Study on Occupancy Prediction for Building Operation using Machine Learning Method

Yuan Jin, Da Yan, Hongsan Sun
Tsinghua University, Beijing, PR China

Abstract

Occupancy in buildings is an important input no matter to building design or building on-site operation. As for building design, the occupancy state is related to the heat gain in buildings and further equipment operation. With room occupied, the comfortable ambient parameter is required and the equipment in buildings are potential to be used. As for the building operation, with the detection of occupancy, the auto control is likely to perform. However, the accuracy needs improved. In this study, occupancy state is predicted by proposed machine learning method. During the occupancy prediction, the temporal sequential characteristic is considered to obtain a better prediction performance. The evaluation indicator is defined and the evaluation is conducted through ten fold cross validation. This study can be further promoted to building operation and performance simulation.

Introduction

With the accumulating insight of occupant behaviour in buildings by researchers and engineers, occupancy has been a crucial input to be considered during the analysis of building energy performance as well as building equipment operation (Dong et al., 2018).

During the building design phase, the occupancy is considered to be related with heat gain in each room. The traditional design handbook and standard consider the occupancy as stable occupied pattern (Standard and Chapter, 2004, GB50189-2005, 2005). This can represent the basic and average performance of the

occupancy pattern. However, from the on-site measurement in offices and buildings, the occupancy possesses the spatial and temporal stochastic characteristic (Wang et al., 2005, Wang et al., 2011). Therefore, many stochastic models depicting the occupancy in buildings are proposed. The models include non-homogeneous Poisson process model (Wang et al., 2005), inhomogeneous two-state Markov Chain (Page et al., 2008), agent-based model (Liao et al., 2012) and first-order homogeneous model (Wang et al., 2011). These models can well represent the stochasticity of the occupancy, but the evaluation is in need for these simulation processes.

During the building operation phase, researchers need the state of occupancy to perform the equipment control and regulate the thermostat. The occupied state of the room refers that the ambient parameter should meet the requirement of the thermal comfort and that the fresh air should be supplied accordingly. There have been many researches about the occupancy prediction and ongoing building system control and operation.

The application of building operation can be divided into two parts. The on-site prediction for the control and future prediction to analyse the control with optimized target.

Kim and Moon (2018) proposed one environmental data driven model for occupancy status detection and an integrated comfort algorithm for operating the indoor devices. The target is to decide three status of human in the room, which are away, active and inactive. Here the

inactive refers to sleeping state. Different status is related to different set points. The occupancy state is predicted by multinomial logistic regression model with three inputs: occupancy data from PIR sensor, the concentration of CO₂ and the energy use of lighting system. The prediction performance is evaluated by prediction error which is:

$$\text{prediction error} = \frac{N_{\text{exact}}}{N_s} \times 100 \quad (1)$$

where N_{exact} refers to the number of exact predictions and N_s refers to total number of predictions.

The method structure of the research is as Figure 1 shows.

Figure 1: Method structure of on-site occupancy prediction and thermal-comfort-based control algorithm.

Another area of application utilizes the results of occupancy prediction in a certain temporal period. The integrated performance and energy consumption in this period is evaluated to choose the optimized control strategy for building operation and control for current timestep (Dong, 2010). In this research, the occupancy state is predicted from two aspects. The amount of occupants is predicted by artificial neural network and hidden Markov chain with the input of CO₂ concentration. The status of the occupants is predicted by semi-Markov model with multi-inputs of acoustics, lighting, motion, CO₂ concentration and temperature & relative humidity. The control method is called non-linear model predictive control and this method uses the Root-Mean-Square Error (RMSE) as the evaluation indicator. The method structure of this research is illustrated in Figure 2.

Figure 2: Method structure of future occupancy prediction and model predictive control.

To conclude, in the field of occupancy in buildings research, the study is categorized as occupancy simulation and prediction. The simulation study contributes to the building design and standard compilation, while the prediction study contributes to the building on-site operation. The summary of occupancy simulation and prediction is as shown in Table 1. This table summarises potential used input data and methods during the process of occupancy simulation and prediction.

Table 1: Input data and prediction summary for study of occupancy in buildings.

Study of occupancy in buildings	Input data	Method (Model)
Occupancy simulation	Occupancy pattern; measured occupancy data	Markov chain model; random walk model; logistic regression; decision tree; agent-based model; survival analysis
Occupancy prediction	Historical occupancy pattern; CO ₂ concentration; temperature; relative humidity	k-nearest neighbor (KNN); random sampling; artificial neural network; support vector machine; Bayesian method

Research in This Study

The reviewed occupancy prediction methods are normally based on ambient parameters as well as the motion sensor. The model input relies on multi-category sensors and the predicted occupancy state is only related to the current time point, without the consideration of the temporal sequential characteristic.

Therefore, this study tries to analyse the features of occupancy state in one office and contribute the analysis results to the occupancy state prediction. The occupancy state prediction is only based on the temporal sequential feature of the room. The occupancy state in this study only focuses on whether the room is occupied or not, but no occupant number prediction. In this study, a systematic prediction and evaluation method are proposed and analysed.

Methodology

The methodology of this research is illustrated in Figure 3. The methodology of this research is divided into three parts:

Figure 3: Research methodology map.

- (1) The acquisition and pre-processing of the occupancy data. In this study, only the historical occupancy data are used for occupancy prediction, for this study attempts to analyze the temporal sequential feature and explore the feasibility of occupancy state prediction based on only historical occupancy data. The occupancy data are collected with one PIR (Passive infrared) sensor. In the research stage, the ground truth data are from the on-site survey. The objective of the data processing is to analyse the occupied routine of the room. The routine is also for further occupancy state prediction.
- (2) The prediction method. In this study, artificial neural network machine learning method is used for the occupancy state prediction. Based on the objective to analyse the temporal feature of occupancy state, the three types of input are considered in this study: historical occupied state, time point and historical

occupied ratio. The model parameter optimization is also considered. In this study, the model selection and parameter tuning work include: network size, activation function, solver and L2 penalty parameter (α). After the search in fitting results, the input and parameter tuning can be confirmed.

- (3) The model evaluation. In this research, three indicators are proposed to evaluate the prediction performance. The confusion matrix is used to analyse the prediction results in detail. The accuracy refers to the proportion for exact prediction. The RMSE is also an indicator depicting the overall prediction performance. The dataset is divided into training set and testing set. The ten-fold cross evaluation is conducted to ensure the results.

Data Acquisition

In this research, only motion sensor of passive infrared is used to get the measurement data. The data quality of the PIR sensor is firstly analysed for raw data pre-processing. The obtained data is shown in Figure 4. The measurement duration is one hour, and the room is continuously occupied. However, the PIR sensor continually reports the room unoccupied.

Figure 4: Comparison of the real occupancy data and occupancy data from PIR sensor.

Figure 5: Duration definitions during PIR occupancy data process: (a) duration of occupied time from PIR, (b) duration of unoccupied time from PIR.

Combined with curve characteristic from PIR sensor, the duration of detected occupied and the duration of detected unoccupied are defined (Figure 5). After the

analysis of histogram of duration (a) and (b), the data pre-processing is on one basic assumption that the staying period will be no less than 5 minutes.

Prediction Method

This study utilized the method, which took multi-layer perceptron based classification as the algorithm (Gardner and Dorling, 1998). Based on historical occupancy routine, this method effectively predicted occupancy state of the room.

Neutral networks are a series of artificial intelligence, and multi-layer perceptron is one of the useful types of network study, defined as a feed-forward neural network (Schalkoff, 1992, Jain et al., 1996). The model of multi-layer perception neural network contains inputs and outputs, and the relationship of the variables needs training through different algorithms. The model can be adapted to non-linear functions and accurately trained. The model consists of a series of inner system with connected neurons. With multiple input variables and output variables, the neurons in between are called hidden layers. The connection refers to different simple nonlinear transfer functions, with different connecting weight coefficient. Figure 6 shows a one hidden layer perceptron neural network structure (Pedregosa et al., 2011). The function can be depicted as:

$$[o_1, \dots] = f([i_1, i_2, i_3, \dots]) \quad (2)$$

The function refers to the relationship between input variables and output variables. The function needs to be trained with the training data set. Here, the input and output variables can be represented as vectors. During the model training, the data of transfer function and connecting weight coefficient are repeatedly changed and determined to optimize the input-output mapping.

The optimization refers to minimizing the error values. The most basic training algorithm is back-proagation using gradient descent procedure to locate the global minimum of the error surface (Rummelhart, 1986), which is the most computationally straightforward algorithm for training the multilayer perceptron.

Figure 6: Perceptron neural network with one hidden layer.

In this research, the prediction model needs optimization to get the best prediction performance. There are four dimensions to optimize the prediction model: neural network size, activation function, solver method and L2 penalty parameter (α). The detailed items are listed in Table 2.

Table 2: Detailed items for model selection and parameter tuning.

Item	Parameters
Neural network size	layer size: 1, 2; neuron size: 10, 20, 50, 70, 100, 120
Activation function	identity, logistic, tanh, relu
Solver method	lbfgs, adam
L2 penalty parameter (α)	0.01, 0.1, 1, 10, 100

For neural network size, in this study, the hidden layer size is set as 1 and 2. In each layer, the neuron amounts are 10, 20, 50, 70, 100, 120 separately.

For activation function, four functions are chosen, the detailed function formation is as follows:

The identity function:

$$f(x) = x \quad (3)$$

The logistic function:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (4)$$

The tanh function:

$$f(x) = \tanh(x) \quad (5)$$

And the relu function:

$$f(x) = \max(0, x) \quad (6)$$

For solver method, two solvers are chosen for error optimization, which are Limited memory BFGS (L-BFGS) method (Le et al., 2011, Liu and Nocedal, 1989) and ADAM method (optimized SGD) (Kingma and Ba, 2015). L-BFGS method is highly competitive mainly for low dimensional problem, while ADAM is more appropriate for high dimensional problem and outperforms L-BFGS solver.

L2 penalty parameter (α) refers to a weight coefficient of regularization term in the equation of minimized function. The value of α is bigger, then the model tends to be simpler, in reverse, the value of α is smaller, the model tends to be more complicated, and even overfitting. In this study, the penalty parameter (α) is set as 0.01, 0.1, 1, 10 and 100 separately.

According to the parameter tuning mentioned above, the prediction model selection and parameter tuning are conducted.

One more important issue is to decide the input of the model. In this study, the historical occupancy routine and temporal sequential relationship of the occupancy state are mainly discussed. So only the data from PIR sensor is used as the occupancy state data, without introduction of other ambient parameters. As for the

input, the historical occupancy state and temporal information is considered. In conclusion, this study takes the following three variables as inputs, and the prediction is conducted with these inputs separately:

- (1) Historical states, in this study, we select the historical states in the former five minutes.
- (2) Time point.
- (3) Historical occupied ratio in this time point.

Evaluation

Confusion matrix, accuracy and Root-Mean-Square Error (RMSE) are selected as the evaluation indicators (Fawcett, 2006). The confusion matrix refers to

$$\begin{bmatrix} \text{True positive} & \text{False negative} \\ \text{False positive} & \text{True negative} \end{bmatrix} \quad (7)$$

And the ratio of confusion matrix is expressed as:

$$\begin{bmatrix} \text{TPR} & \text{FNR} \\ \text{FPR} & \text{TNR} \end{bmatrix} \quad (8)$$

The accuracy refers to the rate of correct prediction, which is:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (9)$$

The dataset is divided into training set and testing set. The indicators can be used for training set to get the fitting performance, as well as the testing set to analyse the prediction performance. The specific procedure is as Figure 7 shows.

Figure 7: Procedure of model training and testing.

As for the validation, the ten-fold cross validation is conducted to the dataset (Kohavi, 1995).

Case Study and Results

Historical Occupancy Dataset

The occupancy data in one office room is measured lasting for two months, with only weekdays picked out.

The temporal resolution of the data is 30 seconds. The office is for three staff.

The data for the first month is analysed about the historical occupied routine. The occupied ratio is shown in Figure 8. This study focuses on prediction of whether the room is occupied. Here the occupied ratio means the

proportion of the room to be occupied in the history days in each specific time point.

Figure 8: Historical occupied ratio in study office room.

From the analysed occupied ratio, the occupancy state in this office is of certain routine. The working time is from 8 am to nearly 7 pm. The occupied ratio increases and decreases all sharply, which means that the routine of the working is quite regular. Another feature lies at noon, when the occupants regularly go out for lunch. As for the working period, the occupied ratio is not stable but fluctuating, indicating that the occupants are not continuously in the room. Possible events include going to the restroom and having a meeting.

Model Selection and Parameter Tuning

For the model selection and parameter tuning, half of the occupancy data set is selected as training set. Three inputs are tested to the model separately. The optimized parameters are decided by the evaluation results of fitting performance.

The RMSE value for fitting data is used to decide which parameter combination is the best for the model fitting. The optimization analysis is from two aspects:

(1) Activation and solver

Figure 9 illustrates the results of the model selection and parameter tuning with different combinations of activation functions and solver methods. The RMSE value refers to the average value for fitting evaluation results. The input of historical occupancy states shows the best fitting performance, with the lowest RMSE value. The influence of the activation function to the fitting performance is not distinct, while normally, the “lbfgs” solver method is better than “adam” method.

Figure 9: Model selection and parameter tuning results of activation and solver.

(2) Neuron size and L2 penalty parameter (α)

Figure 10 and Figure 11 illustrate the optimization results of neuron size and penalty parameter (α), with one and two hidden neuron layers separately. The input of historical occupancy state is of more satisfied fitting performance. The parameter adjustment did not bring about much performance improvement in terms of the time point input. For this study, the increase of neuron size or the layer size did not improve the fitting performance well. While the penalty parameter (α) much influences the results. With the α value decrease, the model tends to be more complicated, and the RMSE value decreases accordingly. However, the improvement is also limited, when the α is smaller than 1, the RMSE value does not decrease any more. Therefore, the α value should not be definitely small to avoid overfitting.

It should be noted that in order to enhance the visualization of the results of model selection and parameter tuning, and because the dimension of the parameters are more than three dimensions, the RMSE values are calculated as the average value for the parameter items not mentioned in the two coordinates in each plot. The figure tries to present the trend of the model selection and parameter tuning.

Figure 10: Model selection and parameter tuning results of activation and solver (with one hidden neuron layer).

Figure 11: Model selection and parameter tuning results of activation and solver (with two hidden neuron layers).

Prediction Results

After the model selection and parameter tuning, the parameters of the best performance for each input are summarized in Table 3.

In this study, the input of historical states possesses the best fitting performance. The prediction performance is evaluated during the parameter tuning. The average RMSE for prediction is 0.075, and the average accuracy

for prediction is 99.4%. The prediction result for one day is illustrated as Figure 12 and the accuracy is shown in Table 4. The third graph shows the difference between real occupancy data and predictive occupancy data, which shows the time point of wrong prediction. It is noted that the figure refers to the occupancy state of the room. 1 refers to the occupied state and 0 refers to unoccupied state.

Table 3: Summary of optimized parameters for each input.

Input	Historical states	Time point	Occupied ratio
n layer	1	2	2
n neuron	10	20	10
Activation	identity	logistic	identity
Solver	lbfgs	lbfgs	lbfgs
α	1	0.01	1
RMSE for fitting	0.0747	0.1832	0.1826

Table 4: Prediction results of the best performance.

Input	Prediction RMSE	Accuracy
Historical states	0.075	99.4%

Figure 12: Comparison result for occupancy prediction in one day.

Conclusion

In both building design phase and operation stage, the occupancy state is one of key inputs. It is important to figure out the occupancy state in the room or building. From measurement, occupancy state in buildings possesses spatial and temporal stochastic feature. In this study, the temporal sequential feature is analysed and used for occupancy prediction. The machine learning method is introduced to the occupancy prediction. Through model selection and parameter tuning, the

optimized parameters are selected. Furthermore, the evaluation method of the prediction is proposed.

This study shall be the preliminary research of occupancy study only based on PIR sensor to detect the occupancy state of the room. However, the data quality from PIR sensor needs careful analysis. How to distinguish errors in the measurement data and how to maintain the prediction quality are key issues. For future work, the performance for combined input shall be evaluated. The amount and state of occupants in buildings are also to be figured out. What's more, self-learning and model update need to be settled for on-site use.

Acknowledgement

This study was supported by National Natural Science Foundation of China (grant number 51778321): Research on the quantitative description and simulation methodology of occupant behavior in buildings.

References

- DONG, B. (2010). Integrated building heating, cooling and ventilation control.
- DONG, B., YAN, D., LI, Z., JIN, Y., FENG, X. & FONTENOT, H. (2018) Modeling occupancy and behavior for better building design and operation—A critical review. *Building Simulation*. Springer, 1-23.
- FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27, 861-874.
- GARDNER, M. W. & DORLING, S. R. (1998). Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32, 2627-2636.
- Design standard for energy efficiency of public buildings (2005). (GB50189-2005)
- JAIN, A. K., MAO, J. & MOHIUDDIN, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, 29, 31-44.
- KIM, S. H. & MOON, H. J. (2018). Case study of an advanced integrated comfort control algorithm with cooling, ventilation, and humidification systems based on occupancy status. *Building and Environment*, 133, 246-264.
- KINGMA, D. & BA, L. (2015). Adam: A Method for Stochastic Optimization.
- KOHAVER, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, Montreal, Canada, 1137-1145.
- LE, Q. V., NGIAM, J., COATES, A., LAHIRI, A., PROCHNOW, B. & NG, A. Y. (2011). On optimization methods for deep learning. *Proceedings of the 28th International Conference on International Conference on Machine Learning*. Bellevue, Washington, USA: Omnipress.
- LIAO, C., LIN, Y. & BAROOAH, P. (2012). Agent-based and graphical modelling of building occupancy. *Journal of Building Performance Simulation*, 5, 5-25.
- LIU, D. C. & NOCEDAL, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45, 503-528.
- PAGE, J., ROBINSON, D., MOREL, N. & SCARTEZZINI, J.-L. (2008). A generalised stochastic model for the simulation of occupant presence. *Energy and buildings*, 40, 83-98.
- PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R. & DUBOURG, V. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12, 2825-2830.
- RUMMELHART, D. E. (1986). Learning internal representations by error propagation. *Parallel Distributed Processing: I. Foundations*, 318-362.
- SCHALKOFF, R. (1992). Pattern Classification: Statistical, Structural and Neural Approaches. John Wiley and Sons, Inc., New York.
- STANDARD, A. & CHAPTER, R. M. (2004). Ansi/ashrae/iesna standard 90.1-2004. *Energy standard for buildings except low-rise residential buildings*.
- WANG, C., YAN, D. & JIANG, Y. (2011) A novel approach for building occupancy simulation. *Building simulation*. Springer, 149-167.
- WANG, D., FEDERSPIEL, C. C. & RUBINSTEIN, F. (2005). Modeling occupancy in single person offices. *Energy and buildings*, 37, 121-126.