Knowledge-Based Systems 82 (2015) 29–40

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Correlation and instance based feature selection for electricity load forecasting

Irena Koprinska [a,*], Mashud Rana [a], Vassilios G. Agelidis [b]

[a] School of Information Technologies, University of Sydney, Sydney, NSW 2006, Australia
[b] Australian Energy Research Institute, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 12 July 2014
Received in revised form 12 January 2015
Accepted 21 February 2015
Available online 28 February 2015

Keywords:
Electricity load forecasting
Feature selection
Autocorrelation
Mutual information
Linear regression
Neural networks

ABSTRACT

Appropriate feature (variable) selection is crucial for accurate forecasting. In this paper we consider the task of forecasting the future electricity load from a time series of previous electricity loads, recorded every 5 min. We propose a two-step approach that identifies a set of candidate features based on the data characteristics and then selects a subset of them using correlation and instance-based feature selection methods, applied in a systematic way. We evaluate the performance of four feature selection methods – one traditional (autocorrelation) and three advanced machine learning (mutual information, RReliefF and correlation-based), in conjunction with state-of-the-art prediction algorithms (neural networks, linear regression and model tree rules), using two years of Australian electricity load data. Our results show that all feature selection methods were able to identify small subsets of highly relevant features. The best two prediction models utilized instance and autocorrelation based feature selectors and an efficient neural network prediction algorithm. They were more accurate than advanced exponential smoothing prediction models, a typical industry model and other baselines used for comparison.

1. Introduction

Forecasting the future electricity load is an important task in the management of modern energy systems. It is used to make decisions about the commitment of generators, setting reserve requirements for security and scheduling maintenance. Its goal is to ensure reliable electricity supply while minimizing the operating cost.

Electricity load forecasting is classified into four types based on the forecasting horizon: long-term (years ahead), medium-term (months to a year ahead), short-term (1 day to weeks ahead) and very short-term (minutes and hours ahead). In this paper we consider Very Short-Term Load Forecasting (VSTLF), in particular 5 min ahead forecasting. VSTLF plays an important role in competitive energy markets such as the Australian national electricity market. It is used by the market operator to set the required demand and its price and by the market participants to prepare bids. The importance of VSTLF increases with the emergence of the smart grid technology as the demand response mechanism and the real time pricing require predictions at very short intervals [1].

Predicting the electricity load with high accuracy is a challenging task. The electricity load time series is complex and non-linear, with daily, weekly and annual cycles. It also contains random components due to fluctuations in the electricity usage of individual users, large industrial units with irregular hours of operation, special events and holidays and sudden weather changes.

Various approaches for VSTLF have been proposed; the most successful are based on Holt–Winters exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) [2], Linear Regression (LR) and Neural Networks (NNs) trained with the backpropagation algorithm [3–7]. The problem of feature selection for VSTLF, however, has not received enough attention, and it is the focus of this paper.

Feature (variable) selection is the process of selecting a set of representative features (variables) that are relevant and sufficient for building a prediction model. It has been an active research area in machine learning [8–10]. Good feature selection improves the predictive accuracy, leads to faster training and smaller complexity of the prediction model. It is considered as one of the key factors for successful prediction.

Most of the existing approaches for VSTLF identify features in a non-systematic way or use standard autocorrelation analysis, which only captures linear dependencies between the predictor variables and the output variable that is predicted. The main goal

* Corresponding author.
E-mail addresses: irena.koprinska@sydney.edu.au (I. Koprinska), mashud.rana@sydney.edu.au (M. Rana), vassilios.agelidis@unsw.edu.au (V.G. Agelidis).

http://dx.doi.org/10.1016/j.knosys.2015.02.017
0950-7051/

of this paper is to show how advanced machine learning feature selection methods can be applied for electricity load forecasting, and more generally to energy time series forecasting. In particular, our contribution can be summarized as follows:

- We adapt and apply three advanced machine learning feature selection algorithms – Mutual Information (MI), RReliefF (RF) and Correlation-Based Selection (CFS) – to the task of load forecasting. We chose these methods as they are appropriate for the nature of the electricity load data – they can identify both linear and non-linear relationships (MI and RF) and capture both relevant and redundant features (CFS, RF), see Section 3. For comparison we also apply a method based on Autocorrelation (AC). We show how these feature selection methods can be applied in a systematic way to energy time series.
- We propose a two-step approach for feature selection. In the first step we form a set of candidate features by applying a 1 week sliding window. A 1 week sliding window greatly reduces dimensionality while still capturing the main characteristics of data. In the second step we use a feature selection method to evaluate the quality of the candidate features and select a final subset of features.
- We use the selected features with state-of-the-art prediction algorithms: NN, LR and Model Tree Rules (MTR). Hippert et al. [11] reviewed the application of NNs for electricity load forecasting and noted the need for systematic and fair comparison between NNs, standard linear statistical methods such as LR and other prediction algorithms.
- We conduct a comprehensive evaluation using two years of Australian electricity data. This includes a comparison with exponential smoothing (one of the most successful methods for load forecasting), a typical prediction model used by industry forecasters and several other benchmarks.
- We investigate additional aspects of the feature selection algorithms such as effect of the number of neighbors in AC and the number of features in MI and RF.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 analyses the data characteristics. Section 4 describes the proposed feature selection methods and how they were applied to our task. Section 5 presents the prediction algorithms we used and their parameters. Section 6 describes the methods used for comparison. Section 7 summarizes the experimental setup. Section 8 presents and discusses the results. Finally, Section 9 concludes the paper.

2. Previous work

VSTLF is a relatively new area that has become important with the introduction of competitive electricity markets, and more recently, with the arrival of the smart grid. In contrast, short-term load forecasting has been widely studied, e.g. see [11–14].

There are two main groups of approaches for VSTLF: traditional statistical and computational intelligence. Prominent examples of the first group are exponential smoothing and ARIMA; these methods are linear and model-based. The most popular examples of the second group are NNs; for a survey on NNs for electricity load forecasting see [11]. NNs are attractive as they can model non-linear input/output relations and can learn them from a set of examples as opposed to the traditional statistical methods that fit a model and estimate its parameters.

One of the first studies on VSTLF was conducted by Liu et al. [3] who applied NN, fuzzy logic and autoregressive models to predict the load for every minute of a 30 min forecasting horizon. They found that the NN and fuzzy rules were more accurate than the autoregressive models.

Charytoniuk and Chen [4] compared several NN-based methods for the prediction of 10 min ahead electricity load using the load in the previous 20–90 min. They forecasted load differences instead of actual load. The best method achieved prediction error MAPE = 0.4–1% and was implemented in a power utility in the United States, showing good accuracy and reliability.

Shamsollahi et al. [5] used a NN for 5 min ahead electricity load forecasting. The data was processed by applying a logarithmic differencing of the consecutive loads; the NN architecture used 1 hidden layer and the stopping criterion was based on a validation set. They obtained an excellent MAPE = 0.12% and the method was integrated into an energy market system for the region of New England in USA.

Chen and York [6] developed a complex hierarchical NN architecture for 15 min ahead prediction. To predict the load for each day of the week, five NNs were used to cover different time intervals of the 24 h period and their decisions were combined using another NN. They reported MAPE = 0.28–0.87%.

Reis and Alves da Silva [15] predicted the load from 1 to 24 h ahead using North American data. They first decompose the load series into several components using wavelet transform and then used NN-based approaches to make the prediction. The best approach achieved MAPE of 1.12% for 1 h ahead prediction.

Taylor [2] used minute-by-minute British electricity load data to predict the load between 10 and 30 min ahead. He studied a number statistical methods based on ARIMA and exponential smoothing. Some of the methods ignored the seasonal patterns, others captured only the weekly cycle or both the daily and weekly cycles. The best forecasting method was an adaptation of the Holt–Winters smoothing for double seasonality, achieving MAPE of about 0.4% for 30 min ahead prediction; the best methods for 5 min ahead prediction were double seasonal Holt–Winters smoothing, restricted daily cycle smoothing and ARIMA, achieving MAPE of about 0.25%. In [16] Taylor, de Menezes and McSharry compared the performance of four methods for predicting the hourly demand for Rio de Janeiro from 1 to 24 h ahead: ARIMA, double seasonal Holt–Winters exponential smoothing, NN and a regression method with principal component analysis. The simplest method, exponential smoothing, was shown to be the most accurate.

In our previous work on 5 min load forecasting [17] we applied autocorrelation analysis to extract and evaluate several nested feature sets of lag variables. The evaluation was limited to data for one month only. In [7] we used a larger dataset and constructed seasonal and yearly prediction models. We applied autocorrelation analysis to the whole training data, without a sliding window, and extracted 50 features. The most accurate prediction model was LR achieving MAPE = 0.29%. We also found that there was no accuracy gain in building separate seasonal models in comparison to using a single model for the whole year. In this paper we extend our previous work by using a two-step feature selection process with a 1 week sliding window, applying and comprehensively evaluating the performance of a number of feature selection methods in addition to autocorrelation, and comparing the results with exponential smoothing and other baselines.

3. Data analysis

We use electricity load data measured at 5 min intervals for a period of two years: from 1st January 2006 until 31st December 2007. Each measurement represents the total electricity load for the state of New South Wales (NSW) in Australia. The data was provided by the Australian Electricity Market Operator (AEMO) [18].

In order to build accurate prediction models, it is important to understand the data characteristics and the external variables affecting the forecasting.