Energy Conversion and Management 103 (2015) 1040–1051

Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Day-ahead load forecast using random forest and expert input selection

A. Lahouar [a,b,*], J. Ben Hadj Slama [b]

[a] National Engineering School of Tunis, University of Tunis El Manar, BP 37, Le Belvedere, Tunis 1002, Tunisia
[b] Advanced Systems in Electrical Engineering, National Engineering School of Sousse, University of Sousse, BP 264, Erriadh, Sousse 4023, Tunisia

A R T I C L E I N F O

Article history:
Received 8 April 2015
Accepted 16 July 2015
Available online 28 July 2015

Keywords:
Short term load forecast
Random forest
Expert input selection
Online learning
Variable importance
Prediction of holidays

A B S T R A C T

The electrical load forecast is getting more and more important in recent years due to the electricity market deregulation and integration of renewable resources. To overcome the incoming challenges and ensure accurate power prediction for different time horizons, sophisticated intelligent methods are elaborated. Utilization of intelligent forecast algorithms is among main characteristics of smart grids, and is an efficient tool to face uncertainty. Several crucial tasks of power operators such as load dispatch rely on the short term forecast, thus it should be as accurate as possible. To this end, this paper proposes a short term load predictor, able to forecast the next 24 h of load. Using random forest, characterized by immunity to parameter variations and internal cross validation, the model is constructed following an online learning process. The inputs are refined by expert feature selection using a set of if–then rules, in order to include the own user specifications about the country weather or market, and to generalize the forecast ability. The proposed approach is tested through a real historical set from the Tunisian Power Company, and the simulation shows accurate and satisfactory results for one day in advance, with an average error exceeding rarely 2.3%. The model is validated for regular working days and weekends, and special attention is paid to moving holidays, following non Gregorian calendar.

1. Introduction

Electrical load forecast is a challenging task for power operators. It is an old research theme that followed the evolution of power installations and computational techniques. Also, it is getting more important by the beginning of the 21th century due to the emergence of renewable energy resources and smart grids. Load forecast means the prediction of the future evolution of the electric load signal of an individual apartment, a local grid, a region or even a whole country. This prediction is performed for a period of time called forecast horizon through one or several time steps. An accurate power prediction of one or several hours ahead is very important to load dispatch, unit commitment and energy exchange decisions. Predicting load for larger time horizon is also useful for maintenance planning and energy management policies. Increasing the forecast error of 1% may entail a spectacular increase in operational costs [1], thus the tiniest improvement in accuracy is interesting. The error may be an underestimation or an overestimation of power, and both entail difficulties to balance supply and demand. The concept of smart grid implies utilization of intelligent computing techniques, including power forecast, to manage supply, in order to match the demand in real time. The supply management is closely related to spinning reserve, which is the total synchronized capacity, minus losses and load. Then, forecasting load means predicting the spinning reserve, which is very important in cases of sudden huge demand, outage or failure of some generators. When the forecast is accurate, the spinning reserve becomes ready to offset rapidly any deficiency. For longer horizons, the prediction of the load profile determines the amount of capacity to add to the overall network, in order to prevent any contingency.

Load forecast becomes harder than before due to two main reasons. First, privatization and deregulation of electricity market in many countries mean that energy consumers are free to choose their provider among several operators. Second, high penetration of intermittent resources in the grid, namely wind and solar energy, increases the degree of uncertainty due to their non-regular behavior. The deregulation of the market entails varying electricity price, which pushes customers to consume when the energy cost is low, and therefore new forecasting schemes are required [2]. In this case, price forecast is performed along with load forecast [3]. Having sparse customers is also a consequence of deregulation. All these aspects may lead to load curves that are non-smooth and poorly correlated with weather variables. This is not the case of the Tunisian electricity market, subject of

* Corresponding author at: National Engineering School of Tunis, University of Tunis El Manar, BP 37, Le Belvedere, Tunis 1002, Tunisia.

http://dx.doi.org/10.1016/j.enconman.2015.07.041
0196-8904/

this research, where privatization is still limited. In 2013, only 3046 GW h was generated by private companies, while 13,947 GW h was produced by the government corporation. The price is therefore fixed and the deregulation problem does not arise. However, intermittent sources are integrated in the grid by the presence of two wind farms. The wind power generation was 357.8 GW h in 2013, which means 2.6% of the total production. Despite this tiny contribution, intermittent energy may affect the grid even with low degree of penetration [4]. To this end, prediction of the wind generation should be carried out along with load forecast, and this was subject of a previous work for the Tunisian wind farms [5]. These two types of forecast, when coupled together, ensure better management of the residual load which must be covered by classical power plants. The small area of the country and the installed power limited to 4425 MW make it possible to forecast the entire country load demand as a whole. It is also possible to use the weather measurements of one station located in Tunis, the capital and the most consuming city, instead of considering many scattered weather stations [6,7].

In order to face the above mentioned challenges, this paper proposes a short term prediction model using the random forest technique. This model is designed for day-ahead prediction by a step of one hour, with respect to the specifications of the Tunisian power installation; such as small area, warm weather, no deregulation and presence of intermittent energy. The main contribution of this paper is to demonstrate the flexibility of random forests when associated with expert selection, to handle any load profile, and in particular to fit with complex customers behavior. The proposed approach shows high accuracy and effectiveness in the four seasons and for particular days, such as weekends and holidays, whether they are moving or not. The remainder of the paper is organized as follows: Section 2 gives a bibliographic overview of existing forecast methods, Section 3 details the necessary mathematical development, Section 4 analyzes the load profile and clarifies the prediction strategy, results and comparisons are given in Section 5 and finally, Section 6 concludes the paper.

2. Literature review

2.1. General approaches for short term forecast

Prediction is a regression analysis applied to time series, which means studying the relationship between several variables, namely future and past samples. The load signal is a time series, and a predictor should estimate its future evolution in terms of past samples and eventually some exogenous variables affecting the future load. Based on this concept, several forecast models were elaborated in the literature for the very short term; less than one hour, the short term; one hour to several days, the medium term; one to several months, and the long term; one to several years in advance. It is possible to classify approaches to conventional statistical methods, artificial intelligence methods and hybrid methods.

2.2. Conventional statistical methods

Statistical methods are white-box models where outputs are explicitly related to inputs through mathematical equations. This family of methods includes the simplest linear regression [8], multiple regression [9,10], the well-known Box–Jenkins models; autoregressive moving average (ARMA) [11] and autoregressive integrated moving average (ARIMA) [12,13], exponential smoothing (Holt–Winters) [14] and Kalman filter [15]. These methods are simple to implement and well adapted for the short term, but unable to handle the non-linearity existing in the load series. This fact pushes towards using intelligent methods.

2.3. Artificial intelligence methods

Artificial intelligence methods are black-box models where the internal dynamic is unknown. This family includes three main approaches, namely fuzzy inference (expert) systems (FIS) [16,17], artificial neural networks (ANN) [18–21] and support vector machines (SVM) [22,23,7]. Here, the relationship between inputs and outputs is determined through a set of linguistic rules for fuzzy systems or training process for learning machines (ANN and SVM). Apart from these three main approaches, little attention was paid to random forest (RF) [24,25], which is also a machine learning technique requiring a training phase.

These methods have the great advantage of non linear estimation. A three-layer neural network is able to achieve any accuracy of continuous function mapping [26]. However, ANN has problems of under-fitting and over-fitting, in addition to the local optimal solution. The SVM uses the empirical risk minimization principle, and overcomes the problems caused by ANN [27,26], which makes all its strength. Expert systems have the advantage of giving good interpretability of the system [28], while the main advantage of RF is low sensitivity to parameter values [24]. All those aspects make these methods very powerful, but they have their limitations such as optimal architecture and parameter tuning. These limits are surpassed by hybridization.

2.4. Hybrid methods and metaheuristics

Metaheuristics are stochastic algorithms that try to find a sufficiently good solution to a hard optimization problem, by sampling an objective function. They include evolutionary algorithms such as genetic algorithms (GA) and differential evolution (DE), as well as particle swarm optimization (PSO), ant colony (AC) and simulated annealing (SA). They are commonly used to tune ANN and SVM parameters or for training purpose. In a similar manner, signal processing techniques, especially wavelet transform (WT), are used in hybrid methods.

In general, almost all recent forecast techniques are combinations of the three main approaches and their derivatives, or hybridization with metaheuristics or signal processing techniques. For example, some derivatives of ANN are spiking [1], abductive [29], structural [30], Elman recursive [31] and generalized neural networks [32]. ANN may be hybridized with exponential smoothing [33], grey theory [34], wavelet transform [35,36] or evolutionary algorithms [37]. A class of ANN called self organizing map (SOM) or Kohonen network is also elaborated [28]. This wide range of applications makes the ANN the most commonly used method for load forecast. Likewise, SVM may be used along with ant colony [38], particle swarm [39], adaptive neuro fuzzy inference system (ANFIS) [40] and wavelet transform [41,3]. Fuzzy systems may also be combined with evolutionary algorithms [42–44]. Exponential smoothing, despite not among intelligent methods, may be combined with WT and achieve results as accurate as other hybrid methods [45].

Hybrid methods that combine artificial intelligence and metaheuristic optimization are the most effective and accurate approaches according to many researchers. Nevertheless, limitations always exist, such as consuming time and resources, varying accuracy according to the context and available data.

2.5. Medium and long term approaches

For longer time horizon forecast, literature is scarcer. The majority of researchers utilize hybrid ANN for both medium term [46,47,27,48] and long term [49,50]. Here, the factors driving electricity consumption are significantly different, such as electricity