Chapter 9

The Impact of Data Segmentation in Predicting Monthly Building Energy Use with Support Vector Regression

William Mounter, Huda Dawood, and Nashwan Dawood

Abstract Advances in metering technologies and machine learning methods provide both opportunities and challenges for predicting building energy usage in the both the short and long term. However, there are minimal studies on comparing machine learning techniques in predicting building energy usage on their rolling horizon, compared with comparisons based upon a singular forecast range. With the majority of forecasts ranges being within the range of one week, due to the significant increases in error beyond short term building energy prediction. The aim of this paper is to investigate how the accuracy of building energy predictions can be improved for long term predictions, in part of a larger study into which machine learning techniques predict more accuracy within different forecast ranges. In this case study the ‘Clarendon building’ of Teesside University was selected for use in using it’s BMS data (Building Management System) to predict the building’s overall energy usage with Support Vector Regression. Examining how altering what data is used to train the models, impacts their overall accuracy. Such as by segmenting the model by building modes (Active and dormant), or by days of the week (Week-days and weekends). Of which it was observed that modelling building weekday and weekend energy usage, lead to a reduction of 11% MAPE on average compared with unsegmented predictions.

Keywords Buildings · Data segmentation · Energy · Prediction · SVR

9.1 Introduction

With greater moves towards using machine learning in predicting building energy usage, there are many direct comparisons of learning techniques using the same data. However, it is considerably rarer when learning techniques are compared multiple times over multiple ranges [1–5]. With multiple ranges usually occurring when only testing a singular learning technique. Whilst a learning machine’s accuracy can be

W. Mounter (✉) · H. Dawood · N. Dawood
Teesside University, Stephenson St, Tees Valley, Middlesbrough TS1 3BA, UK
e-mail: W.Mounter@tees.ac.uk

observed when tested on a single dataset, the observed error is of limited value for comparative purposes with other learning techniques trained upon other data sets. As it is not known how each of the techniques would have performed if they had used the other test's data.

As such this study has been performed as part of a larger research project into investigating the impact of using different machine learning techniques (LR, SVR, ANN and Predictive Algorithms) on the accuracy of predicting building energy usage over a rolling horizon. In addition to investigating how the error that occurs in said predictions may be reduced through the use of data segmentation, such as if it is more accurate to model every single energy meter in a building and summate their predictions, or model the building as a whole. This specific paper focusing on the change in accuracy of building level energy use, by SVR (support vector regression), and the forecast range increase; and if error of long term predictions (Monthly) can be reduced to or below the level of the average error in short term predictions (Daily and weekly) through data segmentation.

Data segmentation, being the process of dividing and grouping data based on chosen parameters, in this case timeframes, so that it can be used more effectively; such as creating two separate prediction models for the separated groups. (As opposed to data splitting, in which data is randomly split for cross validation usage). Or to use an analogy, in cars, winter and summer tyres tend to perform better in their respective seasons than each other and all-season tyres, but poorer than each other and all-season tyres outside of their respective seasons. Or, in the case of machine learning, would a model trained with only with weekend building energy use data be more accurate at predicting weekend building energy use than a model trained with weekend and weekday energy use?

Additionally, Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that evaluate data and identify patterns which can be used for regression analysis and classification; first identified by Vladimir Vapnik and his colleagues in 1992 [4, 7]. One of its main advantages, and reasons for its popularity in predicting building energy usage comes from its ability to effectively capture and predict nonlinearity [6].

To investigate the study's aim the Clarendon Building, part of Teesside University Campus, was selected for use in this study- due to the data rich environment it's BMS (Building Management system) provided. Previous studies into this building utilizing square regression analysis typically had a baseline of "5% Mean Absolute Prediction Error (MAPE)" for the demands of each assets in one day ahead forecasts [8]. With investigations into the impact of data segmentation on ELM (Extreme learning machine) predictions using this data reduced the average of monthly prediction's percent error of the building's cooling system from 44.33 to 19.03%, reducing the error by 25.29% [9].

9.2 Research Method

From the Teesside university campus, amalgamated, building level and specific meter datasets were available of BMS data from January 2018 to December 2018. These datasets contained 15-minute averages of building elements energy usage, as well as sensory data of the internal and external environmental temperatures. From the Clarendon's BMS system, the total energy use of the building over the 2018 period, alongside their relevant timestamps, was extracted for use in investigating the accuracy of predicting building energy usage with support vector regression, over a variety of ranges and segmentations. Specifically:

- The impact of forecasting range on model accuracy, from one day to one month predictions into the future.
- The impact of segmenting the data along the lines of week days and week ends, creating independent prediction models for both, on the accuracy of said models compared with the base model.
- The impact of segmenting the data along the lines of active building periods and inactive periods, creating independent prediction models for both, on the accuracy of said models compared with the base model.

Noting that whilst the accuracy is expected to decrease as the forecast range increases, the point of interest of in studying the change of accuracy over a forecasting range, is the amount it increases by, and how significant that increase is.

To do this end, the data set was then segmented into it's individual seasons, before:

- The first month of each season, is placed into a season specific training data set.
- The second month of each season, is used to create three data sets that are to be used for prediction purposes, containing the first day, week and entirety of the month.

Creating in total four control training sets, and twelve control test data sets for evaluating each model's predictive accuracy over a rolling horizon. For the further segmented data, said control month (Training and testing) datasets would be subdivided into:

- Building active and inactive periods, where in separate models are created for the 8:00 am to 18:00 pm active period and the 18:15 pm to 7:45 am period of the data set.
- Weekdays and weekends, where in separate models are created for the Monday to Friday period and the Saturday to Sunday period of the data set.

Each model (A gaussian kernel SVR machine) is then trained with the first month of each season, to predict the following day, week and month of said season; to establish a baseline accuracy for each model. With the segmented data sets being trained and tested on their own relevant segments, the predicted values being recombined and then compared with the predictions made by the regression models trained on

said months none-segmented data. Recording all of the predicted consumption for use in comparison with the actual consumption, evaluating the respective accuracies in terms of their Mean Absolute Percent Error.

9.3 Results and Discussion

9.3.1 *Daily, Weekly and Monthly Control Building Energy Predictions*

As can be seen in Fig. 9.1, as the forecast ranged increased so did the building energy use prediction error. But the increase in error was not proportional to the increase in forecasting range, e.g monthly predictions having thirty times the errors of daily ones. Instead, the bulk of the increase occurs between moving from daily predictions to weekly ones. Suggesting that the internal variations from the predicted ‘norm’ within a week are greater than the overall energy drift that occurs between the average energy use each week over the month period.

An example of these variations from the predicted energy consumption can be observed when comparing the predicted energy consumption to the actual consumption, as can be seen in Fig. 9.2. The two dips in actual building energy use between event 200 and 400 of the first week of October 2018, can be observed in the majority of weeks through the year period, and are the result of reduced building activity during the weekend periods. As the prediction model has to accommodate for both the weekday and weekend energy use patterns, resulting in an average of the two, weighted in the weekday’s favor, due to outnumbering the weekends.

Fig. 9.1 A comparison of building energy usage prediction error

Fig. 9.2 The first week of autumn’s actual vs predicted energy use

9.3.2 Segmented Monthly Building Energy Usage Predictions

As can be seen in Fig. 9.3, in every tested incnet of weekday/end segmentation, the MAPE was reliable reduced comparatively to that of the unsegmented prediction, whilst the impact of segmenting by building active and dormancy periods had a negligible and erratic impact upon accuracy (<5%). This was unlike previous studies into the Clarendon building [9] where in segmenting between active and dormancy periods resulted in the greater reduction of prediction error. This is most likely due to the distinct difference in energy usage between the building’s chillers being active and dormant, compared to the more continuous relationship and pattern of energy use observed at the summated level. (The Chillers being two separate patterns of energy use, whilst the total energy use being one (ignoring the weekday/weekend divide)).

Fig. 9.3 A comparison of segmented and non-segmented building energy prediction error

Fig. 9.4 The first segmented week of autumn's actual vs predicted energy use

As can be observed in Fig. 9.4, using multiple models to predict the week's building energy usage, allowed for greater accuracy in predicting the two distinct patterns of building energy use. Reducing the average MAPE of the monthly predictions from 31.18 to 19.73%, resulting in predictions with greater accuracy than the unsegmented weekly models (30.65%), having removed the main source of error that occurred in the weekly cycle. Though the segmented monthly prediction does possess 2.33% more MAPE than the average daily predictions (17.4%), that may potentially be attributed to the drift in energy usage from the training period from the changings of HVAC usage as weather patterns change with the progression of time, or from changes in overall building use from holidays, lesson scheduling and shifting building occupancy patterns.

9.4 The Conclusion

In summary, segmenting the training data, in the case of predicting building energy usage with SVRs, produces greater overall accuracy in predicted building energy use, when it is used to account for regular and reliable deviations from the 'normal' pattern of building energy usage. Such as in the case of segmenting between the two different building use patterns of weekends to week days. But having negligible to potentially negative impacts upon predictive accuracy when used to divide singular consistent patterns. Such as in the case of the Clarendon's buildings active and dormancy period, which followed a singular continuous pattern of behavior. Whilst the largest source of error (the internal deviations within each week period), could be addressed through segmenting the data, other data processing techniques or learning machines would have to be used to account for the drift in energy use that occurs over such long periods. As there is no distinct way to segment the training data to account for it.

Based upon these results, the three main areas of future work are as follows:

- The exploration of the data of the Clarendon building's specific meter energy consumption over the course of the year period, to observe how each building element's consumption changes over time.
- The comparison of the building energy model based upon the summated building energy use, with the summated predicted energy use of a set of models based upon each electric meter individually.
- The direct comparison of these results with the results of using the same datasets to train other learning machines previously mentioned.

Acknowledgements This research is a part of a PhD being undertaken at Teesside University under the supervision of Dr Huda Dawood and Prof. Nashwan Dawood. Additional acknowledgements should be given to Dr Chris Ogwumike for his assistance in acquiring the data used in this study.

Appendix

SVR calibration method, kernel selection and results, as well as other data created in this study, are available upon request.

References

1. M. Boegli, Y. Stauffer, SVR based PV models for MPC based energy flow management (2017)
2. Y. Chen, H. Tan, Short-term prediction of electric demand in building sector via hybrid support vector regression (2017)
3. Y. Chen, H. Tan, X. Song, Day-ahead Forecasting of Non-stationary Electric Power Demand in Commercial Buildings: Hybrid Support Vector Regression Based (2017)
4. Y. Ding, Q. Zhang, T. Yuan, Research on short-term and ultra-short-term cooling load prediction models for office buildings (2017)
5. S. Goudarzi et al., Predictive modelling of building energy consumption based on a hybrid nature-inspired optimization algorithm (2019)
6. M. Shen, H. Sun, Y. Lu, Household Electricity Consumption Prediction Under Multiple Behavioural Intervention Strategies Using Support Vector Regression (2017)
7. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995)
8. P. Boisson, S. Thebault, S. Rodriguez, S. Breukers, R. Charlesworth, S. Bull, I. Perevozchikov, M. Sisinni, F. Noris, M.-T. Tarco, A. Ceclan, T. Newholm, *DR Bob D5.1* (2017). https://www.dr-bob.eu/wpcontent/uploads/2018/10/DRBOB_D5.1_CSTB_Update_2018-10-19.pdf. Accessed 2019
9. W. Mounter, H. Dawood, N. Dawood, The impact of data segmentation on modelling building energy usage. In The International Conference on Energy and Sustainable Futures (ICESF). Nottingham Trent (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.