Neurocomputing 170 (2015) 448–465

Contents lists available at ScienceDirect

Neurocomputing

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

A review of datasets and load forecasting techniques for smart natural gas and water grids: Analysis and experiments

M. Fagiani*, S. Squartini, L. Gabrielli, S. Spinsante, F. Piazza

Department of Information Engineering, Università Politecnica delle Marche, Via Brecce Bianche 1, 60131, Ancona, Italy

ARTICLE INFO

Article history:
Received 1 September 2014
Received in revised form
24 March 2015
Accepted 10 April 2015
Available online 13 July 2015

Keywords:
Heterogeneous data forecasting
Short/long-term load forecasting
Smart water/gas grid
Forecasting techniques
Computational intelligence
Machine learning

ABSTRACT

In this paper, experiments concerning the prediction of water and natural gas consumption are presented, focusing on how to exploit data heterogeneity to get a reliable outcome. Prior to this, an up-to-date state-of-the-art review on the available datasets and forecasting techniques of water and natural gas consumption, is conducted. A collection of techniques (Artificial Neural Networks, Deep Belief Networks, Echo State Networks, Support Vector Regression, Genetic Programming and Extended Kalman Filter-Genetic Programming), partially selected from the state-of-the-art ones, are evaluated using the few publicly available datasets. The tests are performed according to two key aspects: homogeneous evaluation criteria and application of heterogeneous data. Experiments with heterogeneous data obtained combining multiple types of resources (water, gas, energy and temperature), aimed to short-term prediction, have been possible using the Almanac of Minutely Power dataset (AMPds). On the contrary, the Energy Information Administration (E.I.A.) data are used for long-term prediction combining gas and temperature information. At the end, the selected approaches have been evaluated using the sole Tehran water consumption for long-term forecasts (thanks to the full availability of the dataset). The AMPds and E.I.A. natural gas results show a correlation with temperature, that produce a performance improvement. The ANN and SVR approaches achieved good performance for both long/short-term predictions, while the EKF-GP showed good outcomes with the E.I.A. datasets. Finally, it is the authors' purpose to create a valid starting point for future works that aim to develop innovative forecasting approaches, providing a fair comparison among different computational intelligence and machine learning techniques.

1. Introduction

In recent years, several Computational Intelligence (CI) techniques and other solutions have been studied and developed for the extraction of key features to enhance and/or guarantee the quality of services provided through different grid types. Specifically, each network of distribution lines (wires and/or pipes), that connects resource consumers to at least one provider, can be identified as grid. The advances achieved so far, and those potentially achievable, rely on the ability to measure the amount of a certain resource at specific grid points (*metering*), and the availability of such measurements to the scientific community.

In contrast to the electrical energy field, where several datasets and approaches for diverse applications already exist [1–4], activities in the water and the natural gas fields are still immature. As confirmed by the 2013 International Energy Outlook (IEO2013) [5], the natural gas remains an essential resource and, until 2040, 80% of the global energy production will be supplied by fossil fuels, with natural gas being the fastest-growing one, increasing by 1.7% per year. Water is a fundamental resource as well, which must be preserved, being at the basis of human life.

To support research activities, aimed at proposing performing solutions in load forecasting for both water and natural gas smart grids, it is undoubtedly relevant to start from an exhaustive state-of-the-art survey. This motivated the work by Fagiani et al. [6] which is presented here in an up-to-date extension, from 2009 to date. A comprehensive collection of recent state-of-the-art works, concerning water and natural gas forecasting, and related datasets are presented. The overview highlights the lack of suitable datasets in the literature, which represents a serious bottleneck for the development of innovative algorithms and techniques for demand prediction specific to gas and water cases. Indeed, as it will be clearer later on, some of the datasets show a short data recording period, a very low sample rate, and most of them are not publicly available, therefore being not useful for scientific

* Corresponding author.
E-mail addresses: m.fagiani@univpm.it (M. Fagiani), s.squartini@univpm.it (S. Squartini), l.gabrielli@univpm.it (L. Gabrielli), s.spinsante@univpm.it (S. Spinsante), f.piazza@univpm.it (F. Piazza).

http://dx.doi.org/10.1016/j.neucom.2015.04.098

exploitation. This does not allow for a standard evaluation criteria and to perform a comparison between different approaches, since each technique has been tested on a different dataset.

After presenting the deficiencies of the current state of the art, the paper proposes an extended study on the available datasets. A wide variety of algorithms for water and natural gas consumption forecasting have been tested and, when possible, the predictions have been generated using heterogeneous data. In particular, the heterogeneous datasets have been composed combining together the available disjoint sets on water, gas, energy and temperature. The water and natural gas grids have been jointly addressed because it is believed that achievements in each field can easily cross-fertilize the other. This issue is surely enhanced by the similar characteristics of the grids themselves and related common nature of metering devices used in both grids. In particular, there is a widespread use of natural gas for residential heating, as shown in the Snapshot of U.S. Natural Gas Consumption (2008).[1] Precisely in 2009, 62 million homes in the U.S. were heated using natural gas, representing about 56% of the total amount of households. Therefore, a strong correlation should be observed between natural gas consumption and temperature and, if exploited properly, it could improve the forecasting performance.

Formally, as reported above, the definition of grid itself connects the water and natural gas grids with the power one. Being the power forecasting widely addressed, many techniques, firstly applied for electricity prediction, e.g. Artificial Neural Networks (ANNs), have been already introduced to address water and natural gas prediction problems. The selected techniques for the analysis go from ANNs, and their variants as the Deep Belief Networks (DBNs) and the Echo State Networks (ESNs), to Genetic Programming (GP) approach and Support Vector Machine Regression (SVR). Up to the authors' knowledge, few studies have dealt with prediction problems using the DBNs and the ESNs. In particular, concerning DBNs, some works have been proposed for financial predictions [7,8], prediction benchmark testing [9], traffic flow prediction [10], and drought index forecasting [11], but none for water or natural gas consumption forecasting. Therefore this can be seen as an innovative contribution of the paper. Similarly, for the ESNs the only presented approach about prediction concerns the short-term load (electricity) forecast [12] and hydro-power plant reservoir water inflow forecasts [13]. The GP and the EKF-GP approaches have been chosen for their promising prediction performance, recently shown for the water demand scenario in [14], and their easy applicability to the case of heterogeneous inputs. A preliminary work has been presented in Fagiani et al. [15]. Concerning the SVR, it is widely used for a large variety of Machine Learning (ML) applications, therefore it seemed appropriate to assess its effectiveness in the consumption forecasting contexts.

This is the paper outline. The problem statement is given in Section 2. In Section 3 an up-to-date overview of the published datasets for water and natural gas are presented and the different features commented. Section 4 presents the state-of-the-art load forecasting techniques for both water and natural gas grids, thus concluding the survey. In Section 5 the approaches used for the comparative analysis are briefly described. The experimental tests are discussed and related results commented in Section 6. Section 7 concludes the paper.

2. Problem statement

Load forecasting aims to provide an estimate for future consumption, at the highest achievable accuracy. In particular, the focus is to develop approaches able to identify specific behaviours in the past data, $x_n \in \mathbb{R}$ with $n = 1, \ldots, N$, to produce models able to generate accurate predictions, $y_j \in \mathbb{R}$ with $j = 1, \ldots, J$, using only the information on recent consumption, x_r with $r = 1, \ldots, R$. Where N is the number of available past data, R denotes the number of inputs assumed for the prediction with $R < N$, and J is the maximum number of predictable values, limited to $N - R$.

Generally speaking, any load forecasting problem in smart grid contexts can be categorized by the prediction horizon, i.e. how far in time, and the target dimension, i.e. the number of consumers determining the load to be predicted. The former can be distinguished in: short-term, medium-term and long-term prediction. It is a loose classification, and slight changes are often encountered among different contributions. Therefore, to avoid confusion and misunderstanding, the classification adopted in this work assumes a *short-term* prediction with a time horizon lower than 48 h. The forecast featuring an horizon up to one to six months is addressed as *medium-term*, and from six months onwards, it is *long-term*.

The target dimension has become more and more important in recent years, with the spreading of the *smart-meter* paradigm. At the beginning, due to the difficulty of collecting the consumption of single consumers with an adequate rate, the forecast techniques have been developed and evaluated only with large amounts of aggregate consumption, e.g. cities, districts or distribution lines. Nowadays, the increasing use of *smart-meters* allows the collection of large amount of data, thanks to automatic readings and network connections, from end users (houses, shops or factories) as well as from primary points of the distribution network. So, the novel prediction approaches have to deal with different origins for the data. In particular, working with city or district consumption reduces the short-term variability of the data, but causes the need to achieve a more precise forecasting, because even a small error can effect dramatically the predicted consumption, and cause wrong decisions. On the contrary, household consumption data presents a high short-term variability, thus, great changes could occur in short times, but larger errors are tolerable due to the reduced resource usage.

The consumption depends on the environmental and climatic conditions. Being able to consider these heterogeneous data is of major concern. Suitable data are, e.g, temperature, humidity, cloudiness values, compatible with the consumption datasets. However, higher complexity models are needed to correctly handle the heterogeneous information, therefore the application of heterogeneous data must be carefully considered. Unfortunately, higher complexity does not always mean better performance.

For these reasons, finding a suitable dataset is crucial to perform a proper analysis for a given scenario. Moreover, the collected data, possibly heterogeneous ones, must have a temporal length adequate to the prediction horizon and suitable to perform a valuable prediction: training and validation data have to be sufficiently populated.

A crucial point for the comprehension of the performance achieved by the prediction methods is the usage of suitable evaluation criteria. As discussed later on, the state-of-the-art contributions fail to produce a clear and understandable comparison between different prediction techniques. The authors select the evaluation criteria to adopt for each experiment, in order to facilitate the comparison and the analysis of the methods. So, the normalized mean square error (NMSE), the determination coefficient (R^2, commonly known as Nash–Sutcliffe efficiency coefficient [16]), the mean square error (MSE), the mean absolute percentage error (MAPE) and the relative root mean square error (RRMSE) have been taken into account. The corresponding formulas are

$$\mathrm{NMSE} = \frac{\sum^N (\tilde{y}_n - y_n)^2}{\sigma_y^2 \cdot N}, \quad R^2 = 1 - \frac{\frac{1}{N}\sum^N (y_n - \tilde{y}_i)^2}{\frac{1}{N}\sum^N (y_n - \overline{y})^2},$$

[1] http://www.aga.org/our-issues/issuesummaries/Pages/SnapshotUSNaturalGas.aspx