Knowlegde Extraction from Smart Meters for Consumer Classification

Gheorghe Grigoras, Florina Scarlatache
"Gheorghe Asachi" Technical University
Iasi, Romania
E-mail: ggrigor@ee.tuiasi.ro

Abstract— **The behavior of an electric customer can be identified through a load profile corresponding to a certain period of time. Information can be available depending on type of customer and customer classification is based only the billing data. But, there are many countries where the Smart Metering Systems are becoming more common and the customer classification and load profiling could be done based on the real consumption data. In paper, is proposed a decision trees based approach for consumer classification in representative categories characterized by typical load profiles using information provided by Smart Meters. For determination of the consumption categories, every customer is characterized by the following primary information: daily (monthly) energy consumption, minimum and maximum loads. The proposed approach was tested using small consumers that don't have a smart metering system implemented. The obtained results it demonstrated that the proposed approach can be used with the success in the optimal operation and planning of distribution systems.**

Keywords-Smart Meters; customer classification; clustering; decision trees.

I. Introduction

In many countries, legislation and regulatory initiatives have been targeted towards the modernization of the grid. Thus, the electricity companies are in a transition phase towards a new infrastructure that will respond the energy needs of the population. This will lead to an improvement of performance corresponding to power systems: increasing efficiency, integration of renewable sources, power quality and reduced costs. These benefits require the reinforcement of equipments, improving of planning and operation, increasing automation environmental sustainability and ways of limiting carbon emissions have led to increased interest in Smart Grids [1] - [3].

Technology has also been a great driver in smart grid development. A Smart Grid monitors electricity delivery and power consumption with smart meters that transmit information to utilities via communication networks, Fig 1. The physical infrastructure, that delivers energy to consumers, represents the first level. The following level is represented by Communications Systems that include Smart Meters. The upper level allows processing information for making timely operational decisions. At the last level, the applications that create values for electrical power system and customers are presented.

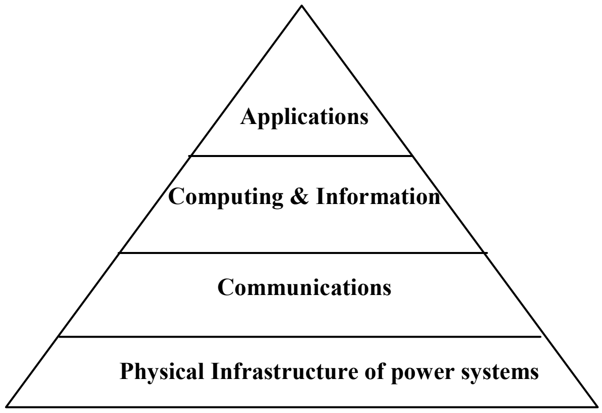

Figure 1. Smart Grid Infrastructure.

Many companies have implemented Smart Metering Systems as a first step in Smart Grids. The deployment of Smart Meter Systems begins with selection of the technology and the planning for installation, operation and maintenance. These systems represent a further step in the relationship between the electricity companies and its consumers. Coupled with an Advanced Metering Infrastructure (AMI), Smart Metering Systems could allow electric distribution companies monitoring the electrical power conditions practically in every point of the network [2], [4]-[6].

Generally, the categories of non-residential and small and medium residential consumers do not have a communication smart meter which to carry out a continuous monitoring. These consumers are monitored by conventional meters and load curves are raised only by campaigns measurements at certain time periods of the year. Concerning larger customers, these have available smart meters because their billing is done at the end of each month, the energy consumption is high and a detailed record of consumption is necessary for the choice of a tariff structure. [7]-[11].

With increasing installations of smart meters, massive amount of residential electric energy consumption data is collected and stored. Exploration of such data emerges as a research direction both in academic and industry sector, such as load aggregation and disaggregation (LD), load forecasting (LF) and demand response (DR) support, are brought back to attention. Using information obtained from smart meters, it is

978-1-4799-5849-8/14/$31.00 ©2014 IEEE

possible to improve the operation and planning of electric distribution systems [12].

The paper presents a decision trees based approach which allows identifying different categories of consumers characterized by the typical load profiles.

For determination of the consumer categories, every customer must be characterized by the following primary information: daily (monthly) energy consumption, minimum and maximum loads. The information can be obtained from load profiles of customers, provided by smart meters. The typical load profiles of the consumers were obtained using k-means clustering method.

II. Decision Trees

A decision tree is a classifier capable of extracting and synthesizes information contained in a database. The widespread use of decision trees in solving various problems due to the ease with which they can be implemented, the ability to generate predictive models that lead to errors very small, the internal structure easy to understand and explain. Thus, based on these features decision trees have become increasingly known. [13], [14].

A simple definition of a Decision Tree is: “A flow chart or diagram representing a classification system or a predictive model” [15]. In Classification and Regression Trees (CART) method, the starting point is represented by a set of objects (elements) described through attributes (features). Every attribute signifies an important characteristic of the object. Each object belongs to one class from a set of classes. The decision tree will be built on a training set which consists of preclassified examples. Each of these examples is fully described by a set of attributes and a class label.

A decision tree contains a root node, one or more internal nodes and one or more leaves nodes (terminal nodes), Fig. 2. If training set contains examples belonging to a single class, the root node will be a leaf. Otherwise, the root node is an internal node with the property that it contains the whole training set. Classification rules for each node are obtained through a mathematical process which aims to minimize impurity resulting nodes, with the help of available training set. Thus, evaluating unbundling rules will lead to a final node [16], [17].

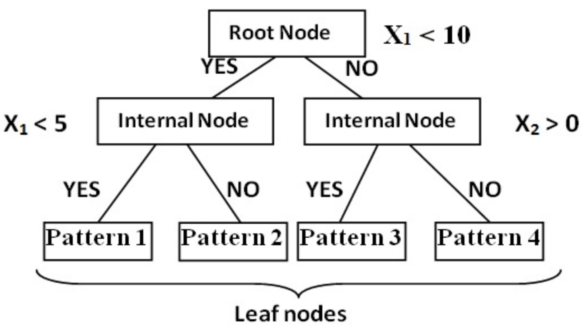

Figure 2. The structure of a decision tree.

All internal nodes have two or more successors and also contain divisions that test the value of a logical or mathematical expression associated some attributes. The branches from an internal node to its successors are labelled with the values of resulting from test associated the respective node, using the labelling convention "YES" of branch toward the left successor and "NO" of branch toward the right successor. Each leaf of the tree has a label associated with a class.

Decision tree generation is represented by two steps: training and classification. In training step of a decision tree, a part of the available data will be saved to assess the capability of generalization of the decision tree. The test data set is formed by saved data, while the training set is represented by the portion of the available data used to build the tree [18].

For to generate the decision tree, the Classification and Regression Trees (CART) method [15], [18], [25] is used in the paper. CART is a classification method which uses historical data to construct so-called decision trees.

III. Algorithm for Customer Classification

An important research topic used in optimal operation and planning of distribution networks by electric companies refers at representative consumption categories of customers. This is corresponding particularly to small residential consumers, whose energy consumption is not usually measured with smart meters. Because of the large number of small customers, sampling is the only possible way to collect data. The problem is how the selection and the analysis of the sample of customers should be made to finally get the most accurate load estimates for practical network calculations.

In this context, the electric companies require a set of load curves to represent all consumption classes. Deciding the optimal number of classes and the type of load curve for one class is a complicated problem. The practical criteria for determination of the consumption classes can be [8], [19], [20], [24]:

- load curve variance in one class customer class should be as small as possible;
- number of classes should be representative;
- classes should be easily linked with the database of electric company.

In a first step, the database will be divided into consumption categories taking into account the type of consumer activity: residential, commercial, industrial, etc, [8], [19], [21]. Inside each category, a refined classification can be identified by taking into account the daily energy consumption, minimum and maximum loads.

Finally the typical load curves, corresponding to consumption categories are determined.

The proposed methodology is based on the structure of the Data Mining in large databases based on Clustering Techniques [8] and decision making using decision trees.

The proposed algorithm has the following steps, Fig. 3:

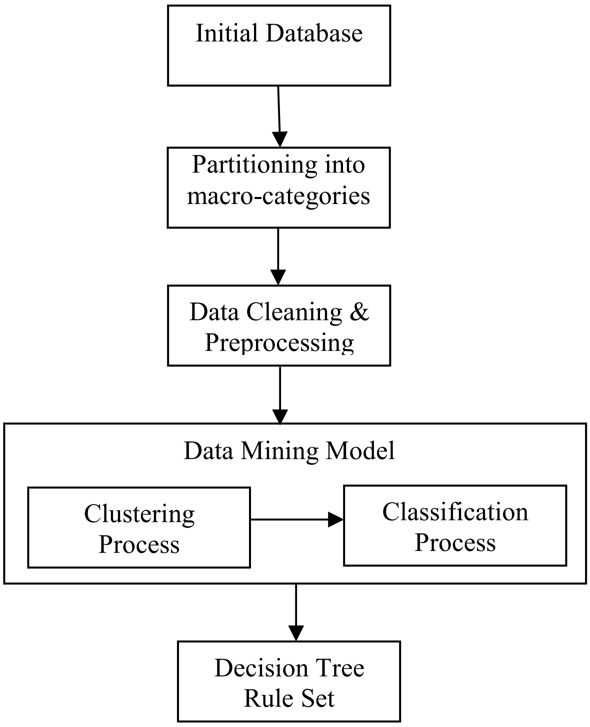

Figure 3. Flow-chart of the proposed method.

Step 1. Database: In this step, a consumer' representative sample will be select from the database. For this, the most important primary characteristics in terms of the consumption energy are identified. These customers have installed the smart meters. The consumers' load profiles, respectively variables that characterize the energy consumption will be recorded in a database. These variables are referred at: the daily energy, minimum and maximum active powers, and consumption category.

Step 2. Partitioning into macro-categories: The database is divided into consumption categories taking into account the type of consumer activity: residential, commercial, industrial.

Step 3. Data cleaning and Pre-processing: In this step, all records that contain missing data or outliers will be excluded. These situations are encountered in practice because the measurements are made for a large number of customers, spread over a wide geographical area, and problems with communication, interruption, failure of equipment or irregular atypical behaviour of consumers can appear. After being cleaned, pre-processed and reduced, the data are used to obtain the division in classes.

Step 4. Data Mining Model: Inside each consumption category a classification into more clusters (classes), in function by the daily energy consumption, minimum and maximum loads of the consumers is done. The classification will be based on a clustering method, namely the k-Means [22]. After obtaining consumer classes, a typical load profile will determine for every class. This will be achieved by averaging the values for each hour of all load profiles of customers from each class.

Step 5. Assignment: In this step, a decision tree is built that explains the patterns of the load profiles attending to the values of the variables determined in the first step. Taking as input the estimations of the input variables (daily energy consumption, minimum and maximum active powers), the decision tree is applied to obtain for customers a typical load profile according to their consumption class.

IV. CASE STUDY

In the case study, a database described by 230 load curves corresponding to the small residential consumers from a rural distribution network from Romania was considered. The measurements of load curves were performed using smart meters. Every measurement is represented by a curve of 24 hourly points describing the behaviour of a consumer during a day. For each consumer, the following consumption data are also known: the daily energy consumption (W), minimum (Pmin) and maximum (Pmax) loads. The consumers with missing values, outliers or energy consumption equal with zero were excluded. In the pre-processing, these consumers were excluded. For clustering process, only 216 consumers were eligible.

Then, using the K-means clustering algorithm, the consumers were grouped in representative clusters. The clusters obtained are represented in Fig. 4 and their characteristics are presented Table I.

TABLE I
THE CHARACTERISTICS OF CLUSTERS

Cluster	No. of customers	Pmax [kW]		Pmin [kW]		W [kWh]	
		m	σ	m	σ	m	σ
1	17	0.23	0.05	0.10	0.03	3.48	0.86
2	6	0.70	0.09	0.03	0.025	4.44	1.31
3	29	0.37	0.08	0.04	0.02	3.84	0.71
4	52	0.06	0.05	0.00	0.00	0.42	0.34
5	112	0.18	0.05	0.02	0.015	1.86	0.51

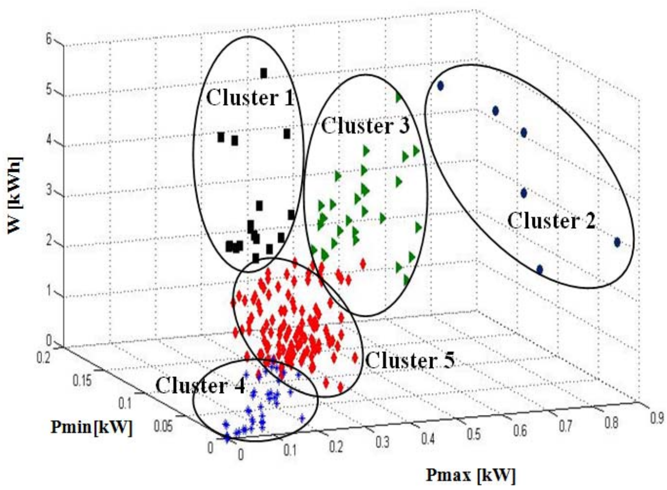

Figure 4. The representation of the clusters.

The values of these characteristics are the medium values obtained through averaging inside each cluster.

The next step includes the normalization of load profiles from each consumption category relatively to daily energy consumption.

Then, for each hour (h = 1, ..., 24), normalized average values are determined for each class of consumers. These values lead to the type of load profiles. The typical load profile for each class from the consumers' categories is obtained by representation of these coefficients, Figs. 5 – 9.

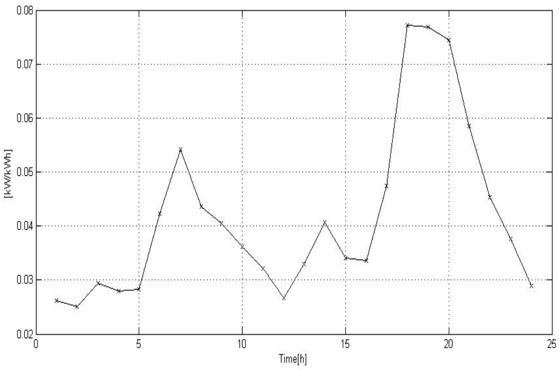

Figure 5. Typical load profile for Cluster 1.

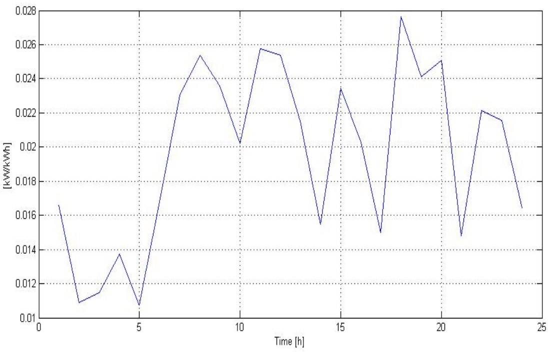

Figure 6. Typical load profile for Cluster 2.

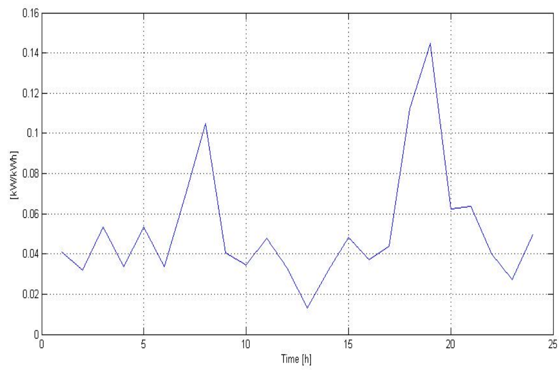

Figure 7. Typical load profile for Cluster 3.

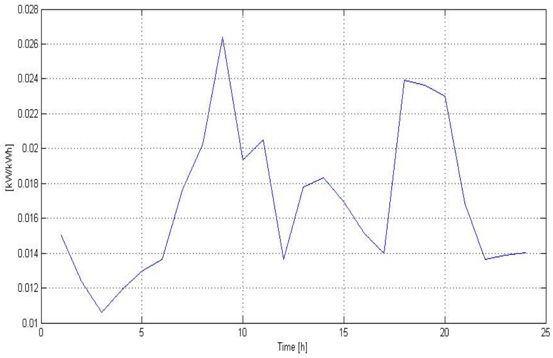

Figure 8. Typical load profile for Cluster 4.

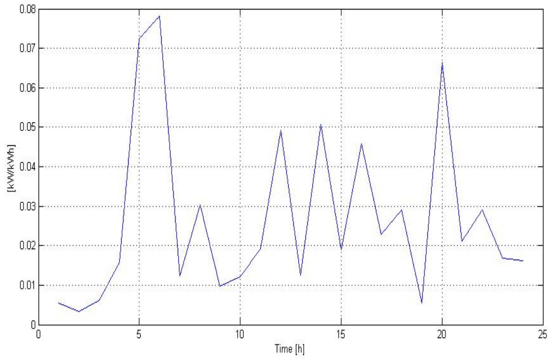

Figure 9. Typical load profile for Cluster 5.

The differences between the shapes of the typical load profiles generated are given by two factors: customer factor (type of consumption, electric heating, size of building etc) and climate factor (temperature, humidity etc).

Further, in order to build decision tree using CART method, a number of patterns must be established. In our case, the number of pattern is obtained using clustering process, namely $K = 5$. The model uses in the training step 70 % of the database (150 customers) and in the testing step 30% of the database (66 customers).

Fig. 10 shows the decision tree obtained in Matlab for learning database. The internal nodes represent test/decisions on one or more variables and the leaf nodes represent decision outcomes.

Further, this structure of decision tree was used for test database. The results corresponding to the classification process of customers from test database are presented as a confusion matrix, Table II.

Information about actual and predicted classifications obtained in classification process is presented in Table II.

The results obtained show that 1 customer from Cluster 3, 2 customers from Cluster 4, and 1 customer from Cluster 5 were misclassified. The test database does not contain customers belonging Cluster 2. The last column presents the rate of cases correctly classified to each cluster. The proportion of customers correctly classified is 0.939.

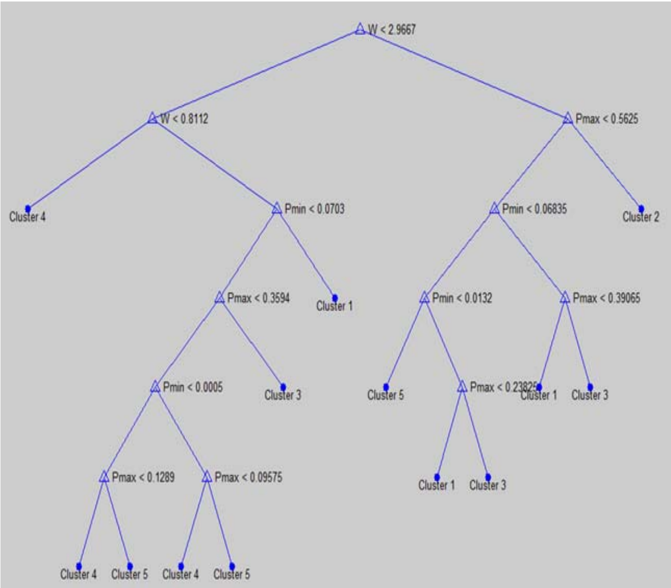

Figure 10. Decision tree for customer classification.

TABLE II
THE RESULTS OBTAINED IN CLASSIFICATION PROCESS (TEST DATABASE)

Actual		Classified with Decision Tree					Correct Rate
		C1	C2	C3	C4	C5	
C1	3	3	0	0	0	0	1.000
C2	0	-	-	-	-	-	-
C3	10	1	9	0	0	0	0.900
C4	17	0	0	15	0	2	0.882
C5	36	1	0	0	0	35	0.972

V. CONCLUSIONS

In this paper a decision tree based approach was proposed for determination of the customer classes using the information provided by smart meters. Based on a set of input data processed through clustering techniques a decision tree was built, through which can be classify with good accuracy customers, depending on input variables (the daily energy consumption, minimum and maximum loads). For to build the decision tree, the Classification and Regression Trees (CART) method is used. A number of patterns was be established using k-Means clustering algorithm, namely $K = 5$. The proportion of customers from test database correctly classified is 0.939.

The results obtained demonstrated that the proposed approach can be used with success in the operation of distribution systems when information about the supplied customers is very poor (based only the data provided by classic meters).

REFERENCES

[1] D. Novosel, "Experiences with Deployment of Smart Grid Projects," Proc. of Innovative Smart Grid Technologies (ISGT), Washington, DC, USA, 2012.

[2] D. R. V. Leite, H. Lamin, J. M. C. de Albuquerque, and I. M. T. Camargo, "Regulatory Impact Analysis of Smart Meters Implementation in Brazil," Proc.of Innovative Smart Grid Technologies (ISGT), Washington, DC, USA, 2012.

[3] T. Taylor and M. Ohrn, "Network management for smart grids. Innovative operations centers to manage future distribution networks", ABB Review, vol. 3, 2009, pp. 45 – 49.

[4] J. Pithart, "Integration and utilization of the smart grid," [Online]. Available: http://www.feec.vutbr.cz/EEICT/2009/sbornik/.

[5] P. Deng and L. Yang A., "Secure and Privacy-Preserving Communication Scheme for Advanced Metering Infrastructure," Proc. of Innovative Smart Grid Technologies (ISGT), Washington, DC, USA, 2012.

[6] R. Piacentini, "Modernizing Power Grids with Distributed Intelligence and Smart Grid-Ready Instrumentation," Proc. of Innovative Smart Grid *Technologies (ISGT)*, Washington, DC, USA, 2012.

[7] G. Hebrail, "Practical Data Mining in Large Utility Company," [Online], Available:: www.imfres.enst.fr /~hebrail / publications/ hdr / Compstat_200.pdf.

[8] G. Grigoras and G. Cartina, "Improved Fuzzy Load Models by Clustering Techniques in Distribution Network Control", International Journal on Electrical Engineering and Informatics, vol. 3, no. 2, 2011, pp. 205-2014.

[9] G. Grigoras and E.C. Bobric, "Clustering based approach for customers' classification from electrical distribution systems," Proc. of International Conference on Energy and Environment, Bucharest, Romania, 2013.

[10] A. Mutanen, M. Ruska, S. Repo, and P. Järventausta, "Customer Classifcation and Load Profilling Method for Distribution Systems", IEEE Trans. on Power Delivery, vol. 26, no. 3, pp. 1755 – 1763, 2011.

[11] J. N. Fidalgoa, M. A. Matosb, L. Ribeiroc, "A new clustering algorithm for load profiling based on billing data," Electric Power Systems Research vol. 82, pp. 27– 33, 2012.

[12] Z. Guo, Z. J. Wang, and A. Kashani, "Home Appliance Load Modeling from Aggregated Smart Meter Data," IEEE Transactions on Power Systems, 2014, in press.

[13] L. Rokach,"Top-down Induction of Decision Trees Classifiers - A Survey", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 35, no. 4 pp. 476 – 487, 2005.

[14] Z. Yu, F. Haghighat, B.C.M. Fung, and Y. Hiroshi, "A Decision Tree Method for Building Energy Demand Modeling," Energy and Buildings, vol. 42, pp. 1637–1646, 2010.

[15] Terri Moore, Carole Jesse, and Richard Kittler, "An Overview and Evaluation of Decision Tree Methodology", ASA Quality and Productivity Conference, Austin, USA, May 23-25, 2001.

[16] E. Lobato, A. Ugedo, L. Rouco and F.M. Echavarren, "Decision Trees Applied to Spanish Power Systems Applications", in Proc. 9th International Conference on Probabilistic Methods Applied to Power Systems, Stockholm, Sweden, 2006.

[17] C. Olaru, and L. Wehenkel, "A Complete Fuzzy Decision Tree Technique," Fuzzy Sets and Systems, vol.138, pp. 221–254, 2003.

[18] R. Timofeev, Classification and Regression trees (CART). Theory and Applications, Master's Thesis, Humboldt University Berlin, 2004.

[19] C. Liua, Z. H. Rathera, Z. Chena, and C. L. Baka, "An Overview of Decision Tree Applied to Power Systems", International Journal of Smart Grid and Clean Energy, vol. 2, no. 3, October 2013.

[20] M. Eremia, G. Cartina, D. Petricica, A.I. Bulac, C. Bulac, I. Tristiu, G. Grigoras, Artificial Intelligence Techniques in Power Systems Operation, AGIR Publishing, Romania, 2006.

[21] A. Jardini, C. Tahan, M.R. Gouvea, S.U. Ahn, F.M. Figueiredo, "Daily Load Profiles for Residential, Commercial and Industrial Low Voltage Consumers", in IEEE Transactions on Power Delivery, vol. 15, no. 1, Jan. 2000, pp. 375 – 380.

[22] G. Chicco, R. Napoli, F. Piglione, "Comparison among Clustering Techniques for Electricity Customer Classification", in IEEE Trans. On Power Systems, vol. 21, no. 2, 2006, pp. 933-940.

[23] A.K. Jain, M.N. Murty, P.J. Flynn, "Data Clustering: A Review", [Online], Available: http://cermics.enpc.fr/ ~keriven/ vision/articles.

[24] G. Grigoras and Fl. Scarlatache, "Use of data from smart meters in optimal operation of distribution systems", 14th International Conference on Optimization of Electrical and Electronic Equipment, OPTIM 2014, May 22-24, 2014, Brasov, Romania

L. Gordon, "Using Classification and Regression Trees (CART) in SAS Entreprise Miner for Applications in Public Health", SAS Global Forum, 2013, San Francisco, April 28 – May 1, 2013.