Chapter 26
A Weight Clustering-Based Pattern Recognition Method for Improving Building's Cooling Load Prediction Reliability

Sihao Chen, Liangzhu Leon Wang, Jing Li, Guang Zhou, and Xiaoqing Zhou

Abstract Accurate building cooling load prediction can effectively guide the start-stop strategies and capacities matching of chillers and is also the basis of model predictive control of the heating, ventilation, and air conditioning (HVAC). Most of the existing literature focused on the structural optimization or selection of cooling load prediction models, and rarely in-depth studies on the matching between data and models. However, the data features determine the upper limit of model prediction performances, thus leading the unsatisfactory prediction accuracy in the existing methods. Aiming at this, the paper proposed a novel weight clustering-based pattern recognition method for improving building cooling load prediction reliability. Firstly, after the outliers were removed, the Pearson correlation analysis was used to select the key input variables for the models. Secondly, the sensitivity analysis was utilized to obtain the weights of input variables on the cooling load, and then the weights were introduced into the K-means clustering algorithm. Finally, the training data of models were classified by the clustering, and the corresponding training set was matched according to the predicted sample's features. The case study showed that the weight clustering-based pattern recognition method has a significant improvement in prediction accuracies to the multiple linear regression (MLR), multiple nonlinear regression (MNR), and artificial neural network (ANN) models (e.g., 35%, 36%, and 15% reduction in mean absolute percentage error (*MAPE*), respectively), In addition, the optimal clustering number, the clustering effects with or without the weights, etc. were also investigated. This paper's method can provide a novel idea for the models' data preprocessing.

S. Chen · X. Zhou (✉)
Guangdong Provincial Key Laboratory of Building Energy Efficiency and Application Technologies, Academy of Building Energy Efficiency, School of Civil Engineering, Guangzhou University, Guangzhou 510006, China
e-mail: zhouxq@gzhu.edu.cn

S. Chen · L. L. Wang · J. Li
Centre for Zero Energy Building Studies, Department of Building, Civil and Environmental Engineering, Concordia University, Montreal H3G 1M8, Canada

G. Zhou
Zhongkai University of Agriculture and Engineering, Guangzhou 510225, China

L. L. Wang et al. (eds.), *Proceedings of the 5th International Conference on Building Energy and Environment*, Environmental Science and Engineering,
https://doi.org/10.1007/978-981-19-9822-5_26

Keywords Cooling load prediction · Pattern recognition · Clustering analysis · Data preprocessing · Mode identification

26.1 Introduction

The prediction process of cooling load mainly includes data preprocessing and optimization of prediction models. Data preprocessing usually comprises removing outliers, filling missing values (for time-series models), eliminating redundant variables, and data normalization. The prediction models can be divided into physical models, data-driven models, and semi-physical models in terms of modeling mechanisms (Xiao et al. 2022). Data-driven models less rely on expert knowledge, thus making the modeling process more simple (Yao and Shekhar 2021), which commonly includes regression analysis (e.g., multiple linear regression MLR and multiple nonlinear regression MNR), artificial neural network (ANN), support vector regression, decision tree regression, etc. (Chen et al. 2022; Zhang et al. 2021). Most of the existing studies focused on models' optimization and comparison while ignoring the preliminary work of models, i.e., data preprocessing. For data-driven models, data quality determines the upper limit of model performances (Xiao et al. 2022). Therefore, in addition to the typical data preprocessing methods, more attention should be paid to the features of the data.

Besides, models' input data may present different statistic distributions, i.e., having variational load patterns so that difficult to guarantee prediction accuracy if the same model is used to predict the data with different load patterns (Chen et al. 2022). Based on this, some researchers introduced unsupervised clustering into the cooling load prediction, i.e., different types of the data are expressed by different models, and therefore obtained a better improvement in prediction accuracy. For instance, Zhang et al. (Zhang et al. 2019) used the K-means clustering to classify the model's input data and then used the *K*-nearest neighbor to determine the training data class of the model. The case revealed that this method could improve the prediction accuracy of the cooling load in the factory workshop by 10%. Ding et al. (Ding et al. 2018) used the K-means and hierarchical methods to cluster the input variables of the ANN and support vector regression, thus improving the results of the cooling load prediction for the office buildings. Ko et al. (Ko et al. 2017) applied the clustering technique to enhance the prediction accuracy of the regression analysis model. However, in these clustering methods, most of them did not consider the influence of the input variables on the cooling load, i.e., they put each input variable equally into the clustering space, so they did not introduce the weights of the variables into the process of the clustering that affects the cooling load. In the production of actual cooling load, the contributions of different input variables to the cooling load are different. If they are all treated equally, it makes the clustering effect getting poor, as a result of little improvement in the prediction accuracy. Aiming at the unsatisfactory prediction accuracy in the existing methods, this paper proposed a novel

weight clustering-based pattern recognition method for improving building cooling load prediction reliability.

26.2 Pattern Recognition Method Based on Weight Clustering

The cooling load pattern recognition method is shown in Fig. 26.1. The original data are first preprocessed and randomly divided into 70% training set and 30% validation set. Secondly, the training set is divided into K classes by using the K-means clustering that is based on the weights of input variables affecting the output cooling load, and then the center points of each class are obtained. Subsequently, the distances between the predicted sample j in the validation set and the centers of each training class are compared, and the training set i with the smallest distance is selected as the training samples for the current model (i.e., the MLR, MNR, and ANN). Finally, the well-trained model is used to predict the corresponding sample j of the validation set. The clustering-based training method can effectively improve the matching between the training samples and the predicted sample for the models, and thus can obtain better prediction accuracy.

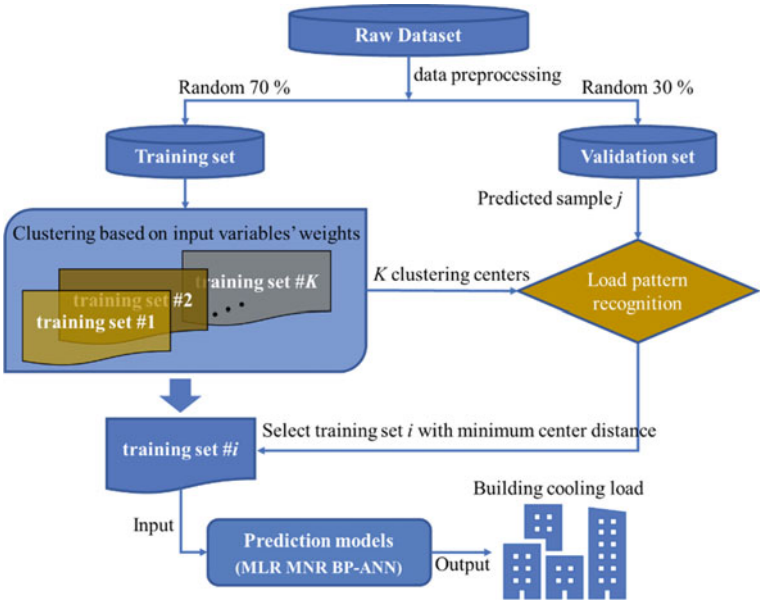

Fig. 26.1 The framework of the training pattern recognition for cooling load prediction

26.3 Data Preprocessing

26.3.1 Outlier Detection

Assuming that the input variables of the model obey the normal distribution, thus the 3σ criterion (Fan et al. 2021) can be used to judge the samples whether are outliers. Equation (26.1) is for solving the mean value $\overline{x}$ of variable x. Equation (26.2) is the solution of standard deviation σ. Equation (26.3) is the discriminant of whether sample x is an outlier value. Note that the outliers removed should be filled especially in the time-series models. The filling method for missing values can adopt linear interpolation.

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{26.1}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2} \tag{26.2}$$

$$|x_i - \overline{x}| > 3\sigma \tag{26.3}$$

26.3.2 Pearson Correlation Analysis

To reduce the complexity of models and decrease the unnecessary investment in redundant points measured, the Pearson correlation analysis method (Ding et al. 2018) is used to reduce the input dimensions of models. In Eq. (26.4), the $\rho_{x,y}$ denotes the size of linear correlation between x and y. The variables with high correlation should be removed.

$$\rho_{xy} = \frac{\sum_{i=1}^{N}(x-\overline{x})(y-\overline{y})}{\sqrt{\sum_{i=1}^{N}(x-\overline{x})^2 \sum_{i=1}^{N}(y-\overline{y})^2}} \tag{26.4}$$

26.3.3 Data Normalization

The variables are normalized as shown in Eq. (26.5) to prevent numerical problems (Fan et al. 2019). Note that the prediction cooling loads of models are needed to be normalized inversely. Where $x_{\min}$ and $x_{\max}$ represent the minimum value and the maximum value of variable x, respectively.

$$x = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{26.5}$$

26.4 Clustering of Training Data

The clustering process of training data is first required to obtain the variables' weights affecting the cooling load and then the weights are introduced into the K-means clustering space.

26.4.1 Determination of Input Variables' Weights

Figure 26.2 shows the two-dimensional attribute space normalized of the cooling loads, which assumes that the cooling loads are only affected by outdoor dry-bulb temperature T and outdoor relative humidity H. Figure 26.2 presents that the cooling load CL_1, cooling load CL_2, and cooling load CL_3 have the same distance from predicted cooling load CL_0, while CL_3 is closer to CL_0 in the outdoor dry-bulb temperature, and CL_2 is closer to CL_0 in the outdoor relative humidity. Then which points of cooling load should be chosen to estimate the CL_0? From the sensitivity analysis of the cooling load, the outdoor dry-bulb temperature has a greater impact on the cooling load than the outdoor relative humidity. Therefore, from the perspective of probability, using the cooling load CL_3 to estimate CL_0 is more preferred. In this paper, the Pearson correlation coefficients after being taken as absolute values are chosen as the variables' weights.

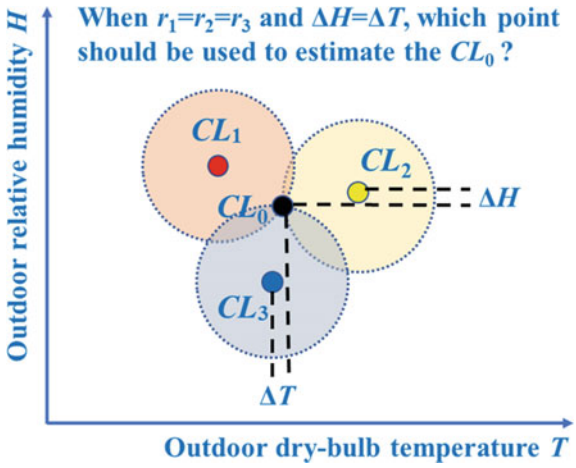

Fig. 26.2 Normalized dimensionless attributes space (a two-dimensional example)

26.4.2 K-means Algorithm for Classifying Training Data

Equation (26.6) is the objective function of the weights-based K-means algorithm, which is to minimize the weighted distance E between class center u and class samples x. Equation (26.7) is the solution to class center point u. Through repeated iterations, when the position of class center point u no longer changes, it is considered that the clustering has been completed. The K-means clustering is a classical unsupervised learning method and is described in detail by Ref. (Zhang et al. 2019; Ding et al. 2018).

$$\arg\min_{u} E = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - u_i\|_2^2 \cdot W_x \tag{26.6}$$

$$u_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{26.7}$$

where C_i denotes the ith class of training samples. K denotes the total cluster number. W_x denotes the weights vector of input variables, which is composed of the Pearson correlation coefficients above.

26.4.3 Pattern Recognition of Validation Data

After dividing each class of training samples via the K-means algorithm, the distance d_{ij} between the predicted sample j and the center point i of each training class are compared, and the training class with the minimum distance d_{ij} is selected as the training set of the current model to predict sample j. The selection of the training class i is shown in Eq. (26.8). Where m denotes the dimensions' number of sample x. For a more intuitive depiction please see Fig. 26.1.

$$\arg\min_{i} d_{ij} = \sum_{l=1}^{m} \left(x_{j,l} - u_{i,l}\right)^2 W_x(l) \tag{26.8}$$

26.5 Case Study

The EnergyPlus software was first used to simulate the energy consumption of the typical office building in Guangzhou to obtain the raw data for training and validating the models above. The Matlab software was second used for implementing the methods of this paper.

Table 26.1 Variables of the prediction models

Variable	Variable name	Unit	Type
T	Outdoor dry-bulb temperature	°C	Input
H	Outdoor relative humidity	%	Input
R	Solar radiation	W/m^2	Input
P	Occupant	P	Input
CL	Cooling load	W	Output

26.5.1 Cooling Load Data of a Typical Office Building

A typical office building energy model in Guangzhou derived from Ref. (Lv et al. 2019), was used to obtain the training and validation data of the models. The hourly cooling load data (1188 samples) of the typical office buildings in Guangzhou during the cooling season (May to September) were obtained by the EnergyPlus.

It is found that the correlation ρ between the three variables of the indoor occupant, indoor lighting power, and indoor equipment power is very high, so the occupant obtained easily is selected as the representative for the other two variables. The input and output variables of models are shown in Table 26.1. The MLR, MNR, and ANN were selected as the prediction models.

26.5.2 Predictive Performance of Models Based on the Load Pattern Recognition

Figure 26.3a shows the variation of prediction performance of the MLR with the increase of clustering number K. It can be seen that the clustering-based pattern recognition method obtained a great improvement in the model performance. Compared with the non-clustering method, its *MAPE* decreased by 35% on average. Due to the introduction of the variables' weights, the clustering performance was further improved, and its *MAPE* decreased by 6% on average. Figure 26.3a also indicates that with the rise of the clustering number, the model performance first increased and then declined. When the clustering number was 7, the model had the optimal prediction accuracy ($MAPE = 4.53\%$). When the clustering number exceeded 9, the model performance began to deteriorate, and the clustering number is recommended between 3 and 7 for the MLR.

Figure 26.3b presents the variation of prediction performance of the MNR with the rise of clustering number K. It can be seen that the clustering-based method makes a great improvement in the model performance. Compared to the non-clustering model, its *MAPE* decreased by 36% on average. Due to the introduction of the weights, the clustering performance was further improved, and its *MAPE* decreased by 8% on average. Figure 26.3b also displays that with the increase of the clustering number, the model performance first increased and then decreased. When the

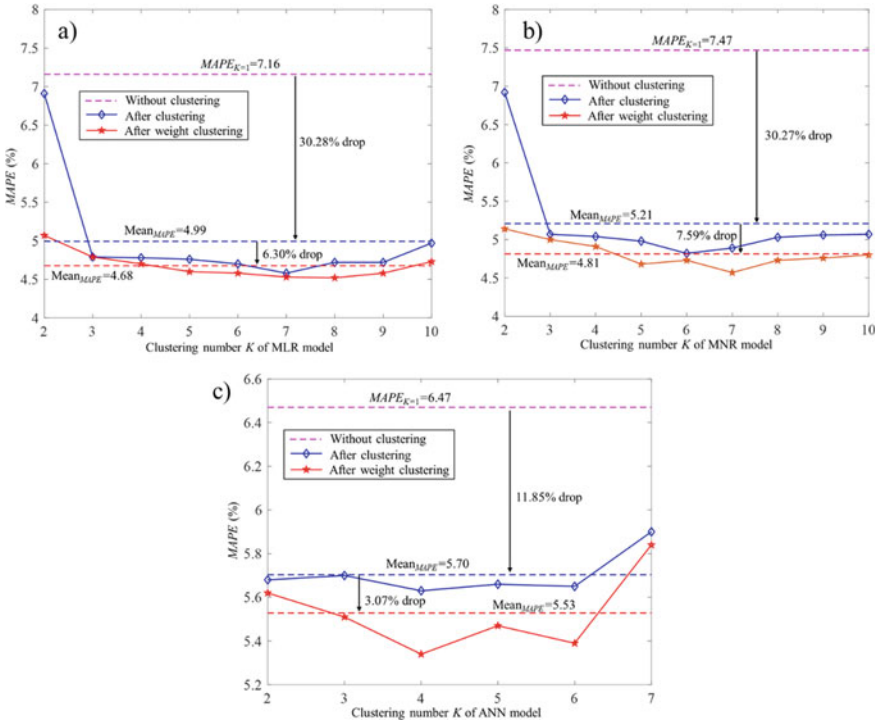

Fig. 26.3 MAPEs varying with clustering number K: **a** MLR; **b** MNR; **c** ANN

clustering number was 7, the model had the best prediction accuracy ($MAPE$ = 4.57%). When the clustering number exceeded 9, the model performance started to deteriorate, and the number of clusters is recommended between 3 and 7 for MNR. Figure 26.3c shows the variation of prediction performance of the ANN with the increase of clustering number K. It can be seen that the clustering-based method obtains a great improvement in the model performance. Compared with the non-clustering, its $MAPE$ decreased by 15% on average. Due to the introduction of the weights, the clustering performance was further improved, and its $MAPE$ decreased by 3% on average, respectively. Figure 26.3c reveals that the regularity of change to the ANN was not obvious with the increase in clustering number. When the clustering number was greater than 6, the model performance would deteriorate. Overall, the clustering number is recommended between 2 and 6 for the ANN.

In summary, the clustering-based pattern recognition method achieves a great improvement in the prediction performance for the MLR, MNR, and ANN. The introduction of the variables' weights into the clustering will further improve the prediction accuracy for the models. Compared to the ANN, the MLR and MNR have a higher improvement in accuracy and better prediction stability when with the variation of the clustering numbers. When the number of clusters is about 4, the robustness of the above models can be guaranteed.

26.5.3 Explanations of the Results

(1) The main reason why the clustering-based pattern recognition method had a great improvement in the model's prediction accuracy is that the unsupervised learning method was used to cluster the model's training samples so that the training samples with the same class had more similarity. After identifying the spatial attributes of the predicted sample, the training class with the greatest correlation was selected to train the sub-model, thus the parameters of the sub-model would better match the current predicted sample. (2) The main reason why the introduction of the variables' weights into the clustering process further improved the model performance is that the weights-based clustering method can achieve a better clustering effect, which thus will further improve the prediction accuracy for the models. (3) In the case study, it can be seen that, under the non-clustering condition, ANN had higher prediction accuracy than the MLR and MNR due to its stronger nonlinear fitting ability. However, after the clustering, the prediction accuracies of the MLR and MNR were much higher than the ANN (e.g., their *MAPEs* less than the ANN by 14% on average). This indicates that the clustering-based pattern recognition method is more suitable for the low complexity models. The main reason is that the number of training samples in each class was reduced due to the division of the total training samples through the clustering. While in the few-shot learning, the generalization ability of the ANN is poor, so the clustering effect on the ANN is not as good as the MLR and MNR.

26.6 Conclusion

Aiming at the unsatisfactory prediction accuracy in the existing methods, this paper proposed a novel weight clustering-based pattern recognition method for improving building cooling load prediction accuracy. The case study showed that this method achieved a significant improvement in the MLR, MNR, and ANN, e.g., *MAPEs* decreased by 35%, 36%, and 15% on average, respectively. Compared to the non-weight clustering, the introduction of the weights further improved the prediction accuracy of the models, such as *MAPEs* reduced by 6%, 8%, and 3% on average, respectively. When the clustering number was about 4, the models had a more stable prediction performance. In future research, the combination of this method with the real-time optimization and online feedback calibration will be investigated.

Acknowledgements This work was financially supported by the National Natural Science Foundation of China (No. 52078146) and the Key Projects of Basic Research and Applied Basic Research of Universities in Guangdong Province (No. 2018KZDXM050).

References

Chen S, Zhou X, Zhou G, Fan C, Ding P, Chen Q (2022) An online physical-based multiple linear regression model for building's hourly cooling load prediction. Energy Build 254:111574

Ding Y, Zhang Q, Yuan T, Yang F (2018) Effect of input variables on cooling load prediction accuracy of an office building. Appl Therm Eng 128:225–234

Fan C, Ding Y, Liao Y (2019) Analysis of hourly cooling load prediction accuracy with data-mining approaches on different training time scales. Sustain Cities Soc 51

Fan C, Chen M, Wang X, Wang J, Huang B (2021) A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. Front Energy Res 9:652801

Ko J-H, Kong D-S, Huh J-H (2017) Baseline building energy modeling of cluster inverse model by using daily energy consumption in office buildings. Energy Build 140:317–323

Lv Y, Peng H, He M, Huang Y, Wang J (2019) Definition of typical commercial building for South China's Pearl River Delta: local data statistics and model development. Energy Build 190:119–131

Xiao T, Xu P, He R, Sha H (2022) Status quo and opportunities for building energy prediction in limited data context—overview from a competition. Appl Energy 305:117829

Yao Y, Shekhar DK (2021) State of the art review on model predictive control (MPC) in Heating Ventilation and Air-conditioning (HVAC) field. Build Environ:107952

Zhang ZT, Ding Y, Lu Y, Niu J (2019) Development and evaluation of cooling load prediction models for a factory workshop. J Cleaner Prod 230:622–633

Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y, Livingood W (2021) A review of machine learning in building load prediction. Appl Energy 285:116452