Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231223

A Building Energy Consumption Prediction Model Based on Data Mining and Cluster Analysis

Yunli ZHANG[a,1], Liang ZHANG[b]
[a] *School of Municipal and Environmental Engineering, Shenyang Urban Construction University, Shenyang, 110168, China*
[b] *Shenyang Municipal Engineering Design and Research Institute Co.Ltd, Liaoning, 110168, China*

Abstract. To predict the building energy consumption data more effectively and scientifically, an energy consumption prediction model based on data mining and clustering analysis is proposed. We first apply data mining to the benchmark evaluation of building energy consumption and propose the process of building energy consumption benchmark evaluation. The energy consumption monitoring model identifies building operating energy consumption patterns through clustering, which mines and matches real-time collected energy consumption data. Then, k-means algorithm is adopted to extract typical daily energy patterns, and the cumulative frequency distribution method is also used to determine the energy consumption baseline value for each type of building. The implementation of the proposed data preprocessing system and method in the case analysis verify that the scheme can accurately identify the relevant factors affecting energy consumption. It improves the accuracy of building energy consumption prediction and has high operational efficiency, thus providing managers with auxiliary decision-making for building energy conservation.

Keywords. building energy; prediction model; k-means; data mining; clustering analysis

1. Introduction

In recent years, Building Automation Systems (BAS) and Building Energy Management Systems (BEMS) have developed rapidly and are widely used, storing a large amount of building energy consumption monitoring data. These data are the most direct and primitive carriers of the actual operation status of buildings, their systems, and equipment, and their data value is self-evident. Data mining on the actual operation data of buildings can analyze system energy consumption and operation status from the data, providing technical support for optimizing control, building energy efficiency, and improving building performance. The monitoring data of building energy consumption itself has the characteristics of large quantity, multi-dimensional nature, and multi measurement. At the same time, the complexity of the structure of buildings and their systems, the diversity of equipment, and the faults of instruments and meters themselves make the actual operating data of buildings extremely susceptible to

[1] Corresponding Author: Yunli ZHANG; School of Municipal and Environmental Engineering, Shenyang Urban Construction University, Shenyang, 110168, China; sycj_zyl@126.com

interference from outliers, missing values, and inconsistent data. Low quality data will lead to low quality mining results. Therefore, in order to obtain high-quality data mining results and guide practice, it is necessary to process the data reasonably [1].

The shortcomings of traditional building energy consumption data analysis methods can be addressed through energy consumption analysis methods based on data mining technology. The clustering algorithm based on data mining not only classifies energy consumption values, but also selects meaningful feature attributes for more accurate classification of different energy consumption patterns. Outlier analysis identifies abnormal energy consumption data through the numerical values of abnormal factors, and determines whether the energy consumption data belongs to distorted or faulty data. Correlation analysis selects important influencing factors of energy consumption data by calculating the correlation between energy consumption data and its correlation factors. Previous studies have proposed new integration methods for multiple data mining techniques to conduct in-depth analysis of building energy consumption data and extract the knowledge contained in energy consumption data [2,3]. Use clustering algorithms to identify energy consumption patterns and classify energy consumption data based on these patterns. Using the energy consumption pattern labels and time generated by clustering analysis as attributes, establish an energy consumption discrimination decision tree. When identifying abnormal energy consumption data, pattern discrimination is first performed on the energy consumption data, and then outlier analysis is used to determine whether the data is abnormal. Using correlation analysis method to identify important factors that affect energy consumption, and using these factors as training attributes in energy consumption prediction, establish an energy consumption prediction model. Predict future energy consumption through predictive models.

This article first discusses the application of data mining technology in building energy consumption prediction, and conducts in-depth research on the general process and modeling methods of building energy consumption benchmark evaluation. Then, based on practical applications, a data mining model for building energy consumption analysis was established, providing the basic steps for data collection and organization, data analysis and preprocessing. In response to the uncertainty of the benchmark value of building energy consumption, cumulative frequency is used to predict the energy consumption level of the building under different conditions and compared with the benchmark value. Finally, building residents are used as the data object for clustering analysis, and the training and testing sets are divided, as well as the selection and optimization of model input variables. Train building household data using the selected clustering algorithm and generate clustering results. The experimental analysis results show that the prediction model proposed in this paper has a small deviation from the actual values, and is less affected by parameters compared to similar algorithms, with stable comprehensive performance.

2. The Basic Method of Building Energy Consumption Prediction Based on Data Mining

2.1 The Process of Building Energy Consumption Benchmark Evaluation

Traditional building energy consumption data analysis is based on a data model analysis method, which has the following shortcomings:

1) The traditional classification of building energy consumption data uses the method of sub item measurement statistics to classify different energy consumption patterns in buildings, which cannot provide a means of verifying results and has low accuracy.

2) Traditional analysis of abnormal points in building energy consumption data identifies and alerts abnormal data by setting thresholds, which cannot determine whether energy consumption data exceeding the threshold belongs to distorted or faulty data.

3) The traditional method of analyzing building energy consumption data can only analyze the energy consumption data itself, without utilizing the relevant factors that affect energy consumption, and cannot analyze the potential relationship between energy consumption data and related factors.

4) The traditional building energy consumption prediction method based on statistics predicts future energy consumption through historical energy consumption data, without considering relevant factors that affect energy consumption, resulting in low accuracy.

To apply data mining algorithms for building energy consumption benchmark evaluation, it is necessary to process the source data to conform to the format required by the data mining algorithms. Data preprocessing and data mining are the two most important steps: data preprocessing is essentially the discretization of numerical attributes, and clustering algorithms can be used for data mining work. The target data is a two-dimensional relational table containing basic building information, and these attributes are all continuous values. However, general clustering algorithms can only classify and mine discrete data, so data preprocessing, namely discretization of numerical attributes, is necessary. In the benchmark evaluation model mentioned above, the preprocessed data is also a two-dimensional relationship table that only contains discrete attributes, allowing for the next step of data mining using decision tree algorithms. After algorithm classification, we obtain some classification rules that are applied to the target building to be evaluated for benchmark evaluation. The benchmark evaluation results can be obtained, such as normal energy consumption, high energy consumption, or low energy consumption [4].

2.2 Application of Data Mining and Cluster Analysis in Building Energy Consumption Analysis

The basic process of data mining includes the following stages [5]:

- Problem definition: At this stage, we need to clarify the data mining goals and problems. Determine what kind of information or problem we would like to obtain from the data.
- Data preparation: we need to collect and prepare data for data mining. This includes steps such as data acquisition, cleaning, integration, and transformation. Ensure the quality and completeness of data for subsequent analysis and modeling.
- Data analysis: use various data analysis techniques and methods to explore the characteristics and relationships of the data. This includes techniques such as descriptive statistics, data visualization, association rule mining, clustering analysis, classification, and prediction.
- Model establishment: select appropriate modeling techniques based on the needs of the problem and use training data to build the model. This can include machine

learning algorithms, neural networks, decision trees, etc. By establishing a model, we can extract patterns and patterns from the data.

- Model evaluation and application: evaluate the performance and accuracy of the model. Use test datasets to validate the predictive ability of the model, and adjust and improve the model based on the evaluation results. Once the model achieves satisfactory performance, we can apply it to practical problems and gain insights and value from it.

Usually, visualization and knowledge representation techniques are used to convert mining results into a form that is easy for users to understand. According to different application fields and user needs, different data mining techniques can be used for data processing. Figure 1 shows the basic process of data mining application in the construction field [6,7].

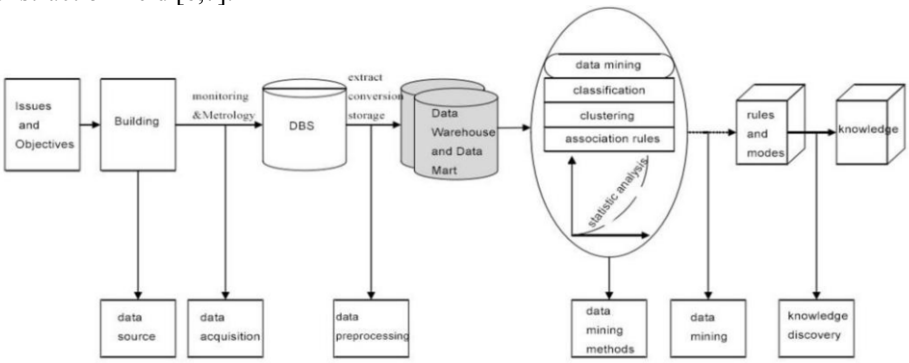

Figure 1. The basic process of data mining application in the construction field.

3. Energy Consumption Prediction Model Based on Data Mining and Clustering Analysis

3.1 Data Acquisition and Preprocessing

The data required for building energy consumption prediction models can be divided into the following categories:

- Building attribute data: These data describe the physical properties and characteristics of a building, including building type, building area, orientation, building structure, insulation materials, window type and area, etc. These attribute data can help the model understand the energy consumption characteristics and potential influencing factors of buildings.
- Meteorological data: Meteorological data is an important external factor in building energy consumption prediction models. These data include temperature, humidity, wind speed, sunshine hours, etc. Meteorological data can help models consider the impact of climate conditions on building energy consumption.
- Energy consumption historical data: Building energy consumption historical data is the key to training building energy consumption prediction models. These data include the actual energy consumption of the building, which can be hourly, daily, or monthly energy consumption data. Historical energy consumption data can help models learn patterns and trends in building energy consumption.

- Time data: Time data is an important factor in building energy consumption prediction models, including date, time, season, etc. Time data can help the model consider time related energy consumption changes, such as daily energy consumption patterns, differences between working and non working days, etc.

In addition to data directly used for modeling, through on-site research, some information about the building itself and energy supply systems can also be learned, which is helpful for comparative analysis between buildings and subsequent error analysis [8].

(1) Missing value analysis

Calculate the number and proportion of missing values in each variable. This can help you understand the distribution of missing values in the dataset and determine which variables are affected by significant missing values. By observing the patterns of missing values, we can understand whether missing values have a specific pattern or association. To use this intuitive method, the formula for defining similarity is

$$\mathrm{d}(x_i, y_i) = \sqrt{\sum_{i=1}^{n} \delta_i(x_i, y_i)} \tag{1}$$

$$\delta(v_1, v_2) = \begin{cases} 1, & \text{Discrete value } and\ v_1 \neq v_2 \\ 0, & \text{Discrete value } and\ v_1 \neq v_2 \\ (v_1 - v_2)^2, & \text{continuous value} \end{cases} \tag{2}$$

Then we select 30 similar samples and fill in the missing values using the first method.

（2）Analysis of outliers

Anomaly analysis refers to the process of analyzing and processing outliers in a dataset. The handling of outliers depends on the specific application scenario and the goal of data analysis. Sometimes, outliers may be real and meaningful data, reflecting special or abnormal situations. This article first conducts a simple statistical analysis of variables when dealing with outliers, and then examines which data is unreasonable. Then, by observing outliers in the box plot, potential outliers are identified, and finally, outliers are removed.

（3）Data protocol

The goal of data protocols is to reduce the amount of data while preserving as much important information and features as possible. Choosing an appropriate data specification method depends on the specific application scenario and the goal of data analysis. In order to eliminate the impact of differences in dimensions and value ranges between indicators, this data is standardized as follows:

$$\mathrm{x}^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{3}$$

（4）Statistical analysis and correlation analysis

Statistical analysis and correlation analysis are commonly used data analysis methods that combine domain knowledge and practical background to explain the results and explore the relationships and interactions between data. This article conducts mean and standard deviation analysis on the data, removing variables with small changes in standard deviation. Then, Pearson correlation is used to analyze the features, and based on the correlation matrix and visualization results, the correlation between the features is explained and redundant features are removed.

3.2 Benchmark value determination

The benchmark value of building energy consumption is a reference value used to measure the energy efficiency and energy-saving level of buildings. The benchmark value of building energy consumption is usually determined based on factors such as building type, area, and usage purpose. By comparing with the benchmark value of building energy consumption, the energy efficiency and energy-saving potential of buildings can be evaluated. The overall process can be described as the figure 2.

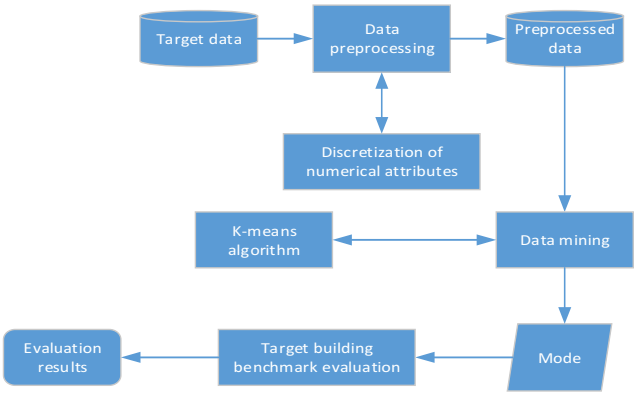

Figure 2. Building energy consumption benchmark evaluation flowchart.

If the energy consumption of a building is lower than the benchmark value, it indicates that the energy management and energy-saving measures of the building are good. On the contrary, if the energy consumption of a building is higher than the benchmark value, corresponding energy-saving measures may need to be taken to improve energy efficiency; Collect energy consumption data: Firstly, collect a representative set of building energy consumption data. These data can be historical energy consumption data or energy consumption data from similar buildings. Then sort the collected energy consumption data in ascending order, and calculate the cumulative frequency corresponding to each energy consumption value based on the sorted energy consumption data. Draw a cumulative frequency curve and determine an appropriate energy consumption reference value based on the cumulative frequency curve. This article uses 50% as the benchmark level to evaluate the energy-saving potential of buildings; 25% is the target level, indicating the expected energy consumption level of the building.

3.3 Clustering Analysis

In our study, building residents are used as the data object for clustering analysis. By determining the similarity/dissimilarity of each feature, they are divided into several categories or clusters, so that the characteristics of building residents in the same category were as similar as possible, while the similarity between different categories was as small as possible. We choose k-means algorithm to analyze the building residents. The principle theory is: assuming there are n objects $x_i\,(i=1,2,...,n)$,

$x_i = (x_{i1}, x_{i2}, ..., x_{im})$ in the dataset X, where x_{ij} is the feature parameter of x_i. Through Euclidean distance function X is divided into k classes, that is, k true subset $C_1, C_2, ..., C_k$ which makes $C_i \subseteq X$ and $C_i \cap C_j$, $1 \le i, j \le k$. Then the detailed procedures of k-means can be described as follows:

Step 1: choose any k points as the original clustering centers as $C_1, C_2, ..., C_k$;

Step 2: compute the Euclidean distance between n objects and each clustering center. Assign them to the nearest center to form k clusters as

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{im} - x_{jm})^2} \tag{4}$$

$$E_{\min} = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, c_i)^2 \tag{5}$$

4. Case Analysis

The factors that affect building energy efficiency can be summarized in three aspects: (1) environmental factors; (2) The factors of the building envelope structure itself; (3) The factors of equipment, climate, and environment. As the building structure has been determined, only the impact of external climate conditions and working conditions on building energy consumption is considered. The main external climate factors that affect building energy consumption include temperature, humidity, wind speed, sunlight, and precipitation. We clean and preprocess the collected data, including handling missing values, outliers, and duplicate values, to ensure the accuracy and completeness of the data. Calculate the correlation between various factors and building energy consumption using the correlation analysis method. Evaluate the established model, including evaluating its degree of fit, prediction accuracy, and stability. After processing the average value of data correlation, the correlation between various factors displayed in the system and building energy consumption is shown in Figure 3.

Figure 3. The relationship between energy consumption related factors and energy consumption data.

This article selects three office buildings with relatively complete data and high quality as case buildings to validate the evaluation method of this study. Assuming that the complex building has 1 underground floor and 24 above ground floors, with a building height of 99.9 meters and a total construction area of 70000 square meters; The upper structure consists of three 19-20 story towers, with a maximum height of 81.5 meters, including buildings 1 and 2 with three story podiums; Except for

underground garages and equipment rooms, all other parts are covered by air conditioning; The project adopts steel reinforced concrete embedded rock foundation and reinforced steel reinforced concrete columns. The structure of the reinforced columns and beam nodes is complex, and the post tensioned prestressed concrete multi-span continuous beam technology is used. Collect electricity consumption data files for each building, including electricity consumption, timestamp, and other information. Preprocess the data, including removing outliers, handling missing values, and so on; Use the k-means algorithm to cluster the electricity consumption data of buildings and divide them into k clusters. Based on the clustering results, the center point of each cluster can be extracted as a typical daily electricity consumption pattern. These center points represent the average daily electricity consumption pattern of each cluster, as shown in Figure 4. By predicting the new building electricity consumption data and assigning it to the corresponding energy consumption level clusters, the optimal number of clusters for the daily electricity consumption curves of the three case buildings can be determined, namely, the number of typical daily electricity consumption modes is 5, 4, and 4, respectively.

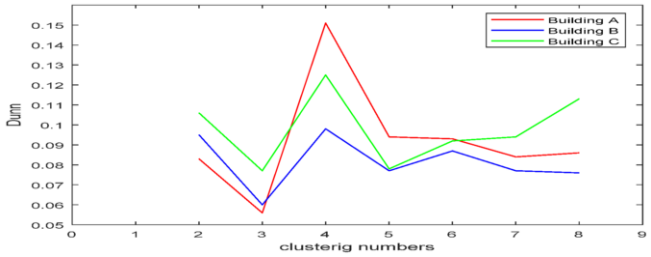

Figure 4. Clustering validation index with different buildings

For each reconstructed data, we adopt three strategies to separately establish a prediction model: the first scenario is to reconstruct the electricity consumption of all users in a time series within seven days; The second scenario is to extract electricity consumption data from a total of 8593 users in all buildings, and use cumulative summation to obtain the time series of electricity consumption; The third method is to collect daily electricity consumption data for each user and aggregate it into a time series. Use time series analysis technology to predict the electricity consumption time series of each group, and finally add up the predicted results of each group. The members of the prediction model include linear regression (LR) model, support vector machine (SVR) model, and K-means model [9,10]. Through the K-Means algorithm, we can obtain the grouping results of users, with each group representing users with similar electricity consumption behavior. In this way, we can further analyze the user characteristics of each group, understand their electricity usage habits and behavior patterns, and provide personalized electricity advice or services for users. Tables 1, 2, and 3 respectively show the prediction indicators of these three algorithms under different power consumption strategies. The experimental results show that the prediction results of the third scenario are basically consistent with those of the second scenario. The K-means method has a closer relationship between its indicators and the model, and can reflect the specific weights of influencing factors, which helps to efficiently and accurately group users based on their electricity consumption behavior, thereby reducing building energy consumption costs.

Table 1. Comparison of prediction performance of three different models under the 1st strategy

Index and model	LR	SVR	K-means
MAE1	2.601552	2.145529	0.454169
MSE1	7.993549	5.650326	5.556410
SMAPE1	0.650889	0.942361	0.488358

Table 2. Comparison of prediction performance of three different models under the 2nd strategy

Index and model	LR	SVR	K-means
MAE2	1.014328	3.335850	0.278469
MSE2	2.054568	9.687119	1.756329
SMAPE2	0.185699	0.353889	0.156564

Table 3. Comparison of prediction performance of three different models under the 3rd strategy

Index and model	LR	SVR	K-means
MAE3	1.098553	0.65189	0.223598
MSE3	3.123654	1.95331	1.658798
SMAPE3	0.283326	0.17006	0.132965

5. Conclusion

Based on the current situation of insufficient energy consumption management and high potential for building energy conservation in China, this paper summarizes the research status of prediction methods for building energy consumption both domestically and internationally Building energy consumption prediction is a complex problem that involves multiple factors such as the structure, materials, and equipment of buildings. Therefore, it is more feasible and accurate to comprehensively apply multiple methods and technologies and combine them with actual situations to predict building energy consumption. This article predicts energy consumption by mining and analyzing historical measurement data, and uses data mining algorithms to predict new building data using trained models to obtain energy consumption prediction results for buildings. We have conducted in-depth research and improvement on the application scenarios, advantages and disadvantages of energy consumption prediction modeling under K-means. In empirical analysis, we have taken the calculation of user electricity consumption as a case, and obtained the final energy consumption prediction value based on the prediction process proposed in this article. The experimental results show that the proposed scheme does not require too much prior knowledge or complex parameter adjustments, and has low computational complexity, which can be quickly applied to building energy consumption prediction problems. Compared with similar algorithms, it has better comprehensive performance in different situations.

Acknowledgement

This project was supported by Scientific Research Project of Liaoning Provincial Department of Education, under grant no.LJKX202009.

References

[1] Zhiguo C, Yong C, Genfeng W U, et al. Research on Preprocessing Technology of Building Energy Consumption Monitoring Data Based on Machine Learning Algorithm. Building Science, 2018, 34(2): 94-99.

[2] Zhichao S, Bo W. AN INTEGRATION METHOD OF BUILDING ENERGY CONSUMPTION ANALYSIS BASED ON DATA MINING ALGORITHMS. Computer Applications and Software, 2017, 34(11): 103-108.
[3] Bogunovic D. Integrated data environment for analysis and control of energy consumption (IDE-ACE) in surface coal mining. Dissertations & Theses - Gradworks, 2008, 25(1):26-26.
[4] HAN Lian-hua, MAO Guo-iun, Sun Xiao-xi. Research on Building Energy Consumption Benchmarking Based on Data Mining. Computer Science, 2008, 35(10): 209-218.
[5] Song N X, Wan D M, Sun Q, et al. Data Mining-Based Smart Industrial Park Energy Efficiency Management System. Applied Mechanics & Materials, 2014, 484-485:585-588.
[6] Zhao, Deyin, et al. Energy consumption predicting model of VRV (Variable refrigerant volume) system in office buildings based on data mining. Energy, 2016, 102:660-668.
[7] Yu, Jiguo, et al. A Local Energy Consumption Prediction-Based Clustering Protocol for Wireless Sensor Networks. Sensors, 2014, 14(12): 17-40.
[8] Aouadj W, Abdessemed M R, Seghir R. A Reliable Behavioral Model: Optimizing Energy Consumption and Object Clustering Quality by Naive Robots. International journal of swarm intelligence research, 2021, 4:12.
[9] Poudel S, Moh S, Shen J. Residual energy-based clustering in UAV-aided wireless sensor networks for surveillance and monitoring applications. Journal of Surveillance, Security and Safety, 2021, 2(3):103-116.
[10] Sharad H V, Desai S R, Krishnrao K Y. Energy-aware multipath routing in WSN using improved invasive weed elephant herd optimization. International Journal of Pervasive Computing and Communications, 2023, 19(3):451-474.