Data Driven Methods for Building Thermal Response Predication

Suxin Mao
Department of Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA
suxinm@andrew.cmu.edu

Zhiang Zhang
School of Architecture
Carnegie Mellon University
Pittsburgh, PA
zhiangz@andrew.cmu.edu

ABSTRACT

The building thermal response model is the core part, and may be the most difficult part, in the model-based building control algorithm such as the model predictive control (MPC). This paper aims at using different data driven methods for building thermal response prediction, i.e. predicting the indoor air temperature of a room given the weather conditions and building system operation conditions. The building operation data of the Center for Sustainable Landscape (CSL) building of the Phipps Conservatory in Pittsburgh, PA is used for training and validating the different methods.

In this paper, linear regression, decision tree regression, and lazy learning are used as the data driven methods. It is found that the simplest model free lazy learning method gives the best prediction accuracy during the testing stage. The linear correlation coefficient between the predicted values and the true values are 0.83; the root mean square error and the mean absolute error achieve 0.07 and 0.05 respectively (after normalizing data to 0-1 scale).

CCS Concepts

•**Computing methodologies** → **Linear regression; Classification and regression trees; Machine learning approaches; Instance-based learning;**

Keywords

Building thermal response prediction; regression; lazy learning; machine learning

1. INTRODUCTION

Model-based building control algorithm, such as the model predictive control (MPC), is popular in recent years because of its potential to achieve better energy efficiency and indoor thermal comfort. Such model-based building control algorithm usually requires a building thermal response (BTR) model to predict future indoor air temperature given weather conditions and control inputs to the Heating, Ventilation and Air-conditioning (HVAC) systems. However, creating an accurate and computationally fast BTR model is challenging. In [7], physical-based building energy simulation model is used for MPC, but the simulation time of such model is so long that the practical value is limited. In [2, 3], resistance-capacitance (RC) thermal network model is used for MPC, but the modeling process is complex which is not practical for large buildings.

With the better availability of the building operation data and the development of the machine learning techniques, data driven BTR models become popular. This kind of model is usually computationally faster and more accurate than the physical-based models [6], and is less complex compared to the RC thermal network models.

In this paper, three data driven BTR modeling methods are presented, including linear regression, decision tree regression, and lazy learning. The methods are evaluated using the real building operation data of an office building in Pittsburgh.

DOI: 10.1145/1235

2. DATA DRIVEN APPROACHES

2.1 Linear Regression

Linear regression is the most straightforward way to predict data. In this method, we consider all features in the dataset, and assume that they have linear relationship. The model can be obtained using algorithm 1. It provides information on the relationship between different inputs. After performing linear regression, the model should be scored on both training dataset and testing dataset. R-squared value is used to evaluate the model performance, with 1 indicating the best performance and 0 indicating zero relationship between input and output.

2.2 Decision Tree Regression

Decision tree regression [4] is an improvement of linear regression for large dataset. It classifies data into a decision tree using X features, and performs linear regression within each leaf. This method is a significant improvement of linear regression as it avoids non-linearity throughout the whole dataset, and only assess linearity within similar data.The model can be obtained using algorithm 2.

Algorithm 1: Linear Regression

input : $D = \{(\mathbf{x_i}, \mathbf{y_i}) \in \mathbb{R}^m \times \mathbb{R}\}$, training dataset.
input : $\mathbf{x_{test}} \in \mathbb{R}^m$, test sample input.
output : y_{pred} = predicted output.
procedure: Construct linear system in the form of $A \cdot \vec{k} = \vec{b}$, where A is all X features in training dataset in matrix form, b is all indoor air temperature in training dataset in vector form. Solve the system by $\vec{k} = A^{-1} \cdot \vec{b}$ to obtain linear relationship between x features and y output, $y = k_1 \times x_1 + k_2 \times x_2 + ...$

Algorithm 2: Decision Tree Regression

input : $D = \{(\mathbf{x_i}, \mathbf{y_i}) \in \mathbb{R}^m \times \mathbb{R}\}$, training dataset.
input : $\mathbf{x_{test}} \in \mathbb{R}^m$, test sample input.
output : y_{pred} = predicted output.
procedure: Construct a decision tree by $(x^k, t_k) = arg\ min \sum_{i|x_i \in R_L} ||y_i - \bar{y}_L||_l^2 + \sum_{i|x_i \in R_R} ||y_i - \bar{y}_R||_l^2$, where x is features, t is threshold and y is indoor air temperature. Perform linear regression within each leaf in the decision tree.

2.3 Lazy Learning with K nearest neighborhoods (LL)

2.3.1 Algorithm

Lazy learning [1] is a simple model-free local learning approach but works well for time series forecasting. It takes the training samples as a "pool", and every time the new test sample comes in, it looks for the K nearest data points to the test sample from the pool, and predicts the value of the test sample based on the values of the K nearest data points. Formally, the algorithm can be written in algorithm 3 [1]:

Algorithm 3: Lazy Learning [1]

input : $D = \{(\mathbf{x_i}, \mathbf{y_i}) \in \mathbb{R}^m \times \mathbb{R}\}$, training dataset.
input : $\mathbf{x_{test}} \in \mathbb{R}^m$, test sample input.
input : K = number of neighbors to consider.
output : y_{pred} = predicted output.
procedure: Sort increasingly the set of vectors $\{\mathbf{x_i}\}$ with respect to distance to $\mathbf{x_{test}}$. $y_{pred} = kernel(\mathbf{x_{test}})$ where the kernel function is locally learned using the data $(\mathbf{x_1}, \mathbf{x_2}, ...\mathbf{x_K}, \mathbf{y_1}, \mathbf{y_2}, ...\mathbf{y_K})$

For the lazy learning algorithm, the value of K, the distance metric and the kernel function should be selected. In this paper, the distance metric is the Euclidean distance. The value of K is chosen by the cross validation. Two kernel functions are studied, including linear regression and the non-weighted average, as shown in the equation (1) and equation (2).

$$\mathbf{y_{pred}} = \mathbf{x_{test}}\mathbf{w} + b \tag{1}$$

where w and b are learned using the K training data that is the nearest to the x_{test}.

$$\mathbf{y_{pred}} = \frac{1}{K}\sum_{j=1}^{K}\mathbf{y_j} \tag{2}$$

where y_j are the K nearest training data points that are the nearest to the x_{test}.

2.3.2 Time Series Forecasting

Two different methods are discussed for the lazy learning for time series forecasting, as shown in the equation (3) and (4).

$$\mathbf{y_{pred,k}} = LL(\mathbf{x_{test,k}}) \tag{3}$$

$$\mathbf{y_{pred,k}} = LL(\mathbf{x_{test,k}}, \mathbf{y_{k-1}}, \mathbf{y_{k-2}}, \cdots, \mathbf{y_{k-l}}) \tag{4}$$

where l is the maximum historical steps (lagged results) to consider to predict the y at the current step k.

The method in equation (3) is straightforward because it uses sample inputs at the current time to predict the output at the current time. For the method in equation (4), the historical outputs are also considered to predict the output at the current time step. The recursive strategy is usually used to predict the value for a long time steps ahead. That is, predicted output at this time step is the input to the model for the next time step. In this paper, the recursive order is chosen to be 1000, which is equivalent to about 1 week if the sample data interval is 10 min. This means equation (4) is applied for 1000 times with the predicted outputs from the previous time steps as the inputs to the model.

3. DATASET

The training and testing data for this paper is from the building automation system (BAS) of the CSL building of Phipps Conservatory, Pittsburgh, PA. The dataset is for the open plan office room on the 1st floor of the CSL building. The dataset includes the following data from 2014-11-01 to 2016-11-01 with 10 minute interval: 1) outdoor air temperature (OA); 2) outdoor relative humidity (RH); 3) total solar radiation (SR); 4) wind speed (WS); 5) damper position of the VAV in the room (DP)-0 to 10, 0 means damper is closed and 10 means the damper is fully open; 6) supply air temperature of the VAV in the room (ST); 7) month of the year, day of the week and hour of the day (MO, DY, HR); 8) (the y) air temperature of the room (AT);.

The dataset will be divided into two parts equally. The first part will be used for training the data driven models, and the second part will be used for testing the model performance.

4. RESULTS AND VALIDATION

4.1 Error Metrics

Three error metrics are used to evaluate the performance of the data driven models, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Linear Correlation Coefficient (R) between the predicted values and the true values. The error metrics will be applied to the normalized data. Therefore, RMSE and MAE has the range from 0 to 1 and closing to 0 means better prediction accuracy. For R, it has the range from -1 to 1 and 1 means the best prediction accuracy. To make the R metric comparable to RMSE and MAE, $norm(1-R)$ which is the $(1-R)$ normalized to the range 0 to 1.

$norm(1-R)$ is mainly used for this paper to compare the performance of the different models.

4.2 Linear Regression

Linear regression is performed using different input and output settings. Model is trained by either all features or outdoor air temperature only. All year data, seasonal data and monthly data is used to test the performance of linear regression model. The models are then assessed using testing data from another year. The result of assessment indicates that the R-squared values for all linear regression models are negative. More studies with respect to right features for the linear regression should be conducted in the future. Auto-regressive linear regression should be considered [5].

4.3 Decision Tree Regression

Decision tree regression is performed in three difference ways. The first analysis is done using all X features and data of all year. The second analysis is done using outdoor air temperature as the only input and data of all year. The third analysis is done using all X features and monthly data. The models are then assessed using data of another year. Models are scored by R-squared value. All methods are subject to over-fitting, since they all have a score of 1.0 using training data, and a very low score using testing data. The results shows that the first method has the best performance, with a score of 0.50. The overall comparison between true value and predicted value is shown in Figure 1, and the detailed comparison within the time frame of January, 2016 is shown in Figure 2

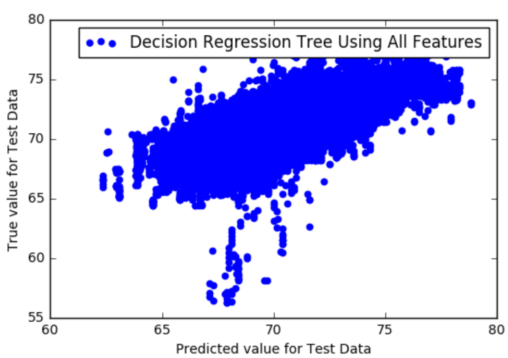

Figure 1: Predicted-True Value Plot of Decision Tree Regression Model 1 on the Test Dataset, $norm(1-R) = 0.290805$

4.4 Lazy Learning

4.4.1 Models

Three alternatives of LL are tested and evaluated, including a 1) LL-Avg: LL with non-weighted average kernel with the original features (equation (2) and (3)), 2) LL-Avg-Lag: LL with non-weighted kernel with lagged features (equation (2) and (4)), 3) LL-LR: LL with linear regression kernel with the original features (equation (1) and (3)). For the second case, the OA, SR and AT are lagged by two time steps, that

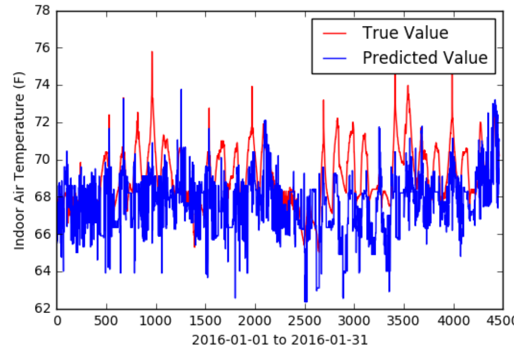

Figure 2: Time Series Comparison between the Predicted and the True Indoor Air Temperature for Decision Tree Regression Model 1

is, the whole set of features are the features described in section 3 plus $\{OA_{k-1}, OA_{k-2}, SR_{k-1}, SR_{k-2}, AT_{k-1}, AT_{k-2}\}$.

4.4.2 Data Normalization

It is found that each feature in the original dataset has very different scales in its value. This will lead to poor accuracy in this study because features with high-value scale dominate the distance metric calculation. Therefore, all data, including training and test data are normalized into 0 to 1 scale using the min-max normalization, as shown in equation (5).

$$z = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{5}$$

where x is the scalar value for normalization, x_{max} is the maximum value that x can take in the dataset, x_{min} is the minimum value that x can take in the dataset, z is the normalized value.

4.4.3 Cross Validation

10 fold cross validation is used for the selection of K based on the training dataset for each LL models. The model with lowest average value of RMSE, MAE and $norm(1-R)$ will be selected for the later testing stage. It is found that $K = 5$, $K = 5$ and $K = 100$ are the best choice of K for LL-Avg, LL-Avg-Lag and LL-LR.

4.4.4 Performance Evaluation

Figure 3, 4, 5 show the predicted value versus true value plot for three LL models testing on the test dataset. It is clear that two models with the non-weighted average kernel outperforms the LL model with the linear regression kernel. The LL-Avg-Lag has the slightly better performance than the LL-Avg model.

Figure 6 shows the part of the prediction results using the LL-Avg-Lag model for the test data. The recursive order is 1000, as described in the section 2.3.2. The error metrics for LL-Avg-Lag over the testing dataset are: $RMSE = 0.069550, MAE = 0.053041, norm(1-R) = 0.084619$ under the normalized dataset.

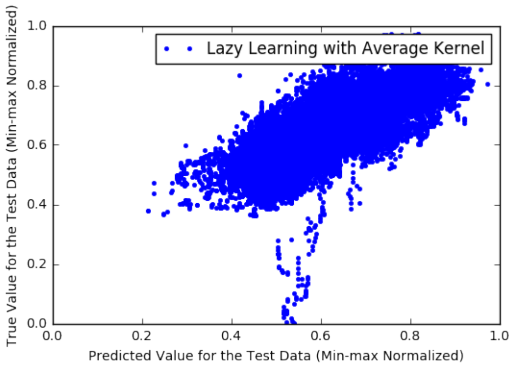

Figure 3: Predicted-True Value Plot of LL-Avg on the Test Dataset, $norm(1-R) = 0.097067$

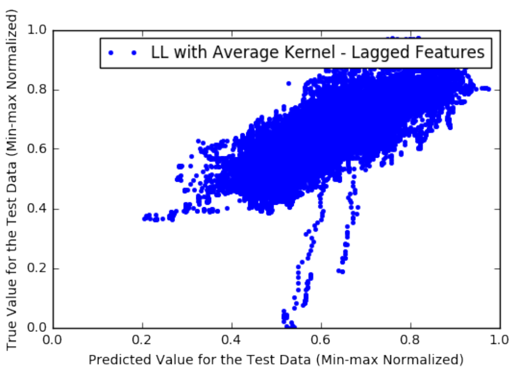

Figure 4: Predicted-True Value Plot of LL-Avg-Lag on the Test Dataset, $norm(1-R) = 0.084619$

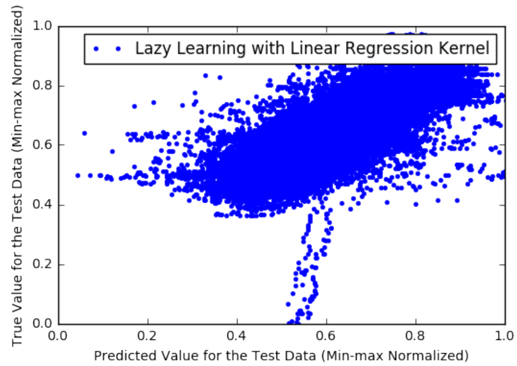

Figure 5: Predicted-True Value Plot of the LL-LR on the Test Dataset, $norm(1-R) = 0.496181$

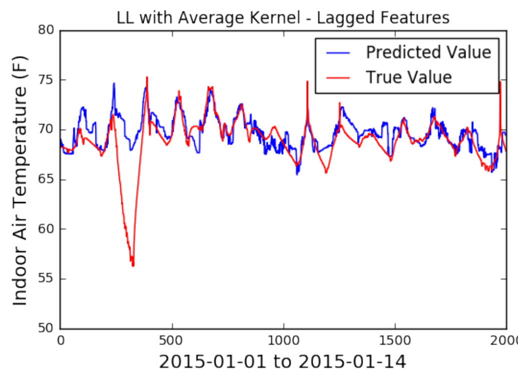

Figure 6: Time Series Comparison between the Predicted and the True Indoor Air Temperature for the LL-Avg-Lag Model (Recursive Order is 1000)

5. CONCLUSION AND THE FUTURE WORK

It has been shown by our analysis that lazy learning outperforms linear regression and decision tree regression given the same level of efforts to tune the each model. Linear regression model cannot easily capture relationship between X features and y output, as they do not present linear relationship. Decision tree regression can be easily over-fitted and the test accuracy achieves only 50%, which is not satisfying as well. Very simple model, Lazy Learning with the non-weighted average kernel, can achieve good prediction accuracy without much efforts to tune the parameters and the selection of features. LL-Avg can achieve the linear correlation coefficient (R) 0.830762, RMSE 0.069550 and MAE 0.053041 (both are based on the 0-1 normalized dataset) based on the test dataset. Thus it is concluded that the Lazy Learning is the best prediction method among the three that we assess.

In the future, systematic way to select the good features should be studied. Also, different kernels for the LL should be tested. Last but not the least, current implementation of the LL is very computational intensive and it is worthy to search for the more efficient implementation of this method.

6. REFERENCES

[1] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne. Machine Learning Strategies for Time Series Forecasting, 2013.

[2] R. De Coninck and L. Helsen. Practical implementation and evaluation of model predictive control for an office building in Brussels. *Energy and Buildings*, 111:290–298, 2016.

[3] H. Huang, L. Chen, and E. Hu. A new model predictive control scheme for energy and cost savings in commercial buildings: An airport terminal building case study. *Building and Environment*, 89:203–216, 2015.

[4] A. Jain and R. Mangharam. Data Predictive Control for Peak Power Reduction. BuildSys, 2016.

[5] Y. Ma, J. Matuško, and F. Borrelli. Stochastic Model Predictive Control for Building HVAC Systems : Complexity and Conservatism. *IEEE TRANSACTIONS ON CONTROL SYSTEMS TECHNOLOGY*, pages 1–16, 2014.

[6] A. Mirakhorli and B. Dong. Occupancy behavior based model predictive control for building indoor climate: A critical review. *Energy and Buildings*, 129:499–513, 2016.

[7] J. Zhao, K. P. Lam, B. E. Ydstie, and V. Loftness. Occupant-oriented mixed-mode EnergyPlus predictive control simulation. *Energy and Buildings*, 2015.