DATA MINING APPLICATION: CLASSIFICATION OF LOAD PATTERN ANALYSIS FOR ELECTRICITY CUSTOMERS

Rupali Meshram*

Prof. A. V. Deorankar**

Dr. P. N. Chatur***

Abstract

This paper deals with the wide range of models (Bayesian model , Support Vector Machines , Neural network) for forecasting the electricity. Electricity load forecasting has been object of vast research since energy load is known to be non-linear and, therefore, very difficult to predict with accuracy. Load forecasting is necessary for economic generation of power. Classification of load pattern is an important task for load forecasting of customers and grouping them into classes according to their load characteristics. The different clustering algorithms (modified follow-the-leader, k-means, fuzzy k-means and two types of hierarchical clustering) and the Self Organising Map to group together customers having a similar electrical behaviour. In the approach, all load curves of customers are first clustered with the clustering algorithms under a given number of clusters. This paper shows the study on Bayesian model, Support Vector Machines, Neural network and k-means algorithm.

Keywords: Classification, Load Pattern Analysis, Clustering, Load forecasting, Typical Load Profile.

* M.Tech 2nd year, Department of Computer Science and Engineering, Government College of Engineering, Amravati (Maharashtra), India.

** Asst. Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati (Maharashtra), India.

*** Head of Department, Department of Computer Science and Engineering, Government College of Engineering, Amravati (Maharashtra), India.

Introduction

Moreover, energy cannot be stored and the power grid must maintain a balance on real-time between the amount electricity produced and consumed. Otherwise, the risk of a blackout cannot be overseen, both caused by defect or excess of energy in the grid (i.e. production does not equal consumption). In this way, if demand of energy stochastic , all participants in this new scenario must deal with the problem of forecasting at least, on the short-run, the amount of energy they will have to generate, transport, distribute, and so on.

In recent years, due to the rapid growth of economy and living standards, peak loads have grown faster than the average loads, the distribution side of the electricity industry has to face new challenges in providing satisfactory service to customers. Along with 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 aim is to classify the load pattern of different types of customers. Conducting load pattern analysis is an important task in obtaining typical load profiles (TLPs) of customers and grouping them into classes according to their load characteristics. When using clustering techniques to obtain the load patterns of electricity customers, choosing a suitable clustering algorithm and determining an appropriate cluster number are always important and difficult issues. Identifying the consumption patterns of the customers and grouping together customers having similar patterns may be significantly more helpful. The system load is a random non-stationary process composed of thousands of individual components. The system load behaviour is influenced by a number of factors, which can be classified as: economic factors, time, day, season, weather and random effects. The objective is utilization of electricity, developing Tariff on different types of electricity customers, Selection of generators. In the approach, all load curves of customers are first clustered with the clustering algorithms under a serial given number of clusters. Clustering methods that can be used for cluster analysis and clustering method may identify groups whose member objects are different. Implementation of load profile , classification of load profile and the typical load profile generation of large electricity customers. This includes such characteristics as average load factor, utilization factor, and responsibility factor. These all can be calculated based on a given load profile. Conducting load pattern analysis is an important task

in obtaining typical load profiles of customers and grouping them into classes according to their load characteristics. When we are using clustering techniques to obtain the load patterns of electricity customers, choosing a suitable clustering algorithm and determining an appropriate cluster number are always important and difficult issues.

It is of common-knowledge that progress affects all aspects of our lives. This fact has become a painful truth when it comes to the way we consume energy: just think of the new gadgets we have acquired in the last 5 years. Indeed, the Malthusian increment worldwide of energy consumption per capita has drawn a new scenario in which the classical dramatis personae have been altered and, moreover, old actors must play new roles. This process, widely known as liberalisation of the energy markets, consists in the separation of electricity generation and retail from the natural monopoly functions of transmission and distribution [23]. In the former case, competing generators offer their electricity output to retailers and, in the latter, end-use customers choose their supplier from competing electricity retailers. Please note that either markets differ from their more traditional counterparts because energy cannot be stored. Consequently, all players are forced to work with consumption prognoses, which as one may think, creates a number of risks. In this scenario, last-mile electricity customers now have the possibility of choosing their retailer: selecting the most convenient one or, directly going for the worst, will definitively make a difference on the energy bill. Moreover, finding a suitable retailer is not an easy task due to many reasons and this aspect has drawn quite an static electricity market. Furthermore, not only energy customers profit from short term load forecasting tools; all participants in the electric system do. For instance, since the balance between generation and consumption must be watched out constantly in the power grid, Transmission System Operators (TSOs) work with global demand prognoses. Any deviation implies a cost because the consumption is not being managed efficiently. Prediction of the demand from the clients' side may help reduce these deviations, reducing the overcosts the TSO must face. Regarding retailers, they always work with client portfolios and being able to foresee their consumption at the short term enables them to buy more accurately what they need.

Load forecasting is a central and integral process in the planning and operation of electric utilities. It involves the accurate prediction of both the magnitudes and geographical locations of electric load over the different periods (usually hours) of the planning horizon. The basic

quantity of interest in load forecasting is typically the hourly total system load. However, according to Gross and Galiana (1987), load forecasting is also concerned with the prediction of hourly, daily, weekly and monthly values of the system load, peak system load and the system energy. Srinivasan and Lee (1995) classified load forecasting in terms of the planning horizon's duration: up to 1 day for short-term load forecasting (STLF), 1 day to 1 year for medium-term load forecasting (MTLF), and 1±10 years for long-term load forecasting (LTLF).

Many methods or techniques for clustering load curves have been proposed. Some clustering methods are: k-means [12], [13], modified follow-the-leader [4], [8], average and Ward hierarchical methods [9], fuzzy c-means (FCM) [13], statistic-fuzzy technique [14], the self-organizing map (SOM) [13], [15], support vector machines (SVM) [2], and extreme learning machine [1]. Some hybrid techniques [16] have been proposed to improve the clustering effect.

The purpose of this paper is to survey and classify electric load classification techniques. Researchers has worked on to decide which clustering algorithm is best for classification of load pattern analysis for different types of electricity customers. They compared most of the clustering algorithms. This paper shows the various clustering models which are used to classify the load pattern of electricity customers.

I. Classification and Clustering Process

Web The classification of different types of electricity customers are achieved by applying clustering techniques, fig.1 shows the flow chart of classification and load profile generation of large electricity customers which include the following basic steps:

A. Data Selection

The power consumption data of customers can be recorded by an automatic meter reading system with time periods in steps of 15 min, 30 min, or 1 h. A preliminary data selection of customers can be carried out by geographical region and voltage level (high, medium, and low). The daily chronological load curves for each individual customer are determined for each study period (month, season, and year).

B. Data Cleaning

The load curves of each customer are examined for normality, in order to modify or delete the values that are obviously wrong (noise suppression). For example, we remove those daily load

curves with 0 MW values and unreasonable load curves for known reasons (such as network failure or meter error).

Data cleaning routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Dirty data can cause confusion for the mining procedure. Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled.

D. Data Preprocessing

Clustering load curves are based on the shape of a load curve but not by absolute MW values, so the data should be normalized. Normalization is particularly useful for classification algorithms involving distance measurements. The different methods for data normalization are: min-max normalization, z-score normalization, and normalization by decimal scaling. In this paper, the data is normalized by z-score normalization [10]. The data were normalized in the range of (0, 1) by using as the normalizing factor the peak value.

E. Load Curve Clustering

Various clustering algorithms are used to cluster the normalized load curves. Based on our observations that under a given number of clusters, the clustering results obtained from a multiple-run in most cases are different when using the same algorithm, the difference can be shown by the number of groups of the load pattern results.

F. Customer Classification

Each load pattern contains a certain number of customers, no "empty" patterns exist. The customer classes can be obtained according to the load patterns. The typical load pattern for each customer can then be generated by the load curves belonging to the same load pattern; each typical load pattern is a centroid curve of a cluster of load curves connected with a load pattern.

II. Clustering Models

Bayesian model (BN): Bayesian Networks (BN) are probabilistic models for multivariate analysis that extend the Bayes' theorem [19]. BN are a very popular solution when tackling a problem that requires predicting the outcome of a variable according to the value that other variables take such as in weather forecasting or spam detection. More specifically, BNs combine an acyclic directed graph with a probability distribution functions [20]: the graphical model

represents the set of probabilistic relationships among the collection of variables modelling the specific problem, whereas the probability function illustrates on each node the strength of these relationships or edges in the graph. The research on BNs has mainly focused on systems with discrete variables, linear Gaussian models or combinations of both since, except for linear models, continuous variables pose a problem for Bayesian networks due to the inherent difficulty of representing a continuous quantity by an estimated magnitude and a range of uncertainty.

Bayesian model selection uses the rules of probability theory to select among different hypotheses. It is completely analogous to Bayesian classification. It automatically encodes a preference for simpler, more constrained models, as illustrated at right. Bayesian network (BN) approach to segmentation. Bayesian networks are also referred to as belief networks, probabilistic networks, probabilistic belief networks (PBN), and probabilistic causal networks. The Bayesian network model makes decisions about how to interpret probabilistic evidence (i.e., non-deterministic information) to support or reject a hypothesis or outcome. Outcomes that yield the highest expected utility are chosen as the optimal solutions. With the Bayesian network model, the definition of expected utility incorporates all the probabilistic uncertainty associated with the outcome, as well as the inherent utility of the outcome. Utility can be defined in dimensions such as monetary cost, entropy, or energy.

Support Vector Machines (SVM): SVM constructs a hyperplane or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks. SVM have been used for load forecasting in buildings ([21]). In essence, an SVM is a mathematical entity, an algorithm for maximizing a particular mathematical function with respect to a given collection of data. The basic ideas behind the SVM algorithm, however, can be explained without ever reading an equation. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships.

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging

to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Neural network (NN): NNs are non-linear circuits whose perceptron (say simple information processors) structure adapts according to the external or internal information that flows through the network during the learning phase[22]. Their output is a linear or non-linear function of the inputs and therefore, they have been widely used for predicting non-linear data. Many NN models are similar or identical to popular statistical techniques such as generalized linear models, polynomial regression, nonparametric regression and discriminant analysis, projection pursuit regression, principal components, and cluster analysis, especially where the emphasis is on prediction of complicated phenomena rather than on explanation. These NN models can be very useful. There are also a few NN models, such as counter propagation, learning vector quantization, and self-organizing maps, that have no precise statistical equivalent but may be useful for data analysis. Fig. 1 show the Neural network model for four inputs. Fig. 1 contains four input ,3 hidden layers and 2 outputs.

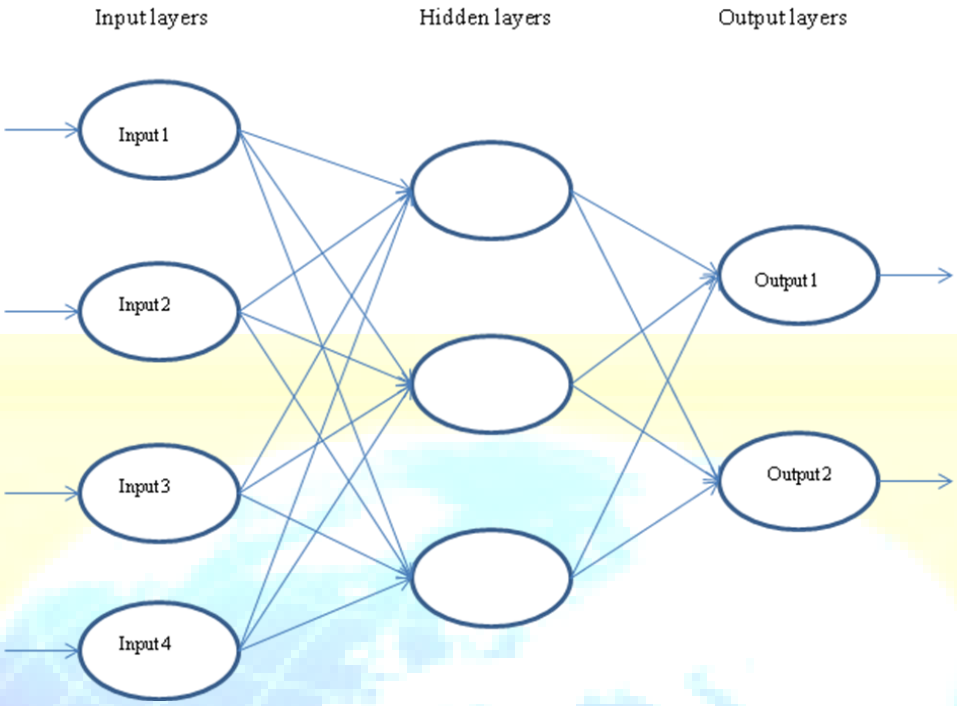

Fig. 1: Neural network Model

NN researchers routinely reinvent methods that have been known in the statistical or mathematical literature for decades or centuries, but they often fail to understand how these methods work. The common implementations of NNs are based on biological or engineering criteria, such as how easy it is to fit the net on a chip, rather than on well-established statistical and optimization criteria. Good performance (e.g. as measured by good predictive ability, low generalization error), or performance mimicking animal or human error patterns, can then be used as one source of evidence towards supporting the hypothesis that the abstraction really captured something important from the point of view of information processing in the brain. Another incentive for these abstractions is to reduce the amount of computation required to simulate artificial neural networks, so as to allow one to experiment with larger networks and train them on larger data sets.

III. Proposed approach

Difficult to select number of clusters for huge amount of electricity data. In supervised algorithm, if the number of cluster size is less then it improve the classification rate. In this work, we focus on the k-means unsupervised clustering algorithm and support vector supervised clustering algorithm.

The classical k-means clustering [12] groups a data set of $\mathbf{x}^{(n)}$ (n = 1, …, N) samples in k = 1, …, K clusters by means of an iterative procedure. A first guess is made for the K cluster centres $\mathbf{c}^{(k)}$ (usually chosen in a random fashion among the samples of the data set). The K centres classify the samples in the sense that the sample $\mathbf{x}^{(n)}$ belongs to cluster k if the distance $\| \mathbf{x}^{(n)} - \mathbf{c}^{(k)} \|$ is the minimum of all the K distances. The estimated centres are used to classify the samples into clusters (usually by Euclidean norm) and their values $\mathbf{c}^{(k)}$ are recalculated. The procedure is repeated until stabilisation of the cluster centres. Clearly, the optimal number of clusters is not known a priori and the clustering quality depends on the value of K. To improve the k-means algorithm include the concept of modified follow-the-leader. In modified follow-the-leader The follow-the-leader[4] 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 of clusters and the number of patterns belonging to each cluster 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.

Support vector clustering is based on the support vector approach introduced in [19]. The application of the SVC method does not require preliminary assumptions on the number of clusters, and operates with any shape of the data patterns. The SVC method is composed of two stages. The first stage is dedicated to the determination of the support vectors. The second stage handles the results obtained at the first stage to form the final clusters. The SVC method is able to provide both a suitable retrieval of the outliers and a meaningful grouping of the remaining load patterns into classes. The combination of unsupervised and supervised algorithm help to improve the classification of load profile of electricity customers.

IV. Conclusion

In this work, we study the three types of clustering models for classifying a load pattern of different types of electricity customers. The paper presents the procedure how to determine the typical load profile based on the clustering methods. Different normalization methods may cause different clustering results. This all models (Bayesian model, Support Vector Machines, Neural network) used for forecasting the electricity. The rapidly increasing power of the personal

computer is making it possible to apply more complicated solution techniques. New load forecasting methods based on fuzzy logic, genetic algorithms, expert systems, and neural networks offer new hopes in this direction of research. Over the last few years, the most active research area has been neural network based load forecasting. Further research will improve the clustering by adding two clustering algorithms.

Reference

[1] A. H. Nizar , Z. Y. Dong , and Y. Wang, "Power utility nontechnical loss analysis with extreme learning machine method," IEEE Trans. Power Syst., vol. 23, no. 3, pp. 946–955, Aug. 2008.

[2] G. Chicco and I. S. Ilie, "Support vector clustering of electrical load pattern data," IEEE Trans. Power Syst., vol. 24, no. 3, pp. 1619–1628, Aug. 2009.

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

[4] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," IEEE Trans. Power Syst., vol. 21, no. 2, pp. 933–940, May 2006.

[5] Hong-Tzer Yang, Shih-Chieh Chen, and Win-Ni Tsai, "Classification of Direct Load Control Curves for Performance Evaluation", IEEE Trans. Power Syst., vol 19, no. 2, pp. 811–817, May 2004.

[6] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas, "Two-stage pattern recognition of load curves for classification of electricity customers," IEEE Trans. Power Syst., vol. 22, no. 3, pp. 1120–1128, Aug. 2007.

[7] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia, "An electric energy consumer characterization framework based on data mining techniques," IEEE Trans. Power Syst., vol. 20, no. 2, pp. 596–602, May 2005.

[8] G. Chicco, R. Napoli, and F. Piglione, "Application of clustering algorithms and self organising maps to classify electricity customers," in Proc. IEEE Bologna Power Tech Conf., Jun. 23–26, 2003, vol. 1.

[9] N. M. Kohan, M. P. Moghaddam, S. M. Bidaki, and G. R. Yousefi, "Comparison of modified k-means and hierarchical algorithms in customers load curves clustering for designing suitable tariffs in electricity market," in Proc. 43rd Int. Universities Power Engineering Conf., Padova, Italy, Sep. 1–4, 2008, pp. 1–5.

[10] J. Han and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2006.

[11] Tiefeng Zhang, Guangquan Zhang, Jie Lu , Xiaopu eng, and Wanchun Yang, "A New Index and Classification Approach for Load Pattern Analysis of Large Electricity Customers," IEEE Trans. Power Syst., vol. 27, NO. 1, Feb. 2012

[12] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," J. Roy. Statist. Soc. Series C (Appl. Statist.), vol. 28, no. 1, pp. 100–108, 1979.

[13] 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.

[14] W. Y. Li, J. Q. Zhou, X. F. Xiong, and J. P. Lu, "A statistic-fuzzy technique for clustering load curves," IEEE Trans. Power Syst., vol. 22, no. 2, pp. 890–891, May 2007.

[15] F. Rodrigues, J. Duarte, V. Figueiredo, Z. Vale, and M. Cordeiro, "A comparative analysis of clustering algorithms applied to load profiling," Mach. Learn. Data Mining Pattern Recognit., Lecture Notes Comput. Sci., vol. 2734, no. 2003, pp. 73–85, 2003.

[16] S. C. Cerchiari, A. Teurya, J. O. P. Pinto, G. Lambert-Torres, L. Sauer, and E. H. Zorzate, "Data mining in distribution consumer database using rough sets and self-organizing maps," in Proc. IEEE/PES Power Systems Conf. Expo., Atlanta, GA, Oct. 29–Nov. 1 2006, pp. 38–43.

[17] D. Hand, H. Manilla, and P. Smyth, Principles of Data Mining. Cambridge, MA: MIT Press, 2001.

[18] V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer,1995.

[19] T. Bayes, "An essay towards solving a problem in the doctrine of chances," Philosophical Transactions of the Royal Society, vol. 53, pp. 370–418, 1763.

[20] E. Castillo, J. Guti´errez, and A. Hadi, Expert systems and probabilistic network models. Springer-Verlag, 1997.

[21] B. Dong, C. Cao, and S. Lee, "Applying support vector machines to predict building energy consumption in tropical region," Energy and Buildings, vol. 37, no. 5, pp. 545–553, 2005.

[22] H. Alfares and M. Nazeeruddin, "Electric load forecasting: literature survey and classification of methods," International Journal of Systems Science, vol. 33, no. 1, pp. 23–34, 2002.

[23] Y. Penya, Optimal algorithms for energy markets: Allocation and Scheduling of Demand in Deregulated Energy Markets. Verlag Dr. Mueller, 2008.