A data mining model for building occupancy estimation based on deep learning methods

Yaping Zhou, Zhun (Jerry) Yu¹,*, Jun Li, Yuanjian Huang¹, Guoqiang Zhang¹

¹ College of Civil Engineering, National Center for International Research Collaboration in Building Safety and Environment, Hunan University, Changsha, Hunan 410082, P.R. China

* zhunyu@hnu.edu.cn

Abstract. Real-time occupancy estimation is of great importance to improve systems control and energy efficiency of buildings. This study proposed a novel real time occupancy estimation model based on CO2 concentration data. A non-neural-network deep learning method (i.e. gcForest) was used to estimate the number of occupants. The gcForest incorporates three classifiers in each level, enabling the estimation performance to be enhanced by exploiting the complementarity among different learning algorithms. To evaluate the effectiveness of the proposed model, this study conducted an experiment in an office room and compared its results with the IHMM model that was widely used in previous studies. The experimental results indicate that the proposed model could achieve higher estimation accuracy and higher detection accuracy of occupant presence or absence.

1. Introduction

Real-time occupancy estimation plays an important role in energy demand prediction as well as operation of systems in buildings [1]. The presence/absence and the number of occupants are especially crucial as they determined the dynamics of various building services systems such as lighting, heating, ventilation and air conditioning [2-4]. In the past years, the occupancy estimation based on CO2 concentration measurements has been studied by many researchers as it is a non-instructive and cost-effective way [5]. Moreover, machine learning approaches were extensively employed for occupancy estimation by automatically mining the relationship between CO2 concentrations and the number of occupants. So far different machine learning models have been developed, mainly including hidden markov models [6], support vector machines models [7], artificial neural networks models [8] and decision trees models [9, 10]. These models have strong non-linear mapping ability. However, their performance was more or less limited by the shallow structure and non-informative inputted features. Therefore, novel occupancy models with proper feature extraction are still needed. This paper aims at developing a deep learning method (i.e. gcForest) for real-time occupancy estimation based on CO2 data. The gcForest incorporates three classifiers in each level, enabling the estimation performance to be enhanced by exploiting the complementarity among different learning algorithms. An experiment in an office room was conducted for the model establishment. To demonstrate its effectiveness, its results were compared with the results of the IHMM model that was widely used.

2. GcForest for real-time occupancy estimation

2.1. GcForest model

As a non-neural-network deep learning method, the gcForest model proposed by Zhou and Feng [11] relies on level-by-level processing of raw CO2 concentration data with less hyper-parameters and automatically identified number of levels. The original implementation of gcForest in python can be found in https://github.com/kingfengji/geForest. In each level, multiple tree-based classifiers are used...
for ensemble learning. In this study, the gcForest model mainly include two major parts for occupancy estimation: multi-grained scanning and cascade forest structure.

2.1.1. Multi-grained scanning. The multi-grained scanning uses pre-defined sliding windows to extract features from input CO$_2$ concentration data. Assume there are $l$ points of CO$_2$ concentration data serving as raw features in multi-grained scanning and sliding window sizes are $d$. A $d$-dimensional feature vector will be generated by sliding scanning at each time step and a total of $(l-d+1)$ feature vectors can be generated. Accordingly, for an $n$-class prediction, one tree-based classifier can produce $n \times (l-d+1)$ dimensional class vectors. For example, if three classifiers were used in gcForest, in total $3 \times n \times (l-d+1)$ dimensional class vectors can be produced which are used as transformed feature vectors to concatenate with the following cascade forest structure. Through such a process, the raw CO$_2$ features can be reconstructed and enhanced by introducing temporal characteristics to improve the performance of the gcForest. Note that the sliding window can be set to different values. In this study, the sliding windows are set to 1, 3, 5 respectively for scanning the CO2 time series data.

2.1.2. Cascade forest structure. The cascade forest structure is a multi-level structure where each level is an ensemble of different tree-based classifiers. Different types of tree-based learning classifiers can be adopted to enhance the diversity of the ensemble methods. In particular, tree-based classifiers in each cascade level receive extracted features generated from its preceding level and generate new outputs (i.e. class vectors). After extending to a new level, the whole model performance can be calculated based on the validation set and the training process will be terminated if the model performance cannot be improved by adding new levels. Therefore, the optimal number of levels can be automatically identified. The overall class distribution is the average of the probability of each class predicted from different classifiers. Based on the predicted class distribution of the last level, the class with the maximum probability is regarded as the final prediction output.

2.2. Learners for ensemble learning

Each level of the gcForest can integrate multiple classifiers. Three algorithms, i.e. XGBoost, Random forests and Extra trees were employed in this study. XGBoost is builds an ensemble of many classification and regression trees [12]. It shows several unique features: (1) inclusion of a penalization term for the model complexity and a randomization parameter to further de-correlate the individual models; (2) a sparsity-aware split finding algorithm to handle different types of sparsity patterns in the data (e.g. missing data); (3) a weighted quantile sketch algorithm to effectively handle weighted data. Random Forests is an ensemble of decision trees combining with the concept of randomized feature selection [13]. Unlike the conventional construction process of decision trees, random forests randomly select a subset of features at each node while splits nodes deterministically by calculating Gini measures on the selected feature subset. The final output is decided by majority votes that generated from multiple decision trees. Extra Trees is an ensemble algorithm of decision/regression trees with two major characteristics [14]: (1) whole training samples are used to grow trees rather than a bootstrap sampling used in RF; (2) randomized feature selection and corresponding nodes are split by randomly choosing cutting-points.

3. Experiment

3.1. Data collection

An experiment was conducted in an office room accommodating four graduate students. A CO$_2$ concentration sensor was installed in the centre of the room and a camera was installed in the entrance. During the experiment, the CO$_2$ concentration data was recorded in the resolution of 1 minute. The actual number of occupants was counted manually and was recorded in the resolution of 10 minutes. The measurement lasted for a month, including 28 days used for training and one day for testing.

3.2. Assessment metrics

To evaluate the model performance, the widely used inhomogeneous hidden markov model (IHMM) was also developed for comparison. Two widely-used metrics were employed to compare the estimation
performance of the two occupancy models, including estimation accuracy (EA) and the detection accuracy of presence/absence (DA). They are explained as follows.

- **EA** is the percentage of correctly estimated number of occupants.

\[
EA = \frac{\sum_{i=1}^{N} \chi(O^* - O_i^e)}{N}
\]

Where:

\[
\chi(O^* - O_i^e) = \begin{cases} 
1, & \text{if } (O^* - O_i^e) = 0 \\
0, & \text{otherwise}
\end{cases}
\]

- **DA** is the percentage of correct binary estimations.

\[
DA = \frac{(A + D)}{(A + B + C + D)}
\]

Where:

- **A**: the sum of true positives if both the estimated state and the actual state are ‘occupied’.
- **B**: the sum of false positives that the estimated state is ‘occupied’ but the actual state is ‘absent’.
- **C**: the sum of false negatives that the estimated state is ‘absent’ but the actual state is ‘occupied’.
- **D**: the sum of true negatives that both estimate state and the actual state are ‘absent’.

4. Results and discussions

Figure 1. Occupancy estimation results with the IHMM model against the ground truth

Figure 2. Occupancy estimation results with the GcForest model against the ground truth

Figure 1 and Figure 2 show the occupancy estimation results generated by the IHMM model and the gcForest model of one day, respectively. As can be seen, both two models can effectively capture the continuous absent state from 0 AM to 8 PM. This may be due to the persistent low CO₂ concentration during these periods. However, the estimation performance of these two models is significantly different in other periods when the office was occupied. It can be observed that the gcForest model tends to have better performance for occupancy estimation than the IHMM model, especially for the first arrival time. As shown in Figure 1 and Figure 2, the actual first arrival time in the test day was 9 AM and it was accurately estimated by the gcForest model, whereas the estimated first arrival time was delayed for 30 minutes for the IHMM model. A possible explanation for this might be that the IHMM model relies on historical occupancy states to build the temporal relationship between the estimated occupancy states,
which may result in difficulties in tracking the occupancy dynamics. Moreover, compared with the IHMM model, the gcForest model can more effectively track the occupancy changes without obvious lags. This can be attributed to the ability of automatically feature extraction and the deep structure that can enhance the modelling ability.

Table 1. EA and DA for the four models

| Criteria | IHMM | GcForest |
|----------|------|----------|
| EA (%)   | 74.3 | 83.3     |
| DA (%)   | 97.9 | 1        |

Table 1 shows the quantified performance of the two models. As can be seen, the gcForest model has a higher EA of nearly 9% than the IHMM model. Moreover, the presence and absence of the office room can be completely detected by the gcForest as DA equals to 1, while errors exist in the IHMM model. Therefore, it can be concluded that the gcForest model have superior performance over the IHMM model for real-time occupancy estimation.

5. Conclusions and future work

This paper presents a novel real time occupancy estimation model based on a non-neural-network deep learning method (i.e. gcForest) using CO2 concentration data. An experiment was conducted in an office room and a comparison with the IHMM model was implemented to evaluate the effectiveness of the proposed model. The results showed that the proposed model achieves superior performance over the IHMM model. In particular, it can effectively track changes in the occupancy dynamics. In the future, it is promising to combine CO2 concentration data with other parameters for occupancy estimation under different scenarios to further improve the occupancy estimation accuracy.

References

[1] Yu Z, Haghighat F, Fung BCM, Yoshino H 2010 *Energy Build*. A decision tree method for building energy demand modeling 42 1637-46
[2] Oldewurtel F, Sturzenegger D, Morari M 2013 *Appl Energy*. Importance of Occupancy Information for Building Climate Control 101 521-32
[3] Yu Z, Haghighat F, Fung BCM, Zhou L 2012 *Energy Build*. A novel methodology for knowledge discovery through mining associations between building operational data 47 430-40
[4] Yu Z, Haghighat F, BCM F, Morofsky E, Yoshino H 2011 *Energy*. A methodology for identifying and improving occupant behavior in residential buildings 36 6596-608
[5] Jiang C, Masood MK, Soh YC, Hua L 2016 *Energy Build*. Indoor occupancy estimation from carbon dioxide concentration 131 132-41
[6] Chaney J, Owens EH, Peacock AD 2016 *Energy Build* An evidence based approach to determining residential occupancy and its role in demand response management 125 254-66
[7] Rana R, Kusy B, Wall J, Wen H 2015 *Energy*. Novel activity classification and occupancy estimation methods for intelligent HVAC (heating, ventilation and air conditioning) systems 93 245-55
[8] Ekwevugbe T, Brown N, Pakka V, Fan D 2013 *IEEE International Conference on DEST*. Real-time building occupancy sensing using neural-network based sensor network
[9] Hailemariam E, Goldstein R, Attar R, Khan A 2014 *Symposium on Spring Simulation Multi-conference*. Real-time occupancy detection using decision trees with multiple sensor types
[10] Dong B, Andrews B, Lam KP, Höynck M, Zhang R, Chiou YS 2010 *Energy Build*. An information technology enabled sustainability test-bed (ITEST) for occupancy detection through an environmental sensing network 42 1038-46
[11] Zhou ZH, Feng J 2017. Deep Forest: Towards An Alternative to Deep Neural Networks
[12] Chen T, Guerstin C 2016 *Acm Sigkdd International Conference on KDD*. XGBoost: A Scalable Tree Boosting System 785-94
[13] Breiman L 2001 *Mach Learn*. Random Forests 45 5-32
[14] Geurts P, Ernst D, Wehenkel L 2006 *Mach Learn*. Extremely randomized trees 63 3-42