Postprint: Space Occupancy Detection Based on Gradient Boosting Decision Models

Authors: Xu Xinwei, Ding Jing' an, Liu Zhicai, Wang Duomei, soaring, Shao Ruirui

Date: 2018-05-18T00:00:00+00:00

Abstract

As economic incentive mechanisms for green buildings and green eco-districts have been basically established, the "Big Data Green Building" energy-saving system has emerged to address large-scale multi-dimensional spatial occupancy data. Nevertheless, extensive multi-dimensional building data have not been fully exploited, and traditional spatial occupancy detection models exhibit insufficient classification accuracy along with high time complexity. Leveraging the UCI occupancy detection dataset, the addition of timestamps to the original dataset enhances the classification accuracy of models. Concurrently, the MCMR (Maximum Correlation Minimum Redundancy) method is utilized for feature selection, with Random Forest employed as a classifier to validate classification effectiveness and acquire the optimal feature subset. Moreover, occupancy detection models are constructed using the selected feature subset, with the XGBoost model compared against the Random Forest model (RF), demonstrating higher classification accuracy and lower time complexity.

Full Text

Preamble

Occupancy Detection Based on Extreme Gradient Boosting Decision Model

Xu Xinwei[1,2], Ding Jing' an[1], Liu Zhicai[1], Wang Duomei[3], Teng Xiang[1], Shao Ruirui[1]

[1]School of Management Science & Engineering, Anhui University of Technology, Ma' anshan, Anhui 243000, China
[2]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210000, China
[3]School of Public Administration, Hohai University, Nanjing 210000, China

Abstract: With the gradual formation of green buildings and green-economic environmental cities, "big data green building" energy conservation systems have emerged. However, large amounts of multi-dimensional building data remain underutilized, while traditional occupancy detection algorithms suffer from insufficient accuracy and high time complexity. This study utilizes the Occupancy Detection dataset from UCI, adding timestamps to the original dataset to improve model classification accuracy. The MCMR (Maximum Correlation and Minimum Redundancy) method is employed for feature selection, with random forest serving as the classifier to verify classification effectiveness and obtain the optimal feature subset. Using the selected feature subset to construct occupancy detection models, the XGBoost model demonstrates higher classification accuracy and lower time complexity compared to the random forest (RF) model.

Keywords: big data of green buildings; occupancy detection; MCMR; XGBoost

Introduction

On March 1, 2017, the "13th Five-Year Plan for Housing and Urban-Rural Development" proposed "developing green buildings and green building materials, and vigorously strengthening building energy efficiency," explicitly requiring that "by 2020, the proportion of green buildings in newly constructed urban buildings should exceed 50%, green building materials application should exceed 40%, and the energy efficiency standard for new buildings should improve by 20% compared to the end of the 12th Five-Year Plan period." On March 21, the 13th International Conference on Green Building and Building Energy Efficiency, themed "Improving Green Building Quality, Promoting Energy Saving and Low-Carbon Development," was held at the Beijing National Convention Center. The conference discussions covered national policies, economic situations, product improvements, and operational management, while technological innovation, platform construction, and big data analysis became hot topics in green building and energy efficiency. With China' s large population and limited resources, energy, economic, and environmental issues have become urban hotspots. How to comprehensively utilize energy rationally, improve urban energy efficiency, optimize resource allocation, and protect the natural environment has become a focus of societal concern. Meanwhile, with the advent of the "big data" era, the combination of green building development concepts and technologies such as data mining and machine learning has achieved certain progress. Through feature extraction from large-scale multi-dimensional heterogeneous building occupancy data to find optimal feature subsets for model construction, occupancy detection can improve classification accuracy, representing a new method for optimizing the allocation of building-related energy resources.

Occupancy detection essentially belongs to the category of pattern recognition,

utilizing multi-sensor monitoring of indoor environments to obtain spatial occupancy data. Through feature extraction and classification algorithm selection, classification models are constructed to predict occupancy status. The predicted accuracy is then used for intelligent HVAC system control to achieve energy savings. Research indicates that intelligent HVAC control through occupancy detection technology can theoretically save approximately 29% to 80% of energy annually. Introducing temporal features through cameras and sensors to obtain real-time data as input for extended Kalman filter algorithms can improve occupancy detection accuracy. Real-time occupancy detection using RFID technology and motion sensors can reduce natural gas consumption and the time when spaces are occupied but rooms remain unwarmed, thereby improving spatial comfort. Different feature combinations and algorithm models also affect occupancy detection accuracy. Extracting existing environmental resources from commercial buildings, including access permits, wireless records, schedules, and communication clients, and using linear regression and C4.5 algorithms for occupancy detection can achieve approximately 90% accuracy. Decision trees applied to single-sensor data for occupancy detection achieve 97.9% accuracy, while multi-sensor data feature combinations can reach 98.4%. However, when sound and CO2 sensor data are added, prediction results become less ideal, suggesting that decision trees can improve single-sensor detection accuracy but may experience precision degradation with excessive sensor data feature combinations. When combining light, temperature, humidity, and CO2 sensor data features for occupancy detection, random forest exhibits overfitting when predicting with all feature combinations, while linear discriminant analysis can achieve 97% accuracy with only two features, indicating that different feature combinations affect prediction accuracy. Through the above literature analysis, first, introducing timestamps can improve occupancy detection accuracy; second, multi-sensor feature combinations lead to classification model accuracy degradation. The main reasons are high correlation between features or redundant features, with some features containing minimal category information, resulting in low classification effectiveness and affecting model performance and time complexity.

Addressing the issues of classification accuracy degradation with multi-sensor feature combinations and accuracy improvement through timestamp introduction, this paper proposes an MCMR (Maximum Correlation and Minimum Redundancy) feature selection algorithm. Timestamps are extracted from the original UCI occupancy detection dataset to refine temporal granularity and form a new dataset. The MCMR method is applied for feature selection on this new dataset, with the selected optimal feature subset serving as input for the XGBoost (extreme gradient boosting) algorithm. XGBoost is an ensemble learning algorithm under the gradient boosting framework, featuring a flexible and portable distributed gradient boosting library that maintains relatively high classification accuracy with lower time complexity when processing large datasets. Few studies have employed this method for detailed research on spatial occupancy information extraction. This paper combines the advantages of

MCMR feature selection with XGBoost's distributed parallel computing, achieving improved accuracy for automatic spatial occupancy recognition models from a feature combination perspective while reducing model time complexity.

The main contributions of this paper are as follows: timestamps are added to the original data, addressing XGBoost and RF algorithms' inability to process temporal variables directly. Experimental results show that models with timestamps achieve improved classification accuracy compared to those without timestamps. The most significant improvement is XGBoost' s classification accuracy on the testing set, which increased by 4.09%, while RF' s accuracy on the testing dataset improved by 2.78%. Additionally, RF_1 demonstrates more reasonable timestamp introduction than literature [19]. The MCMR feature selection method is used to remove the HumidityRatio feature with low correlation and high redundancy. Using random forest as the classifier for iterative optimization, the optimal feature subset is obtained. Through feature and classification algorithm combination, XGBoost achieves the highest classification accuracy on the training dataset at 99.41%; SVM achieves the highest accuracy on test dataset 1 at 97.90%; and BP achieves the highest accuracy on test dataset 2 at 99.07%. Finally, comparing XGBoost with the random forest (RF) classification method, the XGBoost model demonstrates higher classification accuracy and lower algorithm time complexity.

---

1 Theory and Methods

1.1 MCMR Feature Selection Method

Features with low correlation and high redundancy affect classification model accuracy, necessitating feature selection for sample datasets. The purpose of feature selection is to choose a smaller-scale feature subset from the sample data collection that can provide similar or better performance in data mining and machine learning tasks compared to the original set. With fewer features, data becomes more interpretable without changing the amount of category information contained in the features. Traditional feature selection methods only consider linear or nonlinear correlation between features without considering full correlation, and often separate relevance and redundancy judgments, making it impossible to evaluate the combined effect of the entire feature subset.

Based on linear and nonlinear correlation, this paper calculates the full correlation coefficient between features to measure independence and redundancy. Simultaneously, based on information theory, mutual information between features and categories is calculated to represent the amount of category information contained in features, indicating the correlation degree between features and categories. Combining the advantages of wrapper and filter feature selection methods, this paper proposes an MCMR (Maximum Correlation and Minimum Redundancy) feature selection algorithm.

Assuming sample sets x and y represent paired continuous variables of length n, the linear correlation coefficient r between features is calculated through Pearson correlation as shown in Equation (1):

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

The correlation matrix A is constructed. Using distance correlation, the nonlinear correlation coefficient R_n^* between features is calculated to construct correlation matrix B, where $dcov(x,y)$ is the distance covariance of variables x and y, and $dvar(x)$ and $dvar(y)$ are the distance standard deviations of variables x and y, respectively, as shown in Equation (2):

$$R^* = \frac{dcov(x,y)}{\sqrt{dvar(x) \cdot dvar(y)}} = dcor(x,y)$$

Integrating linear and nonlinear correlation, the full correlation between features is calculated to obtain correlation matrix C. The calculation process for full correlation is as follows: the full correlation coefficient is computed as shown in Equation (3), where i and j represent the i-th row and j-th column of the correlation matrix, resulting in the full correlation matrix C.

Based on information theory, mutual information between features and categories is calculated. Let $P(x_i)$ denote the probability of feature x taking the i-th value x_i, and $P(x_i|y_j)$ denote the probability of feature x taking value x_i when category y takes value y_j. The information entropy $H(x)$ of x and the conditional information entropy $H(x|y)$ of x given y are calculated as follows:

$$H(x) = -\sum p(x) \log p(x)$$

$$H(x|y) = -\sum_j p(y_j) \sum_i p(x_i|y_j) \log p(x_i|y_j)$$

The mutual information $MI(x,y)$ between variables x and y is calculated as:

$$MI(x,y) = H(x) - H(x|y) = H(y) - H(y|x) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

However, since mutual information has preference issues, this paper uses the mutual information rate to measure the correlation between feature x and category y:

$$sim(x,y) = \frac{MI(x,y)}{H(x)}$$

The resulting correlation degree $sim(x,y)$ ranges between [0,1], where 0 indicates no correlation and 1 indicates complete correlation.

The MCMR feature selection algorithm is described as follows:

Input: Training set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_m, y_m)\}$

a) **Discretization and Initialization:** Use the ChiMerge method to discretize continuous features in dataset D, with the result still denoted as D.

b) **Calculate Correlations:** According to Equations (1)-(2), calculate the full correlation coefficient w between any two features in dataset D; according to Equations (4)-(7), calculate the correlation coefficient sim between features and categories.

c) **Set Parameter:** Set parameter $\alpha \in [0, 1]$, identify features with $w > \alpha$ correlation, and compare sim values.

d) **Feature Removal:** Remove features with smaller sim values and verify 合理性 through random forest classifier accuracy; otherwise, update $\alpha = \alpha+1$ and return to step 3.

e) **End For**

Output: Feature subset

1.2 Gradient Boosting Algorithm Classification Mechanism

XGBoost (extreme gradient boosting) is an ensemble learning algorithm based on the GBDT (gradient boosting decision tree) framework. GBDT combines "gradient descent"with decision trees, constructing new classifiers in the direction of reducing residuals from the previous model through iterative construction of a series of weak classifiers, whose weighted cumulative output serves as the strong classifier output. The difference between XGBoost and GBDT lies in XGBoost's change from GBDT' s serial sequence solving approach to CPU multi-threaded distributed parallel computing, and its use of Taylor quadratic expansion for residual solving, thereby breaking existing computational speed and accuracy limitations.

The basic steps for training the GBDT classification algorithm are as follows:

Input: Training set $\{(x_i, y_i)\}_{i=1}^n$

a) **Initialize Model:** $F_0(x) = \arg\min_\rho \sum_{i=1}^n L(y_i, \rho)$

b) **For** $m = 1, 2, ..., M$ **do:**

c) **Calculate Negative Gradient:** $g_m(x) = \left[\frac{\partial L(y,F(x))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}$

d) **Update Model:** $F_m(x) = F_{m-1}(x) + \rho_m g_m(x)$

e) **End For**

In the above steps, $F^*(x)$ seeks the decision function that minimizes expected loss, $L(y, F(x))$ is the loss function, $g_m(x)$ is the negative gradient direction of the current model, and m calculates the negative gradient value of the loss function in the current model as a residual estimate. Regression tree leaf node regions are estimated to fit approximate residual values, linear search is used to estimate leaf node region values to minimize the loss function, and regression trees are updated to obtain the final output model $F_m(x)$.

XGBoost performs second-order Taylor expansion on the loss function $obj(t)$ and adds a regularization term $\Omega(f_t)$ outside the objective function to find the optimal solution, balancing objective function reduction and model complexity to avoid overfitting. Equation (8) shows the Taylor expansion of the objective function with the introduction of the regularization term:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

$$\approx \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

First-order derivative g_i and second-order derivative h_i are solved for each sample. The objective function is grouped by leaf nodes:

$$obj^{(t)} \approx \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda) w_j^2] + \gamma T$$

This paper adds timestamps to the spatial occupancy sample dataset, uses the MCMR feature selection method to remove features with low correlation and high redundancy, selects the optimal feature subset, and constructs a spatial occupancy detection model using the optimal feature subset and XGBoost classification algorithm. The flowchart is shown in Figure 1 [Figure 1: see original paper], which mainly includes the following steps:

a) **Add Timestamps:** Re-extract date variables from the original dataset to increase classification features, enabling real-time spatial occupancy detection and constructing sample datasets.

b) **Feature Selection Using MCMR:** Calculate full correlation coefficients between features and mutual information rates between features and categories to select features with high correlation and low redundancy.

c) **Feature Subset Selection and Validation:** Use wrapper-style random forest feature recursive elimination to verify the 合理性 of the MCMR feature selection method and obtain the optimal feature subset.

d) **Classifier Construction and Training:** Input the selected feature subset as training samples to construct XGboost classifiers through iterative building of regression decision trees.

e) **Intelligent HVAC System Control:** Through classifier model learning, obtain models with high classification accuracy. Based on the learned logic, analyze existing indoor environmental variables, predict occupancy status, and intelligently adjust HVAC systems to achieve energy savings.

---

3 Experiments and Analysis

3.1 Experimental Data

Spatial occupancy detection influencing factors often present detection difficulties, making it challenging to obtain large-scale ultra-high-dimensional data. Therefore, only relatively easy-to-monitor data can be obtained. This paper' s data is sourced from the Occupancy Detection dataset on UCI. Both the training set and testing set were measured with doors closed, while the testing2 set was measured with doors open. The dataset includes variables: Date, Temperature (T), Humidity (H), Light, CO2 concentration (CO2), Humidity Ratio (HR), and Occupancy status. The training set contains 8,143 records, the testing set contains 2,665 records, and the testing2 set contains 9,752 records. The training data format is shown in Table 1 .

3.2 Evaluation Metrics

For the spatial occupancy detection model, this paper uses the commonly used confusion matrix for decision models as the performance evaluation metric to measure performance on training and test sample sets. A 2×2 confusion matrix is used to represent prediction accuracy:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP, TN, FP, and FN refer to the number of times model predictions fall into these categories. Therefore, accuracy represents the number of correctly classified instances divided by the total number of predictions.

3.3 Experiments and Analysis

The experimental environment is as follows: operating system Windows 7, CPU Intel® Core™ i5-3210M @2.5GHz, experimental memory 4GB, main experimental platform R version 3.3.3.

1) Data Preprocessing

(1) Timestamp Introduction

Since gradient boosting models and random forest models cannot directly process temporal variables, this paper processes the collected data as follows: use the lubridate package to re-extract temporal variables. In Equation (10), x represents the date in the original data.

Literature [19] processes temporal variables as shown in Equations (11) and (12), where NSM is the total seconds converted from hour, minute, and second in the Date sample, and weekstatus represents weekday status (0 for weekends, 1 for weekdays).

The advantage of timestamp introduction in this paper compared to literature [19] lies in refined temporal granularity. Although literature [19]' s timestamp introduction is easier to interpret, it introduces two temporal features, while this paper uses only one feature without losing temporal information. In literature [19], the NSM variable has large values without eliminating magnitude, affecting model training weights and potentially being excluded during feature selection based on importance ranking. The timestamp introduced in this paper, after processing, yields training sample data as shown in Table 2 .

(2) MCMR-Based Feature Selection Method

This paper proposes a feature selection method combining feature redundancy and feature-category correlation. The main calculation process is as follows: use Pearson coefficient to calculate the linear correlation matrix A for the sample dataset, use distance correlation coefficient to calculate the nonlinear correlation coefficient matrix B, and use Equation (3) to obtain the full correlation coefficient matrix C. Correlation coefficients range in [0,1], where (0.8,1] indicates extremely strong correlation, (0.6,0.8] strong correlation, (0.4,0.6] moderate correlation, (0.2,0.4] weak correlation, and [0,0.2] extremely weak or no correlation. Thus, correlation coefficients measure redundancy between features—the larger the coefficient, the higher the redundancy. Simultaneously, based on information theory, the mutual information rate between features is calculated to measure correlation between features and categories, where higher coefficients indicate stronger feature-category correlation.

The full correlation matrix C obtained through calculation is shown in Table 3 . From the table, the full correlation coefficient between HumidityRatio and Humidity is 0.96, indicating extremely strong correlation and very high redundancy. The correlation coefficient between CO2 and Light is 0.596, indicating moderate correlation, while correlations between other features are relatively small with low redundancy.

To clearly measure the amount of classification information contained in each feature, mutual information is calculated. However, numerical continuous feature variables cannot directly calculate mutual information, so this paper uses the ChiMerge algorithm for numerical feature discretization. ChiMerge is the most

commonly used chi-square-based discretization method, a supervised, bottom-up data discretization technique. It first lists all values within the data range as individual intervals, then recursively finds the best adjacent mergeable intervals to form larger intervals. It uses chi-square statistics to detect correlation between adjacent intervals to determine the best mergeable intervals. Time is discretized into 47 categories, Temperature (T) into 67 categories, Humidity (H) into 274 categories, Light into 56 categories, CO2 concentration into 239 categories, and Humidity Ratio (HR) into 718 categories. The discretized training sample dataset is shown in Table 4 .

Using Equation (3) on the discretized features, the correlation coefficient w between features and categories is calculated, yielding the results shown in Table 5 . From Table 5, the variable containing the largest category information ratio is Time, while the smallest is Temperature. Sorting the correlation coefficients yields: Time > Light > Humidity > HumidityRatio > CO2 > Temperature.

Using random forest as the classifier, the optimal feature subset is found through iteration. Initialize $= 0.4$, incrementing by 0.1 each time. The traversal results are shown in Table 6 . From the table, when the number of variables is 5, the selected feature subset is optimal: Light, CO2, Humidity, Temperature, Time. The deletion of the HumidityRatio feature is thus validated as reasonable.

2) Experimental Results and Analysis

Using the optimal feature subset selected in the data preprocessing stage and the introduced timestamp, classification models are constructed. The experimental results are shown in Table 7 .

(1) Impact of Timestamp on Classification Accuracy

Comparing feature combinations with and without timestamps, as shown in Table 7, XGBoost_1, RF_1, and literature [19] all add timestamps without feature selection. Compared with models without feature selection and without timestamp introduction, XGBoost_1 improves classification accuracy on the testing dataset by 4.09% compared to XGBoost_2, while RF_1 improves by 2.78% compared to RF_2. Models with timestamps show overall improved classification accuracy. Furthermore, RF_1 achieves higher classification accuracy on each dataset than literature [19]' s temporal processing, demonstrating that the timestamp introduction method in this paper is more reasonable.

(2) Experimental Results Comparison

Using the optimal feature subset obtained through the above steps, optimal classification models XGBoost_3, RF_3, C50, SVM, BP, etc. are found by adjusting parameters, as shown in Table 7. XGBoost_3 achieves the highest classification accuracy on the training sample dataset at 99.41%; SVM achieves the highest accuracy on test sample dataset 1 at 97.90%; and BP achieves the highest accuracy on test sample dataset 2 at 99.07%.

The XGBoost_3 model achieves 99.41% accuracy on the training set, 97.75%

on the testing sample, and 97.52% on the testing2 sample set. The RF_3 model achieves 99.38% on the training set, 97.67% on the testing sample, and 97.36% on the testing2 sample set. As shown in Figure 2 [Figure 2: see original paper], the XGBoost_3 model' s classification accuracy on the optimal subset is higher than RF_3' s, with the largest gap on the testing dataset.

The above evaluation uses classification accuracy to assess model performance. Time complexity is further evaluated using time complexity functions to quantitatively describe algorithm model running time. For large-scale data processing, this paper seeks both accurate classification and relatively short data processing time. Using parameter values obtained through caret optimization as original algorithm parameters, the classification models are reconstructed using original algorithm packages. XGBoost and RF model time complexities are compared using the system.time() function, yielding the time complexity shown in Table 8 . The table shows that the XGBoost model takes less time than traditional random forest, primarily because the XGBoost algorithm adopts distributed design, thereby reducing time complexity.

---

4 Conclusion

This paper' s main work focuses on spatial occupancy detection research. Based on the original data, timestamps are added to address XGBoost and RF algorithms' inability to process temporal variables directly. Experimental results show that models with timestamps achieve improved classification accuracy compared to those without timestamps. The most significant improvement is XGBoost' s classification accuracy on the testing set, which increased by 4.09%, while RF' s accuracy on the testing dataset improved by 2.78%. Additionally, RF_1 demonstrates more reasonable timestamp introduction than literature [19]. The MCMR feature selection method is used to remove the HumidityRatio feature with low correlation and high redundancy. Using random forest as the classifier for iterative optimization, the optimal feature subset is obtained. Through feature and classification algorithm combination, XGBoost achieves the highest classification accuracy on the training sample dataset at 99.41%; SVM achieves the highest accuracy on test sample dataset 1 at 97.90%; and BP achieves the highest accuracy on test sample dataset 2 at 99.07%. Comparing XGBoost and RF classification models, XGBoost demonstrates higher classification accuracy and lower algorithm time complexity.

Future work will further investigate spatial occupancy detection influencing factors, seeking alternative soft sensors based on existing hard sensors to obtain occupancy impact variables and reduce resource waste. Simultaneously, by observing influencing factor data, the number of occupants will be predicted and resource utilization standards will be set for rational space allocation, ensuring adequate resource utilization.

References

[1] Qiu Baoxing. Situation and tasks of green building development and building energy efficiency in China [J]. Urban Development Studies, 2012, 19(05): 1-7, 11.
[2] Guo Ping, Li Guogang. Analysis of current problems and countermeasures in China' s green building development [J]. Civil Engineering and Environmental Engineering, 2015, 37(S1): 96-98.
[3] Xu Tao, Wang Qi. Sensor fault diagnosis based on pattern recognition [J]. Control and Decision, 2007, (07): 783-786.
[4] Jin Lianwen, Zhong Zhuoyao, Yang Zhao, et al. A review of deep learning applications in handwritten Chinese character recognition [J]. Acta Automatica Sinica, 2016, 42(8): 1125-1141.
[5] Yang Sai, Zhao Chunxia, Liu Fan. Multi-kernel learning fusion of local and global features for face recognition algorithm [J]. Chinese Journal of Electronics, 2016, 44(10): 2344-2350.
[6] Erickson V L, Carreira-Perpiñán M Á, Cerpa A E. OBSERVE: occupancy-based system for efficient reduction of HVAC energy [C]// Proc of the 10th International Conference on Information Processing in Sensor Networks. 2011: 258–269.
[7] Erickson V L, M. Á. Carreira-Perpiñán, A. E. Cerpa, Occupancy modeling and prediction for building energy management, ACM Trans. Sensor Network, 2014, 10(3): 42.
[8] Dong B, Andrews B. Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings [C]// Proc of Building Simulation. 2009.
[9] Brooks J, Goyal S, Subramany R, et al. An experimental investigation of occupancy-based energy-efficient control of commercial building indoor climate [C]// Proc of the 53rd IEEE Annual Conference on, IEEE, Decision and Control. 2014: 5680–5685.
[10] Brooks J, Kumar S, Goyal S, et al. Energy-efficient control of under-actuated HVAC zones in commercial buildings [J]. Energy Build, 2015, 93() 160–.
[11] Tomastik R, Narayanan S, Banaszuk A, et al. Model-based real-time estimation of building occupancy during emergency egress pedestrian and evacuation dynamics [C]. Berlin: Springer, 2010: 215–224.
[12] Scott J, Brush A B, Krumm J, et al. PreHeat: controlling home heating using occupancy prediction [C]// Proc of the 13th International Conference on Ubiquitous Computing. 2011: 281–290.
[13] Ghai S K, Thanayankizil L V, Seetharam D P, et al. Chakraborty: occupancy detection in commercial buildings using opportunistic context sources [C]// Proc of IEEE Percom Workshops. 2012.
[14] Hailemariam E, R. Goldstein, R. Attar, A. Khan: Real-time occupancy detection using decision trees with multiple sensor types, in: Proceedings of the 2011Symposium on Simulation for Architecture and Urban Design, Society for Computer Simulation International, San Diego, CA, 2011, pp.
[15] Luis M. Candanedo, Véronique Feldheim. : Accurate occupancy detection

of an office room from light, temperature, humidity and CO2 measurements using statistical learning models [J]. Energy and Buildings, 2016, 112(1):
[16] Chen T, He T. Higgs boson discovery with boosted trees [C]// Proc of International Conference on High-Energy Physics and Machine Learning.
[17] Chen T, Guestrin C. Xgboost: a scalable tree boosting system [J]. arXiv preprint arXiv: 1603. 02754, 2016.
[18] Song R, Chen S, Deng B, et al. eXtreme gradient boosting for identifying individual users across different digital devices [C]// Proc of International Conference on Web-Age Information Management. [S. l.]: Springer International Publishing, 2016: 43-54.
[19] Yao Xu, Wang Xiaodan, Zhang Yuxi, et al. Survey on feature selection methods [J]. Control and Decision, 2012, 27(02): 161-166+192.
[20] Mao Yong, Zhou Xiaobo, Xia Zheng, et al. Survey on feature selection algorithms [J]. Pattern Recognition and Artificial Intelligence, 2007, 20(2): 211-218.
[21] Qiu Like, Guo Zhongwen, Liu Qing, et al. Feature selection algorithm based on redundancy analysis [J]. Journal of Beijing University of Posts and Telecommunications, 2017, 40(01): 36-41.
[22] Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy [J]. Journal of Machine Learning Research, 2004: 1205-1224.
[23] Li Yang, Gu Xueping. Transient stability assessment feature selection based on improved maximum relevance and minimum redundancy criterion [J]. Proceedings of the CSEE, 2013, 33(34): 179-186.
[24] Zhao Weiwei, Li Yanying, Zhao Fengqin, et al. Hybrid variable selection algorithm based on mutual information and random forest [J]. Journal of Jilin University: Science Edition, 2017, 55(04): 933-939.
[25] Xu Junling, Zhou Yuming, et al. Unsupervised feature selection based on mutual information [J]. Journal of Computer Research and Development, 2012, 49(2): 372-382.
[26] Zhang Zhenhai, Li Shining, et al. A multi-label feature selection algorithm based on information entropy [J]. Journal of Computer Research and Development, 2013, 50(6): 1177-1184.
[27] Dong Hongbin, Teng Xuyang, Yang Xue. A feature selection method based on association information entropy measurement [J]. Journal of Computer Research and Development, 2016, 53(8): 1684-1695.
[28] Xue Wei. Statistical Analysis and Data Mining Based on R [M]. Beijing: China Renmin University Press, 2014, 1-399.
[29] Zhou Zhihua. Machine Learning [M]. Beijing: Tsinghua University Press, 2016, 1-425.
[30] Zhao Jingxian, Ni Chunpeng, Zhan Yuanrui, Du Ziping. An efficient continuous attribute discretization algorithm [J]. Systems Engineering and Electronics, 2009, 31(01): 195-199.
[31] Yang Ping, Yang Tianshe, Du Xiaoning, et al. A discretization algorithm based on maximizing correlation between categorical attributes [J]. Control and Decision, 2011, 26(04): 592-596.

Note: Figure translations are in progress. See original paper for figures.