A Time-aware Method for Occupancy Detection in a Building

Song Ling [1], Niu Xiao fei [1*], Lyu Qiang [2], Lyu Shun ming [3], Tian Tian [1]
{ song_ling@sdjzu.edu.cu, niuxiaofei2002@163.com, 13356667718@189.cn }
School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China [1],
State Grid of China Technology College，Jinan 250002, China [2], School of Information Technology & Electrical Engineering, University of Queensland, Queensland, 4072,Australia [3]

Abstract. The target of buildings' energy efficient is to facilitate a comfortable environment for occupants while maintaining minimal energy consumption. Occupant behaviors pay a large impact in influencing the energy consumption. Time-aware occupancy detection could give information support for intelligent building energy management. In this paper several building occupancy detection methods, which are based on the temporal analysis of historical data, are proposed for providing different size of prediction window occupancy detection. Each proposed approaches are evaluated against accurate real-life data collected from a building. Experiments have been conducted using actual occupancy data under six different time horizons can be used to perform time-aware occupancy states prediction. The experimental results show that Stochastic Gradient Descent (SGD) and Gaussian mixture models-Hidden Markov Model (GMM-HMM) outperforms the other methods for the evaluation. With proposed more accurate time-aware occupancy prediction algorithms, we hope to develop more energy-efficient HVAC(Heating, Ventilation, and Air Conditioning) scheduling systems in order to save energy consumption.

Keywords: Occupancy Detection; Building Consumption; Time-aware Method.

1 Introduction

Buildings consume approximately 40% of the world's total primary usages to provide a comfortable and healthy indoor environment for occupants. And that building energy consumption has continued to increase. For example, in China, Building energy consumption increased by more than 10% annually. Buildings' energy consumption contributes to more than 70% of the electricity energy. HVAC system, plug loads and lighting loads consume majority of the electricity within buildings.

Occupants' behaviors such as occupancy presence and absence within buildings play a significant role to affect energy consumption. Intelligent controlling the operation time of HVAC systems, lighting control systems and the other appliances in buildings by time-aware occupancy detection will reducing the energy consumption effectively while maintaining thermal comfort.

An energy consumption analysis based on commercial buildings in South Africa showed that more energy is used during non-working hours (56%) than during working hours (44%) in Masoso and Grobler's study[1]. The simplest reducing energy waste suggestions for building occupants is to learn to switch off when they leave the buildings.So building occupancy

This work is supported by the National Natural Science Foundation of China (61503220); Shandong province science and technology plan of Housing and Urban-Rural Development. The corresponding author is Niu Xiao fei.

MOBIMEDIA 2019, June 29-30, Weihai, People's Republic of China

DOI 10.4108/eai.29-6-2019.2282388

presence and absence detection could give data support for intelligent building energy contorl. In smart cities, different sensors are used to record city data. Environmental sensor readings including CO_2, temperature, humidity, light and pressure can be a good indicator for building occupancy detection.

In the past studies, two types of occupancy behavior prediction are widely used:

The first one is occupants behaviour patterns. Dong et al. predict user occupancy presence and absence behavior pattern by Hidden Markov Models (HMM) and connect the behavioral patterns to building energy and comfort management systems through simulation tools. The results suggest potential energy savings of 30% while maintaining an indoor comfort level when compared with other basic energy savings HVAC control strategies [2]. Sangogboye et al. proposed multi-label classification to predict occupancy status in commercial buildings. Their experimental results show that prediction performance for commercial buildings depends more on occupancy frequency[3]. Ortega et al. recognized occupancy and activities of daily living pattern by multiclass Support Vector Machines (SVM) to solve the complex characteristics of the data collected from various sensors [4]. Chaney et al. predicted daytime occupancy behavior by HMM based on electrical power, CO_2 levels, and room dew point, showing that they effectively handled periods of missing sensor data[5]. Peng et al. applied k-Nearest Neighbor (KNN) to predict occupants'stochastic behavior and presented a demand-driven control strategy for reducing energy consumption and maintains room temperature for occupants [6].

The second one is predict the number of occupants at some buildings. With Classification and Regression Tree (CART) and HMM, Ryu et al. proposed a predictor to account for occupancy detection at the current state and occupancy prediction at the future state, respectively[7]. Chen et al. attempt to investigate inhomogeneous Markov chain and multivariate Gaussian to solve commercial building occupancy detection problem, and use autoregressive integrated moving average, artificial neural network and support vector regression to predict the number of occupants [8]. After acquiring a typical week's occupancy patterns from anonymous occupancy data for a monitoring period of four months, Capozzoli et al. investigated an occupancy-based HVAC system operation schedule to enhance the energy management in buildings [9].

From above reviews, we think accurate time-aware occupancy detection that provide short-term prediction demonstrate a guideline to a excellent demand-driven control strategy for reducing energy consumption while a comfortable environment needs.Time-aware occupancy detection not only could give information support for energy efficient operation and planning in time but also could contribute to the reliable energy management of smart grids. For example, if an occupant will get to his room (including commercial or resident room) in half an hour, he hopes his building has adjusted to his personal favorite environment. In a similar way, when the time he leaves the room is greater than 20 minutes, the HVAC systems and lighting control systems should be turn off. In this paper, we focus on how to choose a proper method for the prediction of occupancy status better. For the purpose of time-aware occupancy detection, we conduct occupancy detection under six different prediction time granularities, that is, real time (sample frequency), 10 minutes, 20 minutes, 30 minutes, 45minutes and 1 hour.

2 Methodology

In a classification problem, we have a training sample of n observations on a class variable Y that takes values 1, 2, ... , k, and p predictor variables, $X=x_1,..., x_p$. Our goal is to find a model for predicting the values of Y from new X. In theory, the solution is simply a partition of the X space into k disjoint sets, $A_1, A_2,..., A_k$, such that the predicted value of Y is j if X belongs to A_j , for j = 1, 2,..., k.

The time-aware occupancy detection problem can be defined as following: over time period T, where t_i is a given time interval, $Y(t_i)$ is used to determine the occupancy state(occupied or unoccupied) according to X of a building at time interval t_i. If the building is occupied at t_i, then $Y(t_i)$ = 1, otherwise $Y(t_i)$ = 0. The occupancy detection problem can be solved by classification problem.

2.1 Occupancy modeling

2.1.1 Linear Regression(LR)

In LR, Let Y denote the dependent variable whose values we wish to predict, and let $x_1, \ldots, x_p$ denote the independent variables from which we wish to predict it, with the value of variable x_i in period t denoted by x_{it}. Then the equation for computing the predicted value of Y_t is:

$$\widehat{Y}_t = b_1 x_{1t} + b_2 x_{2t} + \cdots + b_p x_{pt} + b_0 \tag{1}$$

The slopes of their individual straight-line relationships with Y are the constants b_1, b_2, ..., b_p. The additional constant b_0, the *intercept*, is the prediction that the model would make if all the X's were zero. The coefficients and intercept are estimated by least squares, i.e., setting them equal to the unique values that minimize the sum of squared errors within the sample of data to which the model is fitted. And the model's prediction errors are typically assumed to be independently and identically normally distributed.

2.1.2 K-Nearest Neighbor (KNN)

KNN is a method for classifying objects based on the closest training examples in the feature space: an object is classified by a majority vote of its neighbors, whenever we have a new point to classify, we find its K nearest neighbors from the training data. The distance is calculated with measures such as Euclidean Distance, Minkowski Distance and Mahalanobis Distance.

In the training process, each training sample $<X, f(X)>$ is added to the list of training_samples. In the test process, given a query sample X' to be classified, Let X_1, $X_2, \ldots X_k$ denote the k instances from training_smaples that are nearest to X'. Return the class that represents the maximum of the k samples.

2.1.3 Classification and Regression Tree (CART)

CART is built in accordance with the splitting rule that performs the splitting of the learning sample into smaller parts. Each data have to be divided into two parts for maximum homogeneity. Maximum homogeneity of child nodes is defined by the impurity function. The classification problem of CART is to assume a multinomial model and then use deviance as a definition of impurity.

Assume $Y \in \{1, 2, ...k\}$, at each node i of a classification tree with a probability distribution p_{ik} over the k classes. A random sample n_{ik} from the multinomial distribution specified by the probabilities p_{ik} is observed. Given X, the conditional likelihood is then proportional to $\prod_{(leaves_i)}\prod_{(leaves_k)} p_{ik}^{n_{ik}}$. After defining a deviance $D=\sum D_i$,where $D_i=-2\sum_k n_{ik}\log(p_{ik})$, $\widehat{p}_{ik}$ is estimated by $\widehat{p}_{ik}=\frac{n_{ik}}{n_i}$ .Three commonly measures are often used for node impurity in CART, which are Gini index, cross-entropy and deviance.

2.1.4 Random Forest (RF)

RF is an ensemble learning method for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Single decision trees often have high variance or high bias. RF attempts to mitigate the problems of high variance and high bias by averaging to find a natural balance between the two extremes. RF grows many classification trees as following:

1. Randomly select k features from total m features, where $k << m$
2. Among the k features, calculate the node d using the best split point.
3. Split the node into children nodes using the best split.
4. Repeat 1 to 3 steps until l number of nodes has been reached.
5. Build forest by repeating steps 1 to 4 for n number times to create n number of trees.

2.1.5 Stochastic Gradient Descent (SGD)

SGD is a stochastic approximation of the gradient descent optimization and iterative method for optimizing a differentiable objective function. Gradient descent is a way to minimize an objective function $J(\theta)$ parameterized by updating the parameters θ in the opposite direction of the gradient of the objective function ∇θJ(θ) w.r.t. to the parameters. The learning rate η determines the size of the steps we take to reach a local minimum. In other words, we follow the direction of the slope of the surface created by the objective function downhill until we reach a valley. SGD performs a parameter update for each training example X_i and label Y_i, $\theta=\theta-\eta\cdot\nabla_\theta J(\theta;X_i;Y_i)$.

2.1.6 Markov model

2.1.6.1 Hidden Markov Model (HMM)

HMM is one of the Markov processes, which belong to stochastic processes that generate random sequences of states according to certain probabilities. By considering the probability over a sequence of observations, using an HMM, the hidden occupancy state is inferred.

In HMM, the occupancy detection problem is as following: For time period T, t_i is a given time interval, observations are the values of environmental data from recorded sensors, and hidden states are the possible occupancy states of the building that cause the sensor outputs.

Let the state s_t be the occupancy state of the system at time t_i, with a likelihood of an observation, $p(x|s_t)$, where x, is a is feature vector of continuous values derived from the sensors and i is the number of the time interval. If O:$\{x_1,x_2,x_3\ldots x_N\}$, a sequence of observation vectors, at each time interval, t_i, a new state is entered. At each step of the process the model may generate an observation depending on which state it is in and then make a transition to another state. An important characteristic of the Markov model is that the next state depends only on the current state and not on the previous transitions that lead to the current state. In other words, the probability of being in a state at time t depends only on the state at time t-1.

$$P(s_t \mid s_{t-1}, s_{t-2}, \cdots s_1) = P(s_t \mid s_{t-1}) \tag{2}$$

2.1.6.2 Gaussian mixture models-Hidden Markov Model (GMM-HMM)

In GMM-HMM Model, assume that components of the mixture have Gaussian distribution with mean μ and variance σ^2 and probability density function (3).

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{3}$$

We let $o_1, o_2, \ldots, o_L$ denote the observations. Then each observation is assumed to be an independent realization of the random variable O with three-component mixture probability density function (4) and log-likelihood function (5). With the maximum likelihood approach to estimate the model: $M = \langle v, N(\mu_1,\sigma_1), \ldots N(\mu_k,\sigma_k) \rangle$, such values of ν_i, μ_i, andσ_i, need to be found that maximize the function (5). To solve the problem, the Expectation Maximization (EM) Algorithm is used.

$$f(o;M) = \sum_{i=1}^{K} v_i f_i(o;\mu_i,\sigma_i) \qquad \sum_{i=1}^{K} v_i = 1 \tag{4}$$

$$\log L(M) = \sum_{j=1}^{L} \log\left(\sum_{i=1}^{K} v_i f(o_j;\mu_i,\sigma_i) \right) \tag{5}$$

2.2 Evaluation with ground truth data

True Positives (TP) are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes. True Negatives (TN) are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no. False Positives (FP) and False Negatives (FN) occur when your actual class contradicts with the predicted class. FP occurs when actual class is no and predicted class is yes. FN occurs when actual class is yes but predicted class is no.

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall (Sensitivity) is the ratio of correctly predicted positive observations to the all observations in actual class - yes. They are then defined as:

$$Precision = \frac{TP}{TP+FP} \tag{6}$$

$$Recall = \frac{TP}{TP+FN} \tag{7}$$

Accuracy is a ratio of correctly predicted observation to the total observations.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

F_1 score is a measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F_1 = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall} \tag{9}$$

Mean Squared Error (MSE) assesses the quality of a predictor.If $\hat{Y}$ is a vector of n predictions generated from a sample of n data points on all variables, and Y is the vector of observed values of the variable being predicted, then the within-sample MSE of the predictor is computed as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{10}$$

2.3 K-fold Cross Validation

K-fold Cross Validation is one way to improve over the holdout method. The data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subset is used as the test set and the other *k*-1 subsets are put together to form a training set. Then the average error across all *k* trials is computed. The variance of the resulting estimate is reduced as *k* is increased.

3 Experiments

3.1 Datasets

Datasets come from [10]. An office room with approximate dimensions of 5.85m × 3.50m × 3.53m (W× D × H) was monitored for the following variables: temperature, humidity, light and CO_2 levels. The humidity ratio was calculated using the measured temperature and relative humidity. The sampling time logged about 1 min for the occupancy states is the focus of time-aware analysis. The 1 min reporting interval was chosen to be able to capture quick changes in occupancy states. Data training with 8143 samples are taken as training set and dataset2 with 9752 samples are used as testing set.

As show in Table 1, the features include temperature, humidity, derived humidity ratio, light, CO_2, occupancy status (0 for non-occupied, 1 for occupied) and time stamp. As time feature is very important, we extract two time-related features from data and time as follows: 1. Day of the week: Week Index that denoted as WI indicates whether a date is a working day. 2. Time of a day: Minutes Index that denoted as SI indicates number of minutes from midnight to current time for each day.

Let occupancy status denote the feature variable whose values we wish to predict, and let features include WI, SI, temperature, humidity, humidity ratio, light and CO_2 denote the feature variables from which we wish to do prediction.

Table 1 Feature Information of Databases

Attribute	information	

Date	year-month-day	WI
		0(weekend)
		1(weekday)

time	hour:minute:second	MI (minutes)
Temperature	Celsius(℃)	
Humidity	RH	
CO_2	ppm	
HumidityRatio	kgwater-vapor/kg-air	
Light	lux	
Occupancy status	0(non-occupied) 1(occupied)	

3.2 Data analysis

To represent the distribution of numerical data,as shown in Table 2, temperature and humidity have a lower standard deviation that indicates the data points tend to be close to the mean of the set. While light and CO_2 have higher standard deviations that indicate the samples are spread out over a wider range of values. 25%/50%/75% show the feature value that

25%/50%/75% samples‘variable value is less than.For instance, 25% samples’temperature value is less than 19.70.

Table 2 Data distribution information of training set(mean-average values, std-standard deviation, min-minimum value, max-maxmum value)

	Temperature	Humidity	Light	co$_2$	Humidity Ratio
mean	20.62	25.73	119.52	606.55	0.003863
std	1.02	5.53	194.76	314.32	0.000852
min	19.00	16.75	0.00	412.75	0.002674
25%	19.70	20.20	0.00	439.00	0.003078
50%	20.39	26.22	0.00	453.50	0.003801
75%	21.39	30.53	256.38	638.83	0.004352
max	23.18	39.12	1546.33	2028.50	0.006476

The histogram shows the frequency of a given variable in the interval. X axis is the given variable value, Y axis is frequency of a given variable value. The histograms of training set, training set with occupied statues and non-occupied statues are shown in fig.1A, fig.1B and fig.1C respectively, which are accurate representation of the distribution of numerical data. We can see the probability distribution of a given variable, CO$_2$, by depicting the 6000 frequencies of observations occurring in 400~500 ppm when the building is non-occupied. Light occur about 450 lux when the building is occupied and about 0 lux when the building is non-occupied. Temperature was mainly concentrated up 20.5℃ when the building is occupied and mainly concentrated below 20.5℃when the building is non-occupied.

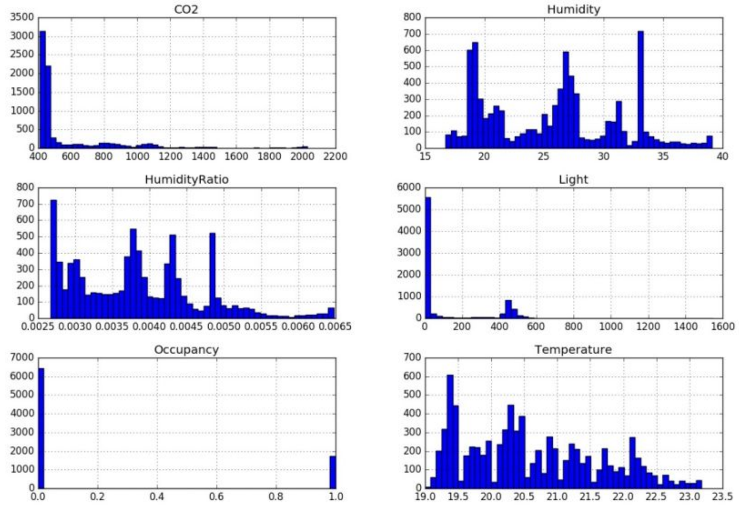

Fig.1A Histogram of the training set

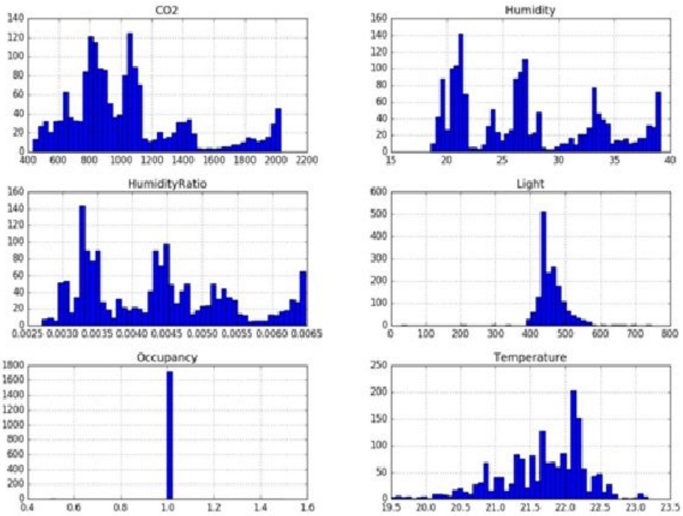

Fig.1B Histogram of the training set under occupied statues

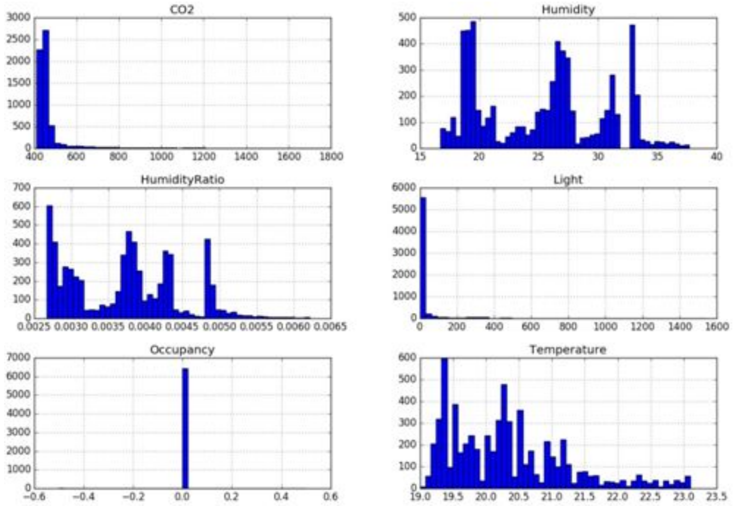

Fig.1C Histogram of the training set under non-occupied statues

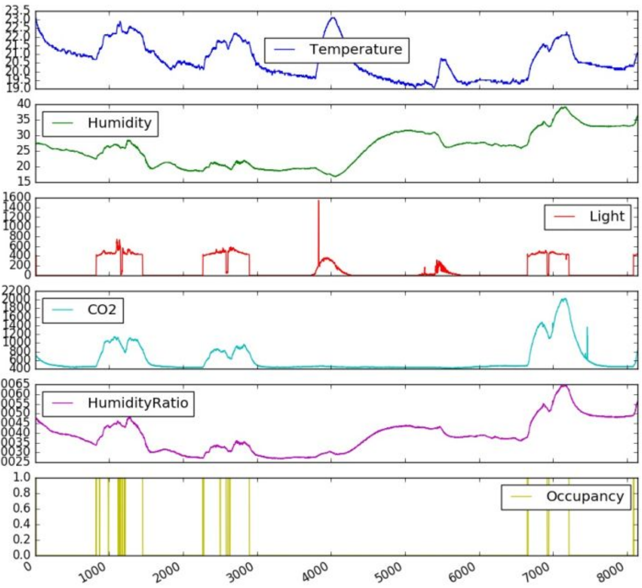

Fig.2 Profiles for the whole period of the training set

Fig. 2 shows the feature data for the whole time period of training set. Horizontal axis represents the number of minutes from start to current time. Vertical axis represents feature values. We can see that feature value of every variable and occupancy status have positive correlation relationship with a synchronous trend of change.

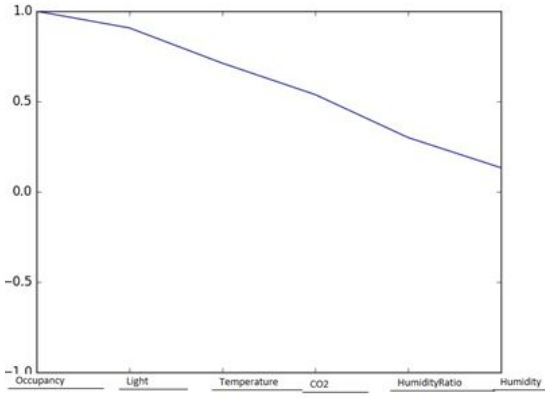

Fig.3 Variable importance for occupied statues with Pearson correlation coefficient

Fig. 3 shows variable importance for occupied statues by Pearson Correlation Coefficient (PCC). X axis is feature variables. Y axis is Correlation Coefficient.It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. Fig.3 is a measure of the linear correlation between variables (such as Temperature, Humidity, Humidity Ratio, Light, CO_2) and (Occupancy

Status). The importance of Temperature, Humidity, Humidity Ratio, Light and CO_2 to occupied statues is decreasing by Light, Temperature, CO_2, Humidity ratio, Humidity.

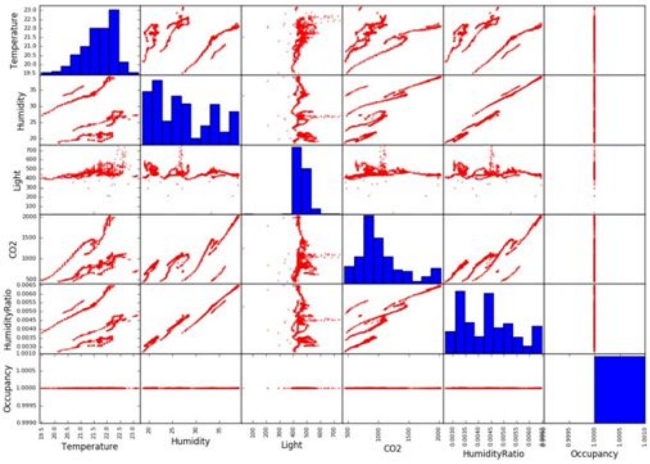

Fig.4A Pairs plot of occupied statues

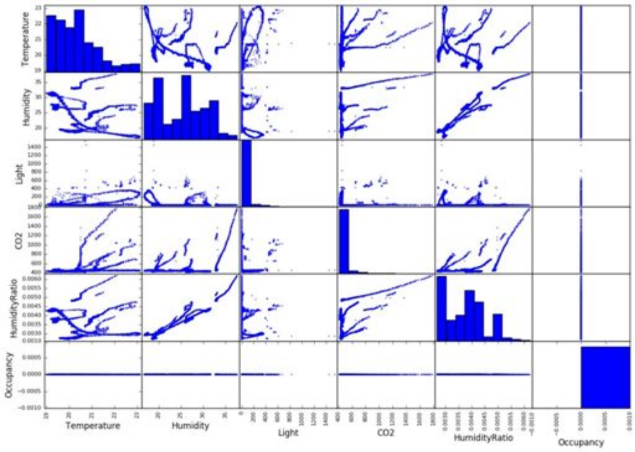

Fig.4B Pairs plot of non-occupied statues

Fig. 4A and Fig. 4B gives a pairs plot showing the relationship for all the variables.Both X axis and Y axis are feature variables. Fig. 5A is the pair plots in occupied status, and Fig. 5B represents is the pair plots in non-occupied status. Pairs plots are a useful way of displaying the pairwise relations between variables, in which one variable in the same data row is matched with another variable's value.We can see:

(1) Fig. 4A and Fig. 4B bring out a striking contrast between occupied and non-occupied status.

There is a clear separation in the temperature and light plot, light and humidity plot, light and

(2) CO_2 plot and light and humidity ratio plot, which indicates that these pair combinations are good candidates for training and testing the occupancy status.

(3) All variables including temperature, humidity, humidity ratio, light and CO_2 have linear relation with occupancy status.

(4) There is a clear correlation between light and all other variables.

3.3 Experiment results and discussion

With the help of above described data analysis,we could understand the time-aware occupancy states better. The below experiments under six different prediction horizons had been conducted for time-aware prediction, that is, 1 minute (real time), 10 minutes, 20 minutes, 30 minutes, 45minutes, 60 minutes and 90 minutes.

3.3.1 Real time detection

The sampling time of the databases logged about 1 min can be able to capture quick changes in occupancy states. So 1 minute reporting interval was chosen to be real time detection.

Table 3 Accuracy of HMM and GMM-HMM

Model	Accuracy
HMM_Temperature	80.09%
HMM_Humidity	50.92%
HMM_Light	98.14%
HMM_CO_2	94.90%
HMM_Humidity ratio	41.66%
GMM-HMM(including 5 features)	96.48%

Table 3 presents an accuracy of HMM and GMM-HMM. In HMM, each feature is used to determine occupancy status. And in GMM_HMM, five features including temperature, humidity, light, CO_2 levels and humidity ratio are used together to determine occupancy status. From the table 3, it can be seen that the best performance, 98.14%, is found using Light feature with HMM, which verified that Light play the most important role for occupancy detection in fig. 3. The second best performance is GMM-HMM at 96.48%.

Table 4 Evaluation criteria of GMM-HMM

	GMH-HMM(5 features)	GMH-HMM(7 features)
Precision	0.858	0.970
Recall	0.998	0.997
F1	0.923	0.984
Accurancy	96.48%	99.48%

Table 4 shows the number of features difference with the GMM-HMM. As shown in Table 1, except for temperature, humidity, light, CO2 levels and humidity ratio, another two features WI and MI are also used to determine a probability of occupancy. In Table 4, for the GMM-HMM model, 7 features perform better than 5 features from four evaluation criterion. It can be seen that time is one of the most important features in influencing occupancy detection. After adding two features, WI, reflecting whether a day is a working day, MI, representing time series by the number of minutes from midnight to current time, GMH-HMM with 7 features with day and time performs better than 5 features. Thus in the following experiments, we use 7 features as shown in Table 1. Table 5 show evaluation criteria of six machine learning models. LR, KNN, CART, RF, SGD and GMM-HMM are as machine learning models, and Precision, Recall, F1, Accuracy and MSE are as evaluation criteria. All evaluation criteria are performed by 10 cross validation. The results are as shown in table 5.

Table 5 Evaluation criteria of 6 models

Model	Precision	recall	F1	Accuracy	MSE
LR	0.968	0.984	0.976	98.98%	0.180
KNN	0.910	0.937	0.923	95.45%	0.140
CART	0.915	0.934	0.924	95.43%	0.135
RF	0.930	0.971	0.950	97.87%	0.123
SGD	0.975	0.993	0.984	99.17%	0.122
GMM-HMM	0.970	0.997	0.984	99.48%	0.119

From table 5, it can be seen that the best performance is found by GMM-HMM, in which the precision is 0.970, the recall is 0.997, the F1 is 0.984, the accuracy is 99.48% and MSE is 0.119. SGD is only worse than GMM-HMM with minimal difference. And the precision is 0.975, the recall is 0.993, the F1 is 0.984, the accuracy is 99.17% and MSE is 0.122 by SGD. As shown in fig.5, real and predicted occupancy states by GMM-HMM have a very high coincidence.

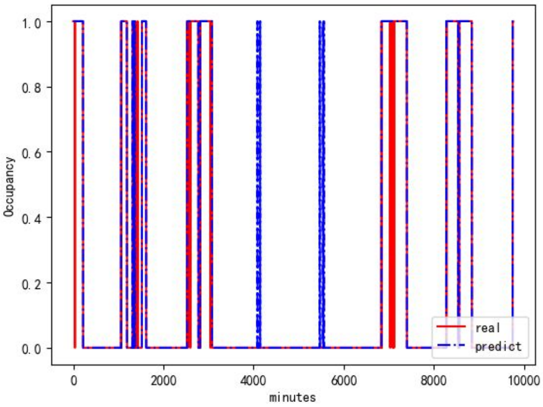

Fig. 5 Real and predicted occupancy states by GMH-HMM

3.3.2 Time-Aware detection

Time-aware occupancy detection could give a guideline for intelligent building energy management. Six different size of prediction window occupancy detection are provided by GMM-HMM and SGD because of the best two performances in precision, recall, F1, accuracy and MSE. Time-aware occupancy detection at 10 min, 20 min, 30 min, 45 min, 60 min and 90 min were conducted. The experiment results are as in Table 6.

Table 6. Evaluation criteria of SGD and GMM-HMM at 10 min, 20 min, 30 min, 45 min, 60 min and 90 min time-aware occupancy detection.

		10 min	20 min	30 min	45 min	60 min	90 min
Precision	SGD	0.980	0.980	0.971	0.952	0.926	0.881
	GMM	0.981	0.981	0.972	0.955	0.925	0.910
Recall	SGD	0.990	1.0	1.0	0.973	0.970	0.952
	GMM	0.995	1.0	1.0	0.980	0.972	0.961
F1	SGD	0.985	0.990	0.985	0.962	0.947	0.915
	GMM	0.987	0.990	0.986	0.967	0.948	0.935
Accuracy	SGD	99.28%	99.6%	99.4%	99.4%	99.2%	99.0%
	GMM	99.48%	99.6%	99.4%	99.4%	99.3%	99.1%
MSE	SGD	0.129	0.132	0.141	0.146	0.170	0.193
	GMM	0.125	0.125	0.132	0.145	0.163	0.182

From the table 6, we can see:

(1) GMM-HMM perform better than SGD. The F1 of GMM-HMM ranges from 0.987 at 10 minutes to 0.935 at 90 minutes, the F1 of SGD ranges from 0.985 at 10 minutes to 0.915 at 90 minutes. Precision, recall, accuracy also tend to decrease with the time window expands. While MSE of GMM-HMM increases from 0.125 at 10 minutes to 0.182 at 90 minutes, while the MSE of SGD increases from 0.129 at 10 minutes to 0.193 at 90 minutes.

(2) It is worth noting that occupancy detection arrive the best performance at 20 minutes. So 20 minutes occupancy detection with automatic control the operation time of HVAC systems, lighting control systems and the other appliances in buildings will reducing the energy consumption effectively while maintaining thermal comfort.

4 Conclusion

Occupancy detection is considered a critical impact factor of energy consumption in commercial and residential buildings. By presenting several machine learning models for more accurate occupancy detection, we hope to promote the development of more energy-efficient HVAC scheduling systems to reduce overall energy consumption.

In this paper, several occupancy detection methods are proposed with different size of prediction window. The experimental results show that GMM-HMM at 20 minutes occupancy detection outperforms the other prediction window, which demonstrated a guideline to control the operation time of HVAC systems, lighting control systems and the other appliances in buildings.

References

[1] Masoso O T, Grobler L J. The dark side of occupants' behaviour on building energy use[J]. Energy & Buildings, 2010, 42(2):173-177.

[2]Dong B, Andrews B. Sensor-based occupancy behavioral pattern recognition for energy and comfort management in intelligent buildings[J]. Proceedings of Building Simulation, 2009.

[3] Sangogboye FC, Imamovic K, Kjærgaard MB. Improving occupancy presence prediction via multi-label classification. In: 2016 IEEE int conf pervasive comput commun Workshop PerCom Workshop; 2016. p. 1–6.

[4] Ortega JLG, Han L, Whittacker N, Bowring N. A machine-learning based approach to model user occupancy and activity patterns for energy saving in buildings. Sci Inf Conf SAI 2015:474–82. http://dx.doi.org/10.1109/SAI.2015.7237185.

[5]Chaney J, Owens E H, Peacock A D. An evidence based approach to determining residential occupancy and its role in demand response management[J]. Energy & Buildings, 2016, 125:254-266.

[6] Peng Y, Rysanek A, Nagy Z, et al. Using machine learning techniques for occupancy-prediction-based cooling control in office buildings[J]. Applied Energy, 2018, 211.

[7] Ryu S H , Moon H J . Development of an occupancy prediction model using indoor environmental data based on machine learning techniques[J]. Building and Environment, 2016, 107:1-9.

[8] Chen Z, Soh YC. Comparing occupancy models and data mining approaches for regular occupancy prediction in commercial buildings. J Build Perform Simul 2016:1–9.

[9] Capozzoli A, Piscitelli MS, Gorrino A, Ballarini I, Corrado V. Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. Sustain Cities Soc 2017;35:191–208.

[10] Candanedo L M, Feldheim V. Accurate occupancy detection of an office room from light, temperature, humidity and CO 2, measurements using statistical learning models[J]. Energy & Buildings, 2016, 112:28-39.