide a clear and nonoverwhelming amount of information to the customers concerning the tariff options.

A. Clustering Validity Assessment

Repeated executions of the clustering algorithms have been performed by varying the number of customer classes from 5 to 100 and computing the clustering validity indicators for all algorithms. This range of analysis has been chosen for the purpose of comparing the methods, even though 100 classes correspond nearly to one half of the total customers and are a number practically too high for a real application aimed at associating each customer class to a different tariff.

The results illustrated in Fig. 3 show that the information provided by the clustering validity indicators is highly consistent, with a clustering technique ranking (for increased values of the indicator with the same number of customer classes) nearly similar for the same number of customer classes. All of the methods are able to form the required number of clusters. The only exception is the SOM, for which the cluster formation requires post-processing of the map components under a given criterion, making it difficult to choose the cluster elements when the number of clusters is relatively high, and the smoothing effects during the creation of the map lead to uncertainties in the map

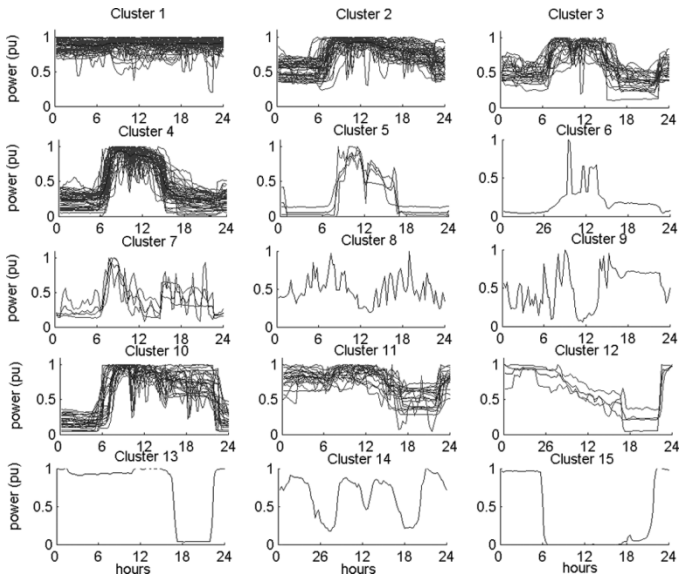

Fig. 4. Clustering results for the hierarchical clustering (with average linkage criterion) with $K = 15$ clusters.

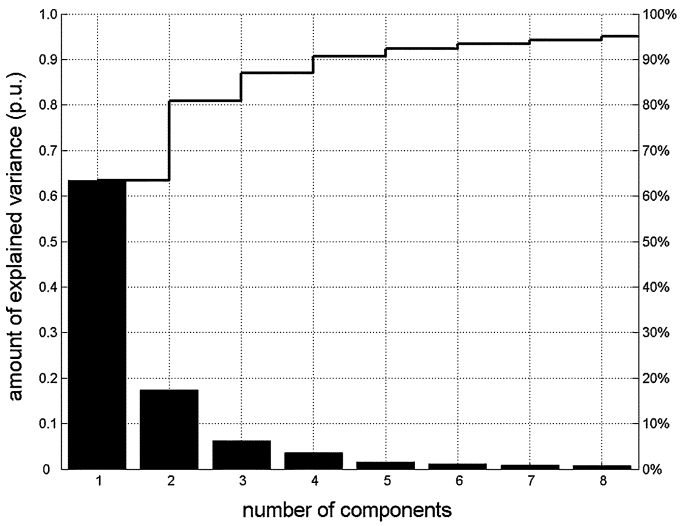

Fig. 5. Pareto diagram for the PCA application.

elements of the same order of magnitude of the differences between the cluster components. In fact, if an automated procedure is employed (clustering first the map prototype vectors and then

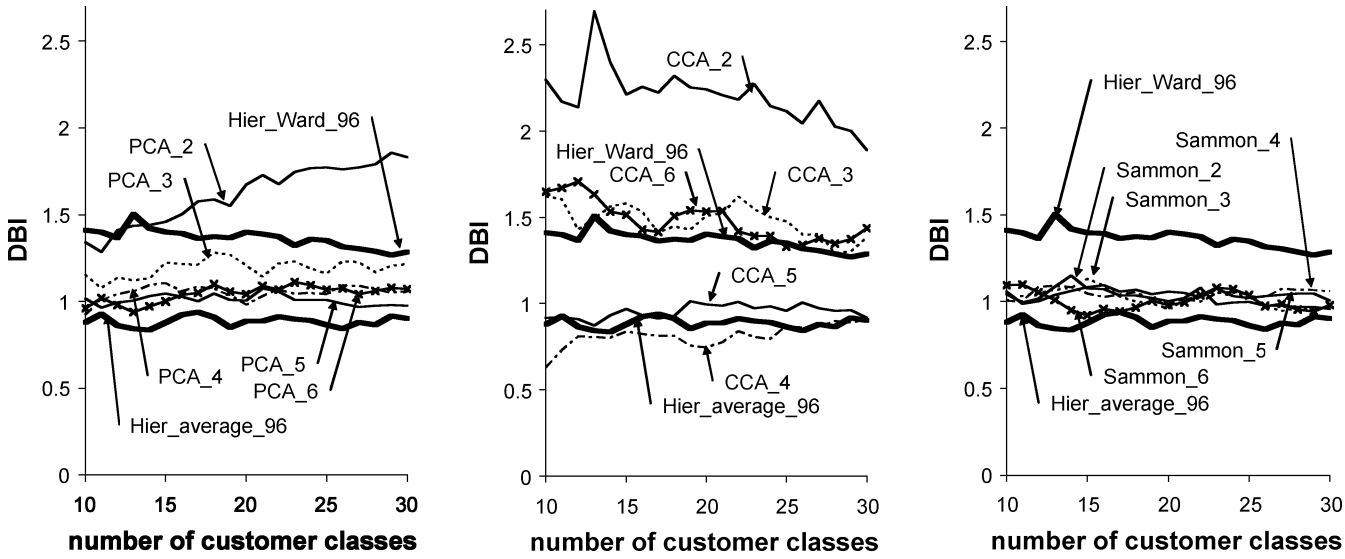

Fig. 6. Performance of the DBI indicator with different types of data sets, for $K = 10$ to 30.

assigning each sample at its best-matching prototype vector), it is not possible to obtain an arbitrary number of clusters, because the smoothing properties of the SOM form some void clusters (centroids with no best-matching units). According to this fact and due to the comparatively weak results obtained, the SOM performance has been indicated in Fig. 3 (thick line) only for low values of the number of clusters.

The results clearly show the superiority of the modified follow-the-leader and of the hierarchical clustering (with average linkage criterion) over the other techniques. This superiority mainly depends on the ability of these techniques to build well-separated classes and to isolate uncommon load patterns, as shown in Fig. 4 for the hierarchical clustering with average linkage criterion in the case resulting in $K = 15$ clusters. Note also that the various clustering techniques provide exactly the same results only when the number of clusters is equal to the number of subjects, i.e., each load pattern is a cluster center. However, following any slight decrease in the number of clusters, each method adopts a different strategy for grouping the load patterns into clusters, so that some differences in the clustering validity indicators emerge immediately.

B. Reduction of the Input Vector Size

The use of PCA, Sammon map, and CCA allows for reducing the size of the input vector formed by the RLPs. The Pareto diagram of Fig. 5, related to the use of the PCA, shows the amount of the total variance of the data set that is explained by using the first R components of the projection onto the H dimensions. The staircase-like draw represents the cumulative amount of the expected variance. For instance, the PCA with $R = 6$ is able to explain about 92% of the total variance.

The possibility of explaining a relatively large amount of the total variance by using a reduced number of components deriving from a transformation of the initial data set provides the rationale for the investigation carried out in this paper, aimed at checking the effectiveness of compacting the data set into a small number of features before clustering. In order to compare the results obtained from the various methods operating the size reduction, for each method, the hierarchical clustering with average linkage criterion has been run by using as input data the sets $\boldsymbol{Z}_R = \{\mathbf{z}_R^{(m)}, m = 1, \ldots, M\}$ resulting from the projection of the original set of data $\boldsymbol{X}$ onto a number of dimensions R of the method under test variable from 2 to 6.

The customer partitioning resulting from clustering has been used to compute the centroids on the basis of the data of the initial set $\boldsymbol{X}$. The clustering validity indicators of Section IV have been used to assess the effectiveness of the methods for data size reduction. Again, the various indicators provided consistent information in terms of ranking the alternatives. Fig. 6 shows the values of the DBI indicator obtained for a number of clusters K variable from 10 to 30. For the sake of comparison, both the indicators $DBI(\boldsymbol{Z}_R, \boldsymbol{X}, K)$ computed for $R = 1$ to 6 with the hierarchical clustering with average linkage criterion and the indicators $DBI(\boldsymbol{X}, \boldsymbol{X}, K)$ related to the hierarchical clustering with average linkage and Ward criteria on the data set with $H = 96$ (already shown in Fig. 6) are illustrated. It emerges that the Sammon maps are able to provide significantly good results with any of the reduced data sets, whereas the PCA exhibits good performance for a number of dimensions $R > 3$, and the CCA can provide the most effective results but only in some cases (i.e., for $R = 4$ and $R = 5$).

VI. Concluding Remarks

Clustering techniques are extremely useful for assisting the distribution service providers in the process of electricity customer classification on the basis of the load pattern shape. The study carried out in this paper has shown that the most relevant aspects for the suppliers are related to the possible reduction of the data set size and to the choice of the clustering method, whereas the automatic determination of the best number of clusters based on theoretical considerations is an issue of relatively low practical impact. An indicative number of customer classes not higher than 15–20 could fit the supplier's needs, even though the application of theoretical criteria to find out the best number of clusters could result in higher values.

The results of the clustering validity assessment performed in this paper show that two algorithms—the modified follow-the-leader and the hierarchical clustering run with the *average distance* linkage criterion—emerge as the most effective ones. Both algorithms are able to provide a highly detailed separation of the clusters, isolating load patterns with uncommon behavior and creating large groups containing the remaining load patterns. These properties make the two algorithms particularly suitable for a customer classification oriented toward grouping the customers into a small number of customer classes for tariff formulation purposes. The other algorithms tend to distribute the load patterns among some groups formed during the clustering process and, as such, are less effective.

This paper has also shown how it is possible to reduce the size of the data set used as input in the clustering procedure, in order to store a smaller amount of data in the electricity company's database and to speed up the clustering calculations. In general, the counterpart of the benefits of data size reduction is a lower classification effectiveness, in terms of higher clustering validity indicators. In particular, the Sammon maps have provided slightly similar clustering validity for different data size reductions, so indicating a robust behavior, whereas the CCA has exhibited the best performance but is less robust. On the basis of the results, the validity of the data size reduction methods can be generally indicated as acceptable. Trading-off between clustering validity and number of components of the input data set is left to the decision of the distribution service provider.

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, no. 1, pp. 381–387, Feb. 2003.

[2] P. Stephenson, I. Lungu, M. Paun, I. Silvas, and G. Tupu, "Tariff development for consumer groups in internal European electricity markets," in *Proc. CIRED*, Amsterdam, The Netherlands, Jun. 18–21, 2001, paper 5.3.

[3] C. S. Chen, M. S. Kang, J. C. Hwang, and C. W. Huang, "Synthesis of power system load profiles by class load study," *Elect. Power Energy Syst.*, vol. 22, pp. 325–330, 2000.

[4] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

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

[6] 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 Transmission Distribution Conf.*, vol. 1, New Orleans, LA, Apr. 11–16, 1999, pp. 320–325.

[7] D. Gerbec, S. Gasperic, and F. Gubina, "Determination and allocation of typical load profiles to the eligible customers," in *Proc. IEEE Bologna Power Tech*, Bologna, Italy, June 23–26, 2003, paper 302.

[8] R. Lamedica, L. Santolamazza, G. Fracassi, G. Martinelli, and A. Prudenzi, "A novel methodology based on clustering techniques for automatic processing of MV feeder daily load patterns," in *Proc. IEEE/PES Summer Meeting*, vol. 1, Seattle, WA, Jul. 16–20, 2000, pp. 96–101.

[9] G. Chicco, R. Napoli, F. Piglione, M. Scutariu, P. Postolache, and C. Toader, "Load pattern-based classification of electricity customers," *IEEE Trans. Power Syst.*, vol. 19, no. 2, pp. 1232–1239, May 2004.

[10] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Determining the load profiles of consumers based on fuzzy logic and probability neural networks," *Proc. Inst. Elect. Eng., Gener., Transm., Distrib.*, vol. 151, no. 3, pp. 395–400, May 2004.

[11] C. S. Özveren, C. Vechakanjana, and A. P. Birch, "Fuzzy classification of electrical load demand profiles—a case study," in *Proc. IEE Power System Management Control*, Apr. 17–19, 2002, pp. 353–358.

[12] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organizing maps to classify electricity customers," in *Proc. IEEE Bologna Power Tech*, Bologna, Italy, Jun. 23–26, 2003, paper 333.

[13] M. R. Anderberg, *Cluster Analysis for Applications*. New York: Academic, 1973.

[14] B. S. Everitt, *Cluster Analysis*, 3rd ed. London, U.K.: Edward Arnold and Halsted, 1993.

[15] J. H. Ward, "Hierarchical grouping to optimize an objective function," *J. Amer. Stat. Assoc.*, vol. 58, pp. 236–244, 1963.

[16] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.

[17] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[18] Y.-H. Pao and D. J. Sobajic, "Combined use of unsupervised and supervised learning for dynamic security assessment," *IEEE Trans. Power Syst.*, vol. 7, no. 2, pp. 878–884, May 1992.

[19] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. Berlin, Germany: Springer-Verlag, 1989.

[20] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, May 2000.

[21] *SOM Toolbox for Matlab 5*. Helsinki, Finland: Helsinki Univ. Technol., 2000.

[22] J. E. Jackson, *A User's Guide to Principal Components*. New York: Wiley, 1991, pp. 1–25.

[23] J. W. Sammon Jr., "A nonlinear mapping for data structure analysis," *IEEE Trans. Comput.*, vol. C-18, no. 5, pp. 401–409, May 1969.

[24] P. Demartines and J. Herault, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 148–154, Jan. 1997.

[25] J. C. Dunn, "Well separated clusters and optimal fuzzy partitions," *J. Cybern.*, vol. 4, pp. 95–204, 1974.

[26] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAM-1, no. 2, pp. 224–227, Apr. 1979.

[27] D. Gerbec, S. Gasperic, I. Smon, and F. Gubina, "Allocation of the load profiles to consumers using probabilistic neural networks," *IEEE Trans. Power Syst.*, vol. 20, no. 2, pp. 548–555, May 2005.

[28] G. Schwarz, "Estimating the dimension of a model," *Ann. Stat.*, vol. 6, pp. 461–464, 1978.

[29] H. Bozgodan, "On the information-based measure of covariance complexity and its applications to the evaluation of the multivariate linear models," *Commun. Stat., Theory, Methods*, vol. 19, pp. 221–278, 1990.

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

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

Roberto Napoli (M'74) received the laurea degree in electrotechnical engineering at the Politecnico di Torino (PdT), Torino, Italy, in 1969.

He is a Professor of electric power systems at the PdT and the former 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.

Federico Piglione received the laurea degree in electrotechnical engineering at the Politecnico di Torino (PdT), Torino, Italy, in 1977.

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