Energy and Buildings 111 (2016) 299–310

Contents lists available at ScienceDirect

Energy and Buildings

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

The role of data sample size and dimensionality in neural network based forecasting of building heating related variables

Martin Macas[a,*], Fabio Moretti[b], Alessandro Fonti[c], Andrea Giantomassi[c], Gabriele Comodi[c], Mauro Annunziato[b], Stefano Pizzuti[b], Alfredo Capra[d]

[a] Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Czech Republic
[b] Italian National Agency for New Technologies, Energy and Sustainable Economic Development, Italy
[c] Universitá Politecnica delle Marche, Italy
[d] Universitá degli Studi Roma Tre, Italy

ARTICLE INFO

Article history:
Received 27 July 2015
Received in revised form
22 November 2015
Accepted 23 November 2015
Available online 26 November 2015

Keywords:
Heating
Power consumption
Temperature
Thermal comfort
Neural networks
Prediction
Feature selection
Input selection

ABSTRACT

Energy consumed in buildings represents a challenge in the context of reduction of greenhouse gases emission. For this reason and due to the growing interest in operative costs reduction the energy used by buildings (tertiary and privates) for heating, ventilating, and air conditioning (HVAC) is even more investigated. Due to the nature of the energy consumption profile a predictive optimization method is one of the solution the scientific literature spreads even more. However optimization techniques need a good and reliable prediction of the variables of interest over a time horizon. This work focuses on methods to obtain a robust and reliable predictor based on artificial neural networks. For the optimization purposes the neural model predicts total heating energy consumption (gas), internal air temperature and aggregated thermal discomfort 12 h ahead. Training and testing data are simulated using a simulator based on heat, air and moisture model for building and systems evaluation (HAMBASE), by which a real office building was modeled. Influence of training data sample size and selection of predictor inputs is examined. Several combinations of early stopping condition and network complexity are tested for different training sample sizes. It is observed that the early stopping mechanism is crucial especially but not only for small training data, because it reliably overcomes overfitting problems. Surprisingly, relatively small networks were sufficient or performed best, although examined range of training sample covered up to five heating seasons. The use of a model tuning is thus supported by the results. Further, two strategies of selection of suitable input variables are demonstrated. While the input selection does not degrade the prediction performance, it is able to reduce the dimensionality and thus to save computational, communication, time, and data acquisition demands. The importance of inputs selection in HVAC modeling is thus pointed out and demonstrated.

1. Introduction

Heating, ventilation, and air conditioning (HVAC) in buildings are the most important consumers of energy. In developed countries, buildings account for 20–40% of the total final energy consumption [1], HVAC consumption accounts for half the energy use in buildings and one fifth of the total national energy use [1]. In Europe, buildings accounts for 40% of total energy use and 36% of total CO_2 emissions [2]. Moreover, 46% of energy is used for heating and cooling, 33% is used for mobility reasons, and 21% is used for electrical power; also in the United States, HVAC systems constitute even over 50% of the building energy. Within the commercial sector, office and retail buildings are together the biggest energy consumers. All those examples suggest that the energy efficiency in building heating should be one of world's priorities.

This article follows this goal by focusing on predictive models for one-step ahead forecasting of total gas consumption, internal temperatures, and thermal discomfort in one particular gas-heated office building. The modeling and prediction of heating related variables has been intensively studied for a long time. A very comprehensive review of modeling methods for HVAC systems can be found in [19]. The predictive models can be further used for an optimization of heating operation [5,20]. In such scenario, a building management system can find the inputs of those predictive

* Corresponding author.
E-mail address: mmacas@seznam.cz (M. Macas).

http://dx.doi.org/10.1016/j.enbuild.2015.11.056
0378-7788/

models that lead to proper values of their outputs. Such inputs can be future set points for indoor air temperature and supply water temperature. The optimization task can be further defined as constrained optimization or multi-objective optimization. An example of the former is the minimization of the predicted gas consumption subject to a pre-defined maximum level of the predicted thermal discomfort [9]. An example of the latter is the simultaneous minimization of predicted gas consumption and thermal discomfort and selection of one of multiple pareto-optimal solutions in terms of multi-objective optimization [21]. Another example can be a direct control of thermal discomfort by minimization of the difference between its predicted value and a reference [22]. In those approaches, it is crucial to reach a good prediction accuracy of consumption, thermal discomfort and temperature, because the control errors cannot be reduced under the limit imposed by the prediction performance of the model the controller is based on.

An important novelty of this article is that the prediction of three different heating related variables in multiple zones is examined (although some studies predict both the consumption and air temperature [4,14]). In HVAC literature, the main focus is naturally devoted to variables related to consumption (see [23] or [6] for literature survey). In some cases, however, it can be also required to predict internal temperature. Since the current internal temperature is one of the inputs of consumption and discomfort models, in case of multiple steps ahead prediction, the temperature must be predicted and fed back as an input to predict the next step. In another scenario, temperature model can be used for estimation of the time needed to pre-heat the building to some comfort temperature [24]. The temperature prediction is thus also important for many heating management systems. Finally, thermal comfort is also forecasted here. In HVAC studies and artificially ventilated buildings, it is often assessed in terms of a predicted mean vote (PMV) model [25]. A forecasting of PMV values is not so common, because it is difficult and expensive to measure training data. Here, this expensiveness and lack of the data is overcame by the use of simulation. In real case, the thermal comfort can be assessed by questionnaires or some simplified measurements of thermal comfort [26]. It should be remarked that there are other variables that can be worth to predict. Examples of such variables are air quality [27], relative humidity, electrical consumption [5], or occupancy and behavior of people [28,29]. Those variables are however not focused in this article.

Three main types of HVAC predictive models exist – physical, semi-physical and data-driven [23]. In this article, purely data-driven black-box models are considered. The focus on data-based statistical modeling technique can be motivated by some practical features of the heating management problem. Physical or semi-physical models of HVAC related variables can give precise predictions, but are highly parameterized and computationally expensive for control applications. Moreover, measured data are necessary also for those physical models and their calibration and validation. Therefore, the data-based predictive models of HVAC related variables in buildings are of great importance for many HVAC control and energy management problems. A very representative example is the model predictive control, which optimizes forecasts of target variables over a certain time horizon and applies only the first time step of this horizon. With some exceptions [9], linear models are mostly used that make it possible to solve the optimization problem using fast quadratic programming methods. The parameters of such model are usually estimated by system identification methods (e.g. identification of parameters of state space equations). On the other hand, modeling of the energy variables using ANNs is a very popular approach especially in last few years, because of their ability to represent the non-linear dependencies. A review on the use of ANN in energy applications in buildings can be found in [30], where one can observe that common feed-forward networks (FFN) are the most popular models. This has been also supported by predictor shootout contests organized by ASHRAE, where feed forward ANNs trained by the back propagation algorithm were the most efficient and effective models for energy prediction [31]. This trend is also followed here, only feed-forward neural networks (multi-layer perceptrons) are considered. Another very common method is the non-linear auto regression with exogenous inputs [11], which can be understood as a recurrent neural network. The recurrent neural networks, although used less often than FFNs, are also popular (e.g. [32–34]). Other much less frequent approaches are artificial neural fuzzy interface system [35], combination of multiple ANNs [36], pseudo-dynamic transitional characteristics used as one of inputs for a feed-forward network [6,37].

However, the non-linear character of those models implies a need of more complex, often population-based optimization heuristics [5,38]. Those optimization methods repeatedly run the model in a loop, which causes high computational demands that can make the real-world implementation impossible. This is a very strong motivation to make the model as simple and fast as possible, this goal can be partly reached by using the techniques like input selection or model tuning. The reduction of complexity is however not the only potential benefit of those approaches, those techniques can also improve the generalization abilities and performance of the final predictive models. The model tuning is related to the fact that an optimal predictive model depends on the available training data, especially the sample size (number of data instances) and dimensionality (number of model inputs) are easily observable properties. While the first can be hardly changed without measuring more data and implies a need for model tuning, the second can be reduced by input selection methods.

To motivate our research by a literature evidence, we summarized some examples of recent studies dealing with prediction of HVAC related variables in Table 1. Because of the limited space, the table does not comprehensively summarize all the research, but only those similar to our approach with iteratively trained predictors. The table examines if a division of the data into test/train set is described (column 4), if a selection of input variables was performed (column 5), if the predictor was tuned by some means (column 6) and if an early stopping was used (column 7).

Although a vast majority of studies clearly describes a division of available data into training and testing set and time periods spanned by those data sets (column 4 of Table 1), most of those studies avoid discussing of the satisfactory sample size and mostly satisfy with momentary accessible data. The size of the training and testing data thus spans different values from couple of days [6,10], months [17,18] to several years [3,14]. We found only one study that examines an impact of the sample size on forecasting of heating related variables in [9]. Therefore, the influence of sample size on the prediction performance is focused in this article. Moreover, this article shows a strong dependence between data sample size and feasible model complexity (number of hidden neurons). A direct consequence of this is that a model tuning should be always performed to reach a good predictive performance. As can be seen from Table 1, some studies do not perform (or do not describe) model tuning and simply guess a suitable model complexity. Moreover, the methodology of model tuning is often not correct or sufficiently described. Some models are tuned on the testing data, which usually gives biased estimate of the final performance. This occurs in [3] or [4], where different numbers of neurons are compared on testing data, but the results of comparison are not further validated on independent data. In [5], five different model architectures are compared, but it is not clearly explained what is the testing set.

From survey described in [6,30] it appears that the most common inputs of heating consumption prediction studies are internal and external temperature, solar radiation, relative humidity, wind