Purdue University
Purdue e-Pubs

International High Performance Buildings Conference

School of Mechanical Engineering

July 2018

Apply Active Learning in Short-term Data-driven Building Energy Modeling

Liang Zhang
Drexel University, United States of America, lz356@drexel.edu

Jin Wen
Drexel University, United States of America, jw325@drexel.edu

Application of Active Learning in Short-term Data-driven Building Energy Modeling

Liang ZHANG[1]*, Jin WEN[1]

[1]Drexel University, Department of Civil, Architectural and Environmental Engineering,
Philadelphia, Pennsylvania, USA
E-mail: lz356@drexel.edu
E-mail: jw325@drexel.edu

* Corresponding Author

ABSTRACT

For better building control, and for buildings to be better integrated with the grid operation, high fidelity building energy forecasting model that can be used for short-term and real-time operation is in great need. With the wide adoption of building automation system (BAS) and Internet of things (IoT), massive measurements from sensors and other sources are continuously collected which provide data on equipment and building operations. This provides a great opportunity for data-driven building energy modeling.

However, the performance of data-driven based methods is heavily dependent on the quality and coverage of data. The collected operation data are often constrained to limited applicability (or termed as “bias” in this paper) because most of the building operation data are generated under limited operational modes, weather conditions, and very limited setpoints. The fact impedes the development of data-driven forecasting model as well as model-based control in buildings.

The proposed framework of active learning in short-term data-driven building energy modeling aims to choose or generate informative training data, either to defy data bias or to reduce labeling cost. In the framework, a disturbance categorization is applied to divide the disturbance space into several categories. Then, in each disturbance category, independently apply active learning strategy to decide the controllable inputs in the current time step. In this way, the variations of controllable inputs and disturbances are both considered.

In the case study, A virtual DOE reference office building with large-size and simulated in EnergyPlus environment is used as the testbed. A group of hierarchical setpoints, including zone temperature setpoint, supply air temperature and static pressure setpoints and chiller leaving water temperature setpoint, are the controllable inputs in this study. Regression tree is used as disturbance categorization algorithm and estimated error reduction is used as active learning algorithm. Improved model accuracy (lower testing error) is observed in the model trained by data from proposed framework, compared with models trained by normal operation data.

1. INTRODUCTION

In the United States, the buildings sector accounted for about 41% of primary energy consumption. Building control and operation strategies have a great impact on building energy efficiency and the development of building-grid integration. For better building control, and for buildings to be better integrated with the grid operation, high fidelity building energy forecasting model that can be used for short-term and real-time operation is in great need.

With the wide adoption of building automation system (BAS) and Internet of things (IoT), massive measurements from sensors and other sources are continuously collected which provide data on equipment and building operations. This provides a great opportunity for data-driven building energy modeling. However, the performance of data-driven based methods is heavily dependent on the quality and coverage of data. The collected operation data are often constrained to limited applicability (or termed as “bias” in this paper) because most of the building operation data are generated under limited operational modes, weather conditions, and very limited setpoints (often one or two fixed values, such as a constant zone temperature setpoint). For nonlinear systems, a data-driven model generated from biased data has poor scalabilities (when used for a different building) and extendibility (when used for different weather and operation conditions). The fact impedes the development of data-driven forecasting model as well as model-based control in buildings.

The design of task that aims to describe or explain the variation of information under conditions that are hypothesized to reflect the variation is termed as active learning in machine learning. The purpose is to choose or generate informative training data, either to defy data bias or to reduce labeling cost (when doing experiments in building is too expensive). Research on applying active learning in building energy modeling is relatively unexplored. Cai et al. generated an optimal training data set for varying zone temperature setpoints that maximizes the accuracy of parameter estimates for an intended building model structure. Optimal zone air temperature setpoints were determined to maximize the Fisher information matrix [1]. Jingran Ma et al. proposed and demonstrates the effectiveness of an economic model predictive control (MPC) technique in reducing energy and demand costs for building. A pseudorandom binary sequence (PRBS) is generated as the excitation input. The binary levels of the PRBS are the lower and upper bounds of the thermal comfort region, which in this work are set as 21 and 25 °C as zone temperature setpoint [2]. From the few existing researches, most of them only consider single operational setpoint (e.g. in [1-3], only zone temperature setpoint in considered in the active learning research), which is impractical for most real buildings where multiple setpoints in chillers, air handling units and air-conditioning terminals are used for building operation and control. Moreover, disturbances, especially weather and occupancy, in most cases are not considered.

In this research, an active learning framework in short-term data-driven building energy modeling is proposed. Introduced in Section 2, 3 and 4, the proposed framework considers the design of both controllable inputs (e.g. zone temperature setpoint, chiller leaving water temperature setpoint, etc.) and disturbances (outdoor air dry-bulb temperature, diffusive solar radiation, etc.). In the framework, a disturbance categorization is applied to divide the disturbance space into several categories in offline stage (Section 3). In online stage, at the start of each time step, the disturbance category of the current time step will be decided first. Then in each category, independently apply active learning strategy to decide the controllable inputs in the current time step (Section 4).

In the case study (Section 5), a virtual DOE reference office building with large-size and simulated in EnergyPlus environment is used as the testbed. A group of hierarchical setpoints, including zone temperature setpoint, supply air temperature and static pressure setpoints and chiller leaving water temperature setpoint, are the controllable inputs in this study. Regression tree is used as disturbance categorization algorithm and estimated error reduction is used as active learning algorithm.

2. FRAMEWORK OF ACTIVE LEARNING IN BUILDING ENERGY FORECASTING

The scope of this paper is active learning in short-term (hours ahead) building energy forecasting where buildings are currently operated under limited range of controllable setpoints but will be operated with a much wider range of controllable setpoints in the future, for applications like demand response with slowly changing setpoints where system dynamics is not considered. The task of active learning in short-term building energy forecasting model has two key concerns.

First, the time duration of active learning should be short enough so as not to hinder too much the normal use of building. Since the basic idea of active learning is to manipulate a model's controllable inputs to experience as much situations as possible to cover a large data space of controllable inputs, active learning tends to design the controllable setpoints go to extremes (such as the maximum or minimum value of controllable inputs). In this way, it will affect thermal comfort and unexpected situations in building operations. Another reason to consider time duration of active learning is that in practice, we only have limited time to do active learning, which is termed as "budget" in active learning. For example, when building cooling energy forecasting model in summer time, if a small part of summer time can be used for training, then we only have weeks or even days of time for generating training data using active learning. This fact demands the active learning to be efficient enough to generate most informative controllable setpoint design in a short period of time. And that is why active learning is needed instead of full factorial design where we can exhaust every possibility of training data regardless of training time.

Second, active learning should consider disturbances like weather and occupancy factors into the design of controllable setpoints. It is well-known that building energy is a weather-driven model. Weather not only directly impact the cooling/heating load of building, but also indirectly impact the human behavior which also has a great impact on building energy consumption. However, as the dominant factor impacting building energy as well as uncontrollable inputs, disturbances like weather and occupancy cannot be directly designed. Not like controllable setpoints like zone temperature setpoint, weather disturbances cannot be designed in advance. For example, we cannot say that we want to "manipulate" outdoor air temperature to be 20℃ during 8:00 to 9:00 tomorrow. Only passive and indirect method could be applied to consider weather and occupancy disturbance (For example, as soon as outdoor air temperature is 20℃, we start to do certain designed experiments).

The proposed framework of active learning in building energy forecasting model starts from the two key concerns mentioned above, and is designed to ensure an efficient way to get informative experiment design of controllable input, and at the same time, considers the dominant disturbances which cannot be actively designed in building energy forecasting model.

To visualize the framework, as can be seen from Figure 1, it starts from dividing the data into three categories: disturbances, controllable inputs and output. Disturbances are features that cannot be controlled, and most of disturbances are weather or occupancy related. Typical disturbances that impact building energy consumption are but not limited to: outdoor dry-bulb temperature, outdoor wet-bulb temperature, direct solar radiation, diffusive solar radiation, wind speed, occupancy, etc. Controllable inputs (setpoints) include features that are used for building operation control. They are the primary target of active learning in this study. In other words, we want to study the strategy to manipulate controllable setpoints so that informative training data can be generated. Typical controllable inputs include zone temperature setpoint, supply air static pressure setpoint, supply air temperature setpoint, chiller leaving water temperature setpoint, lighting schedule, etc. The third category of data is output, which is normally building energy consumption (in certain period of time) or power (in certain point of time) in building energy forecasting modeling.

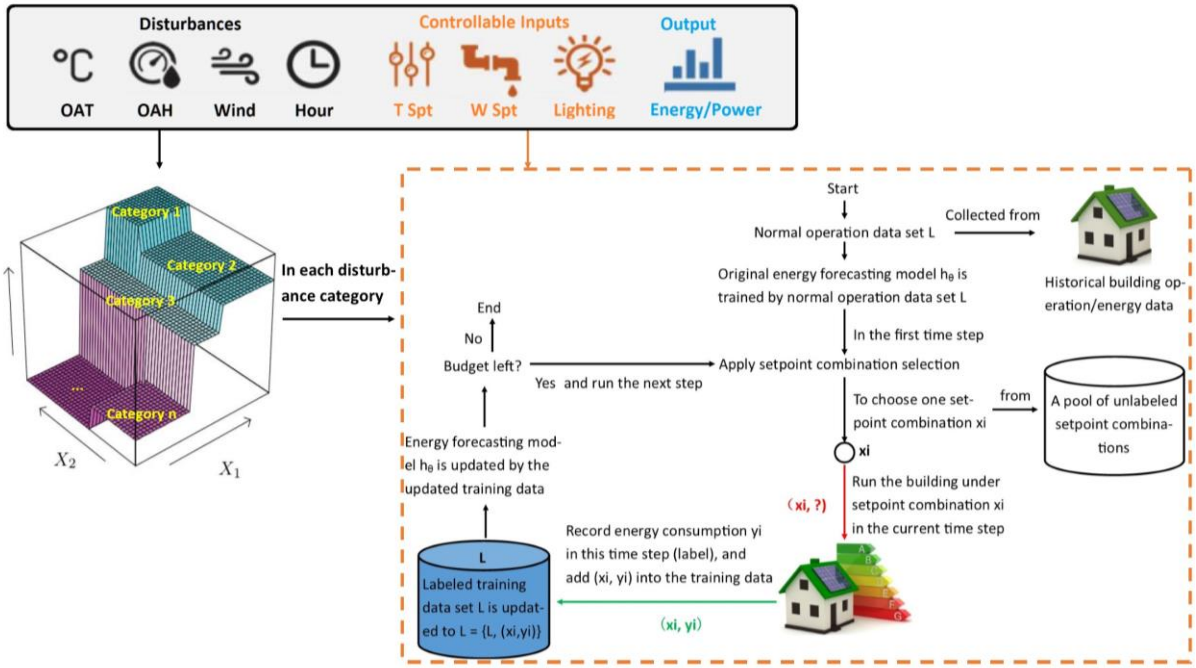

Disturbances
OAT
OAH
Wind
Hour
Controllable Inputs
T Spt
W Spt
Lighting
Output
Energy/Power
Category 1
Category 2
Category 3
...
Category n
X_2
X_1
In each disturbance category
Start
Normal operation data set L
Collected from
Historical building operation/energy data
Original energy forecasting model h_θ is trained by normal operation data set L
In the first time step
Apply setpoint combination selection
End
No
Budget left?
Yes and run the next step
To choose one setpoint combination xi
from
A pool of unlabeled setpoint combinations
xi
Run the building under setpoint combination xi in the current time step
(xi, ?)
Energy forecasting model h_θ is updated by the updated training data
Record energy consumption yi in this time step (label), and add (xi, yi) into the training data
L
Labeled training data set L is updated to L = {L, (xi,yi)}
(xi, yi)
A
B
C
D
E
F
G

Figure 1: Framework of active learning in building energy forecasting

Next, disturbances categorization will be conducted. Since the impact of disturbances is very important but cannot be manipulated, disturbance categorization (or block design) is used here to solve this conflict. In the statistical theory of the design of experiments, blocking is the arranging of experimental units in groups (blocks) that are similar to one another. Blocking reduces variability. Its principle lies in the fact that a variability that cannot be overcome. In this study, the variability that cannot be overcome is the disturbances such as weather features that cannot be manipulated in the building energy modeling process. Since they cannot be manipulated, to reduce their variability that twisted with the impact of controllable inputs to building energy, the disturbance categorization (block design) is used to group disturbances into several categories (blocks). In the same category (block), the impact of disturbance on energy is similar or linear. Each category (block) represents a set of disturbances or weather under which the relationship between controllable inputs and energy consumption is clearer (has less variability). During the disturbances categorization, classification algorithm will be applied to decide the number of disturbance categories, key disturbance used for categorization and their specific values to divide the disturbance space. Details of disturbances categorization (block design) used in this case is further introduced in Section 4.

After the disturbance space is divided into k categories, in the online stage, active learning on controllable inputs will be conducted in each category individually. Then, setpoint combination selection strategy, or query selection strategy, is applied. Setpoint combination (query) selection strategy is the core of active learning. It decides the next values of setpoint combination set as the controllable input values in the next time step. (e.g. 26℃ for zone temperature setpoint and 6℃ for chiller leaving water temperature setpoint). The main idea of setpoint combination (query) selection strategy is to selection the most potentially informative or uncertain controllable inputs in the setpoint combination pool (pool-based method) as the next query. Details of setpoint combination (query) strategy will be further illustrated in Section 3.

To sum up, the proposed framework of active learning in building energy forecasting model considers the two key concerns in active learning. It is designed to ensure an efficient way to get informative experiment design of controllable input using query selection strategy, and at the same time, considers the dominant disturbances which cannot be actively designed using disturbance categorization (block design). In the next two sections, setpoint combination (query) selection strategies and disturbance categorization (block design) will be discussed in detail.

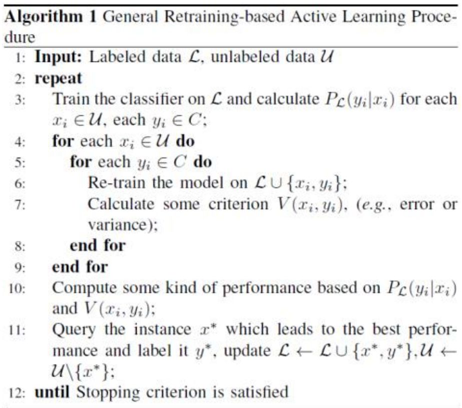
3. SETPOINT COMBINATION (QUERY) SELECTION STRATEGIES

Active learning evaluates the informativeness of unlabeled instances (unknown building energy consumption with known disturbances and controllable inputs, in the context of building energy forecasting), which can either be generated from scratch (without any historical labeled data) or sampled from a given distribution (with historical labeled data). There have been many query strategies in the literature.

Lewis and Gale [4] proposed an active learning strategy that queries the instances about which it is least certain how to label. This approach is often straightforward for probabilistic learning models. For example, when using a probabilistic model for binary classification, uncertainty sampling simply queries the instance whose posterior probability of being positive is nearest 0.5. This method is suitable for probabilistic learning models.

Another, more theoretically-motivated query selection strategy is the query-by-committee (QBC) algorithm. Proposed by Seung et al.[5], the QBC approach involves maintaining a committee of models which are all trained on the current labeled set L, but represent competing hypotheses. Each committee member is then allowed to vote on the labelings of query candidates. The most informative query is considered to be the instance about which they most disagree.

Expected model change, proposed by Settles et al.[6],uses a decision-theoretic approach, selecting the instance that would impart the greatest change to the current model if we knew its label. The intuition behind this method is that it prefers instances that are likely to most influence the model (i.e., have greatest impact on its parameters), regardless of the resulting query label. This approach has been shown to work well in empirical studies, but can be computationally expensive if both the feature space and set of labelings are very large.



Algorithm 1 General Retraining-based Active Learning Procedure

```
Input: Labeled data $\mathcal{L}$, unlabeled data $\mathcal{U}$
repeat
    Train the classifier on $\mathcal{L}$ and calculate $P_{\mathcal{L}}(y_i|x_i)$ for each $x_i \in \mathcal{U}$, each $y_i \in C$;
    for each $x_i \in \mathcal{U}$ do
        for each $y_i \in C$ do
            Re-train the model on $\mathcal{L} \cup \{x_i, y_i\}$;
            Calculate some criterion $V(x_i, y_i)$, (e.g., error or variance);
        end for
    end for
    Compute some kind of performance based on $P_{\mathcal{L}}(y_i|x_i)$ and $V(x_i, y_i)$;
    Query the instance $x^*$ which leads to the best performance and label it $y^*$, update $\mathcal{L} \leftarrow \mathcal{L} \cup \{x^*, y^*\}, \mathcal{U} \leftarrow \mathcal{U} \backslash \{x^*\}$;
until Stopping criterion is satisfied
```

Figure 2: General framework and procedure of decision-theoretic approach [7]

Another decision-theoretic approach, expected error reduction, aims to measure not how much the model is likely to change, but how much its generalization error is likely to be reduced. Roy and McCallum first proposed the expected error reduction framework for text classification using naive Bayes [8]. In most cases, unfortunately, expected error reduction is also the most computationally expensive query framework. Not only does it require estimating the expected future error over U for each query, but a new model must be incrementally re-trained for each possible query labeling, which in turn iterates over the entire pool. This leads to a drastic increase in computational cost.

The general framework and procedure of the above two decision-theoretic approach (expected model change and expected error reduction) is shown in Figure 2.

Another widely used approach is called variance reduction. It is derived from statistical theories of optimal experimental design. A key ingredient of this approach is Fisher information, which is sometimes written as I(θ) to make its relationship with model parameters explicit. Formally, Fisher information is the variance of the score, which is the partial derivative of the loglikelihood function with respect to the model parameters:

$$I(\theta) = N \int_x P(x) \int_y P_\theta(y|x) \frac{\partial^2}{\partial \theta^2} log P_\theta(y|x) \tag{1}$$

where there are N independent samples drawn from the input distribution. θ is model parameters, and x/y is model inputs/outputs. To minimize the variance over its parameter estimates, an active learner should select data that maximizes its Fisher information (or minimizes the inverse thereof).

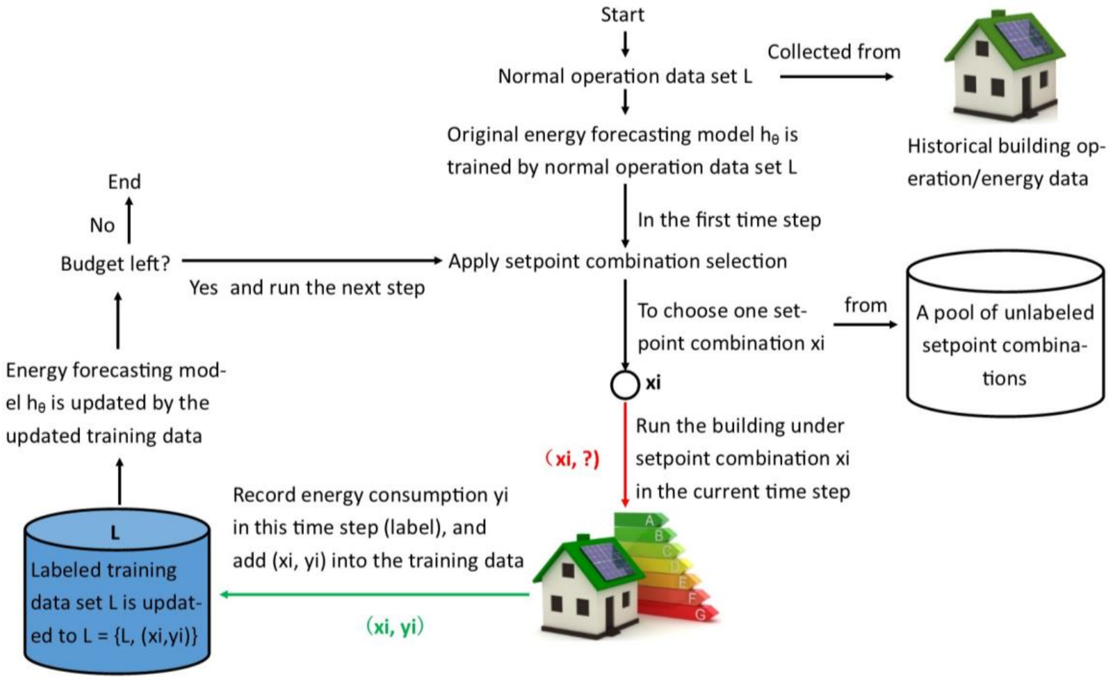
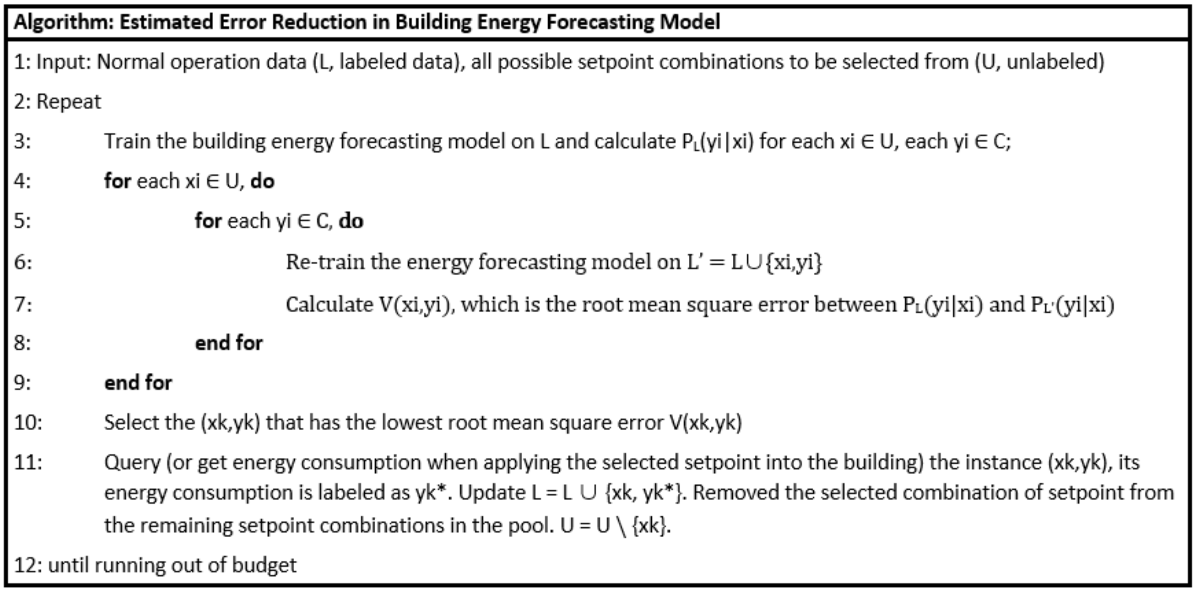

Start
Normal operation data set L
Collected from
Historical building operation/energy data
Original energy forecasting model h$_\theta$ is trained by normal operation data set L
In the first time step
End
No
Budget left?
Yes and run the next step
Apply setpoint combination selection
To choose one setpoint combination xi
from
A pool of unlabeled setpoint combinations
xi
Run the building under setpoint combination xi in the current time step
(xi, ?)
Energy forecasting model h$_\theta$ is updated by the updated training data
Record energy consumption yi in this time step (label), and add (xi, yi) into the training data
L
Labeled training data set L is updated to L = {L, (xi,yi)}
(xi, yi)
A
B
C
D
E
F
G

Figure 3: Proposed setpoint combination selection strategy

Figure 3 shows the procedures of setpoint combination selection strategy proposed in this paper. The procedures start with training energy forecasting model by normal operation data, which is generated under very limited setpoints. The setpoint combination selection strategy is applied to choose one setpoint combination xi from a pool of setpoint combinations, making it a pool-based active learning (different from stream-based active learning). After selecting the most informative setpoint combination xi, operate the building under setpoints of xi, and record energy consumption in this time step (yi). (xi,yi) is the new labeled data. Include (xi,yi) into the existing training data set and form the new training data set L = {L, (xi,yi)}. And updated energy forecasting model with the updated training data. Repeat the above procedures until we are running out of budget (maximum time steps for active learning).

In this study, expected error reduction will be used as the setpoint combination selection strategy because this approach has been shown to work well in empirical studies. Although it can be computationally expensive, both the feature space and set of labeling are not very large in the case study and the computational cost will be acceptable. The algorithm of estimated error reduction in the context of active learning of building energy forecasting model is shown in Figure 4.



```
Algorithm: Estimated Error Reduction in Building Energy Forecasting Model
Input: Normal operation data (L, labeled data), all possible setpoint combinations to be selected from (U, unlabeled)
Repeat
    Train the building energy forecasting model on L and calculate $P_L(yi|xi)$ for each $xi \in U$, each $yi \in C$;
    for each $xi \in U$, do
        for each $yi \in C$, do
            Re-train the energy forecasting model on $L' = L \cup \{xi,yi\}$
            Calculate V(xi,yi), which is the root mean square error between $P_L(yi|xi)$ and $P_{L'}(yi|xi)$
        end for
    end for
    Select the (xk,yk) that has the lowest root mean square error V(xk,yk)
    Query (or get energy consumption when applying the selected setpoint into the building) the instance (xk,yk), its
    energy consumption is labeled as yk*. Update $L = L \cup \{xk, yk^*\}$. Removed the selected combination of setpoint from
    the remaining setpoint combinations in the pool. $U = U \setminus \{xk\}$.
until running out of budget
```

Figure 4: Pseudo-code of estimated error reduction in the context of active learning of building energy forecasting

4. DISTRUBANCE CATEGORIZATION (BLOCK DESIGN)

Disturbance categorization in this research is used to consider disturbance in active learning framework. Disturbance categorization itself is a classification problem. It classifies different disturbances (especially weather) into different categories. In each category, the variance caused by disturbance is negligible or explainable.

There is a constraint in this particular classification problem. As mentioned in Section 2, since time duration of active learning should be short enough so as not to hinder too much the normal use of building, the number of groups should be limited. The reason is that, in the limited number experiments, too many categories will lead to too few experiment in each block. Too few experiments in each category will cause the insufficient variation of controllable variables, which will make the model lose power to reflect the nonlinear impact of controllable inputs (setpoints) on building energy consumption. Meanwhile, too few number of blocks will also cause problems. The most extreme case is that there is only one category. In this way, the disturbances will not be considered in the active learning process. This will lead to poor model performance because some disturbances (especially weather) in building energy forecasting model are dominant factors (even more important than certain controllable inputs).

In the context of building energy forecasting model, the procedures of disturbance categorization are described as follows. As an offline procedure, the blocks are designed before the setpoint combination selection strategy. Collect existing normal operation data, including their controllable inputs, disturbances, and building energy consumption. Take disturbances as independent variables and building energy consumption as dependent variables to perform a classification algorithm. Then we get a classification model that categorizing similar disturbances that causes relatively similar (or linear) building energy consumption. Before each time step when we do setpoint combination selection, the classification model is used to first decide which category the disturbance belongs. And in each disturbance category, the setpoint combination strategy is conducted individually.

Since disturbance categorization is a classification problem, many mature classification algorithms can be used in disturbance categorization. In this study, decision tree algorithm is used. A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning. It has also been used in short-term building energy forecasting model[9].

5. CASE STUDY

5.1 Virtual Building Testbed

The case uses a virtual building testbed developed by the U.S. Department of Energy (DOE) and National Renewable Energy Laboratory [10] using EnergyPlus [11] environment. The virtual building is a large-sized office building. It is a twelve-story office building with 46320 m^2 floor area. The building uses water cooled chiller for cooling and boiler for heating. Table 1 provides more detailed building description and data. Figure 5 illustrates the outlook of the building. The weather file used for the simulations is based on Philadelphia, PA (USA_PA_Philadelphia.Intl.AP.724080 TMY3). In the case study, the key controllable inputs for future building control are zone temperature setpoint, air handling unit supply air temperature setpoint, and chiller leaving water temperature setpoint. In this case, all data used for normal operation, training, validation, and testing are generated using this virtual building testbed. July 1st to July 31st (31 days) using the above discussed weather file are simulated in the case study. Among these simulated days, the first 21 days (July 1st to July 21st) are functioned as normal operation data, when the building is operated under single setpoint combination: zone temperature setpoint = 24℃, supply air temperature setpoint = 12.8℃, chiller leaving water temperature setpoint = 6.7℃. The classification model for disturbance categorization is established using normal operation data. The following seven days (July 22nd to July 28th) are used for active learning. The operation of three setpoints considering disturbances will be designed according to the proposed active learning framework. The final three days (July 29th to July 31st) are used for testing. The simulation time step is 60 minutes. Therefore, all virtual measurements (features) have a sample interval of 60 minutes.

Table 1: Description of the virtual testbed.

Reference Large-size Office Building	
Floor area (m^2)	46320
Length-to-width ratio	1.5
Number of floors	12
Floor-to-floor height (m)	3.96
Floor-to-ceiling height (m)	2.74
Window-to-wall ratio	0.38
Construction	
Roof	Insulation entirely above deck
Wall	Mass
HVAC Equipment Types	
Heating	Boiler
Cooling	Water cooled chiller
Air distribution	Multi-zone variable air volume
Lighting and Elevators	
Lighting assumptions	ASHARE 90.1-2004
Number of elevators	12
Motor power (W/each)	18537
Hot Water	
Peak use rate (per floor) (L/h)	80.6
Temperature and fixture (°C)	43
Building Activity	
Data source of principal building activity	EIA 2005

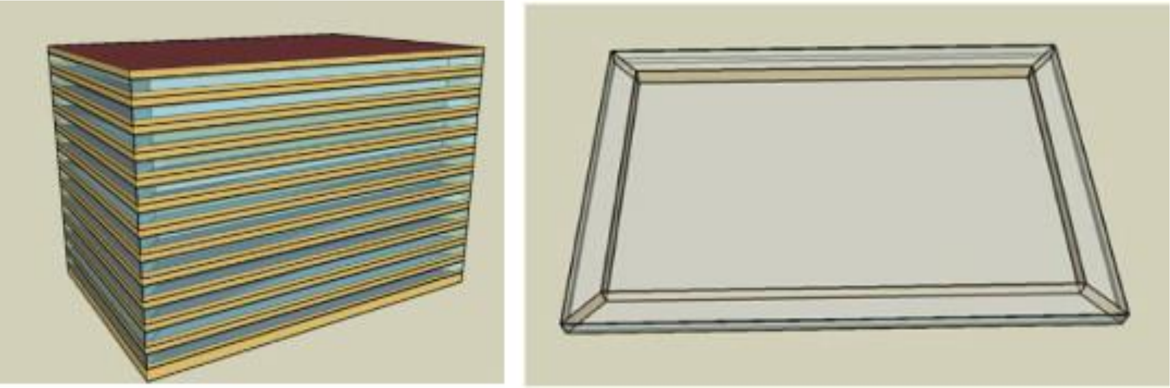


Figure 5: The 3D and plan views of large-size office building

5.2 Disturbance Categorization

In the case study, decision tree algorithm is used. Standard CART algorithm [12] is used to create decision trees. It performs the following steps: (a) Start with all input data, and examine all possible binary splits on every predictor. (b)Select a split with best optimization criterion. (c) Impose the split. (d) Repeat recursively for the two child nodes. The explanation requires two more items: description of the optimization criterion, and stopping rule. For the optimization criterion, Gini's diversity index is used as the split criterion and minimizing mean-squared error is used as regression criterion. The decision tree CART algorithm is realized in MATLAB using built-in function of *"fitctree"*.

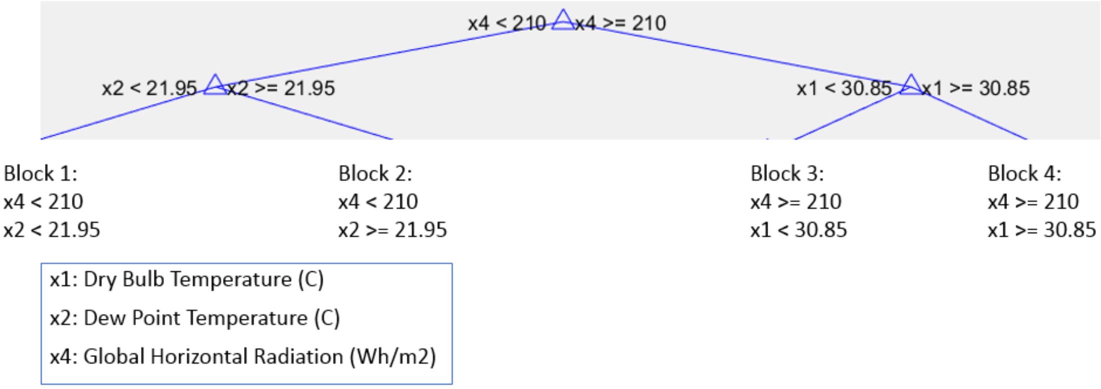


Figure 6: Result of disturbance categorization in virtual testbed during normal operation (July 1st to 21st)

The disturbances considered in this case is limited to weather disturbances. The potential disturbances used in the block design include: dry bulb temperature, dew point temperature, relative humidity, extraterrestrial horizontal radiation, extraterrestrial direct normal radiation, horizontal infrared radiation intensity from sky, global horizontal radiation, direct normal radiation, diffuse horizontal radiation, global horizontal illuminance, direct normal illuminance, diffuse horizontal illuminance, zenith luminance, wind direction, wind speed, total sky cover, opaque sky cover, and visibility. The values of these weather disturbances are from TMY 3 of Philadelphia International Airport. Together with the energy consumption that operated under normal operation (from July 1st to July 21st, zone temperature setpoint = 24℃, supply air temperature setpoint = 12.8℃, chiller leaving water temperature setpoint = 6.7℃), a regression tree method is conducted in MATLAB, and Figure 6 shows the result. Among eighteen weather disturbances, three variables: dry bulb temperature, dew point temperature and global horizontal radiation, are selected to divide the data space into four parts.

5.3 Setpoint Combination (Query) Selection Strategies

Blocks are designed offline in Section 5.2. Next, the setpoint combination selection strategies will be applied online. As time proceeds, one combination of setpoints will be selected in each time step. It is selected as the following procedures. First, get disturbances in the current time step and decide the disturbance category it belongs according to the classification model built in Section 5.2. Then, in each disturbance category, conduct setpoint combination selection strategy to decide one combination of setpoint to be operated in the building. At the end of this time step, record energy consumption of the building under the selected combination of setpoint. Include the selected setpoint combination and its energy consumption as new training data generated in this time step into the existing training data. Finally, update the building energy forecasting model trained by updated training data, until it is running out of budget (maximum time step of active learning).

The length of time step in this case is 3 hours because of the stabilization time of the reference office building is 88.0 minutes[13]. In this way, we can make sure that steady states account for most of the operation time and the system is simplified to non-dynamic system. Since each experiment lasts 3 hours, and the total active learning lasts 168 hours (7 days). The budget, which is the maximum number of active learning, can be calculated to be 56 experiments. In other words, 56 setpoint combinations will be selected from the setpoint combination pool.

Unlabeled setpoint combination pool in this case is full-factorial design of three controllable inputs: zone temperature setpoint (20-25.6°C), AHU supply air temperature setpoint (12.8-18.3°C) and chiller leaving water temperature setpoint (6.7-10 °C). The full factorial design is in 1.1 °C discretization. Altogether, 6*6*4=144 combination of setpoint is in the pool. Active learning strategy will select 56 setpoint combinations from the pool with 144 setpoint combination.

Estimated error reduction is applied to select setpoint combination in each time step. The algorithm of estimated error reduction in the context of active learning of building energy forecasting model is shown in Figure 4.

To train the energy forecasting model, Multivariate Adaptive Regression Splines (MARS) is used as the data-driven modeling algorithm. MARS is a nonparametric regression which constructs underlying relationship from a set of coefficients and basis function that are determined by training data. MARS operates as multiple piecewise linear regression, where each split defines the region of application for a particular linear regression equation. MARS has been used in short-term building energy forecasting and shown high forecasting performance in previous research [14, 15]. In this research, ARES Toolbox v1.51[16] operated in MATLAB environment is used for MARS model development.

5.4 Results and Comparisons

Test days (July 29th to July 31st) are used to evaluate the proposed method. Test days are operated under two scenarios. The first scenario is a typical demand response operation with changing zone temperature setpoints, air handling unit supply air temperature setpoint and chiller leaving water temperature setpoint. The operation of setpoints are the same in Braun's research of setpoint for building optimal control (demand response) in Iowa Energy Center[17].

The second test scenario is normal operation with constant setpoints: zone temperature setpoint = 24℃, supply air temperature setpoint = 12.8℃, chiller leaving water temperature setpoint = 6.7℃.

The model trained by proposed active learning method will be compared with the model trained by normal operation data. The criteria to evaluate model performance is normalized root mean square error (NRMSE, Equation 4) between forecasted building energy consumption with measured building energy consumption in the test period.

$$\mathrm{NRMSE} = \sqrt{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2 /n}/(y_{max} - y_{min}) \tag{4}$$

where $\hat{y}_t$ hat and y_t are predicted and measured values of model output (hourly cooling energy consumption in this case), y_{max} and $y_{min}$are maximum and minimum values of model output, and n is the total number of training data. The result of test is shown in Table 2

Table 2: Comparison between model built by normal operation data and by proposed active learning in test period

(NRMSE)	Model built by normal operation data	Model built by proposed active learning framework
Training Error	0.0353	0.0476
Scenario 1: normal operation test	0.0489	0.0537
Scenario 2: demand response test	0.1312	0.0895

6. CONCLUSIONS

The proposed framework of active learning in building energy forecasting model considers the two key concerns in active learning. It is designed to ensure an efficient way to get informative experiment design of controllable input using setpoint combination selection strategy, and at the same time, considers the dominant disturbances which cannot be actively designed using disturbance categorization. The case study has shown that:

- Though with good performance in normal operation test, the very poor performance in demand response test has shown that energy forecasting model trained only by normal operation data have very poor extendibility and generalization.
- Poor performance of building energy forecasting model built by normal operation data in demand response test has indicated the importance of active learning in building energy forecasting model.
- Compared with model built by normal operation data, the model trained by data generated by proposed active learning framework has much better model accuracy in demand response test and similarly good performance in normal operation test.
- Energy forecasting model trained only by data with proposed framework has good extendibility and generalization.

REFERENCES

1. Cai, J., et al. *Optimizing Zone Temperature Setpoint Excitation to Minimize Training Data for Data-driven Dynamic Building Models*. in *American control conference, submitted, Boston, July*. 2016.
2. Ma, J., et al., *Demand reduction in building energy systems based on economic model predictive control.* Chemical Engineering Science, 2012. **67**(1): p. 92-100.
3. Zhang, L., et al., *Experiment Design and Training Data Quality of Inverse Model for Short-term Building Energy Forecasting.* 2016.
4. Lewis, D.D. and W.A. Gale. *A sequential algorithm for training text classifiers*. in *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*. 1994. Springer-Verlag New York, Inc.
5. Seung, H.S., M. Opper, and H. Sompolinsky. *Query by committee*. in *Proceedings of the fifth annual workshop on Computational learning theory*. 1992. ACM.
6. Settles, B., M. Craven, and S. Ray. *Multiple-instance active learning*. in *Advances in neural information processing systems*. 2008.
7. Yang, Y. and M. Loog. *Active learning using uncertainty information*. in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. 2016. IEEE.
8. Roy, N. and A. McCallum, *Toward optimal active learning through monte carlo estimation of error reduction.* ICML, Williamstown, 2001: p. 441-448.
9. Yu, Z., et al., *A decision tree method for building energy demand modeling.* Energy and Buildings, 2010. **42**(10): p. 1637-1646.
10. Deru, M., et al., *US Department of Energy commercial reference building models of the national building stock.* 2011.
11. Crawley, D.B., et al., *EnergyPlus: creating a new-generation building energy simulation program.* Energy and buildings, 2001. **33**(4): p. 319-331.
12. Breiman, L., et al., *Classification and Regression Trees. Wadsworth, 1984.* Intelligence, 1993: p. 1002-1007.
13. Li, X., *Net-zero Building Cluster Simulations and On-line Energy Forecasting for Adaptive and Real-Time Control and Decisions*. 2015: Drexel University.
14. Williams, K.T. and J.D. Gomez, *Predicting future monthly residential energy consumption using building characteristics and climate data: A statistical learning approach.* Energy and Buildings, 2016. **128**: p. 1-11.
15. Cheng, M.-Y. and M.-T. Cao, *Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines.* Applied Soft Computing, 2014. **22**: p. 178-188.
16. Jekabsons, G., *Adaptive Regression Splines toolbox for Matlab/Octave.* Version, 2013. **1**: p. 72.
17. Braun, J.E., *Load control using building thermal mass.* Journal of solar energy engineering, 2003. **125**(3): p. 292-301.