Energy 263 (2023) 125575

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

A dynamic anomaly detection method of building energy consumption based on data mining technology

Lei Lei [a], Bing Wu [b], Xin Fang [c], Li Chen [c], Hao Wu [c], Wei Liu [d,*]

[a] School of Civil Engineering and Architecture, Zhejiang Sci-Tech University, Hangzhou, 310018, China
[b] College of Civil Engineering and Architecture, Guangxi Vocational and Technical College of Communications, 1258 Kunlun Avenue, Nanning, 530216, China
[c] Alibaba Cloud, Alibaba Group, 969 West Wen Yi Road, Hangzhou, 311121, China
[d] Division of Sustainable Buildings, Department of Civil and Architectural Engineering, KTH Royal Institute of Technology, Brinellvägen 23, Stockholm, 10044, Sweden

ARTICLE INFO

Keywords:
Building energy consumption
Dynamic anomaly detection
Semi-supervised algorithm
Particle swarm optimization
K-medoids algorithm
KNN algorithm

ABSTRACT

Due to the equipment failure and inappropriate operation strategy, it is often difficult to achieve energy-efficient building. Anomaly detection of building energy consumption is one of the important approaches to improve building energy-saving. The great amounts of energy consumption data collected by building energy monitoring platforms (BEMS) provides potentials in using data mining technology for anomaly detection. This study proposes a dynamic anomaly detection algorithm for building energy consumption data, which realizes the dynamic detection of point anomalies and collective anomalies. The algorithm integrates unsupervised clustering algorithm with supervised algorithm to establish a semi-supervised matching mechanism, which avoids the influence of error label and improves the efficiency of anomaly detection. A particle swarm optimization (PSO) is used to optimize the unsupervised clustering algorithm. This investigation tests the effectiveness of the proposed algorithm and evaluates the performance of the energy consumption clustering algorithm by using the annual electricity consumption data of an experimental building in a university. The results show that the clustering accuracy of the algorithm can reach more than 80%, and it can effectively detect the building energy consumption data of two different forms of outliers. It can provide reliable data support for adjusting building management strategies.

1. Introduction

As the demand for residential and commercial buildings continues to expand globally, the corresponding building energy consumption is steadily increasing. In recent years, the average annual growth rate of China's civil building energy consumption is 11.82%, accounting for about 20% of the country's total energy consumption [1]. According to the data of Eurostat in 2021, the energy consumption of residential buildings accounted for about 26.3% of the total energy consumption in Europe [2]. According to the statistics of energy consumption data in 2020 released by the US Department of Energy (DOE), the energy consumption of residential buildings accounted for 22.38% of the total energy consumption [3]. The operation of a building is influenced by many factors such as climate, structure, materials, space utilization, lighting system and the operation of HVAC (heating, ventilating, and air-conditioning) system and inappropriate operation would result in a loss of about 16% of building energy consumption [4–6]. Therefore, it is very important to improve the efficiency of building operations for reducing the building energy consumption.

At present, one of the most effective strategies to mitigate the energy wasting during building operations is the anomaly detection of energy consumption [7–9]. Anomaly detection can also be called fault detection and diagnosis (FDD), which refers to the process of detecting data that do not conform to the expected pattern [10]. It can accurately identify samples or sample sets with large deviations (outlier) in data sets. These outliers can be classified as point anomalies or collective anomalies [10]. In a data set with the same characteristics (such as similar weather and electricity consumption, etc.), point anomaly refers to an energy consumption data that differs a lot with the others. The collective anomaly is a certain set of data that is abnormal compared with the rest of the data set, but there is no point anomaly [11]. It is also referred to the anomaly mode. Accurate anomaly detection provides data support for energy-saving measures, timely repair for equipment failures, and adjusting operation strategies, which finally reduces unnecessary

* Corresponding author.
E-mail address: weiliu2@kth.se (W. Liu).

https://doi.org/10.1016/j.energy.2022.125575
Received 10 May 2022; Received in revised form 20 August 2022; Accepted 24 September 2022
Available online 4 October 2022

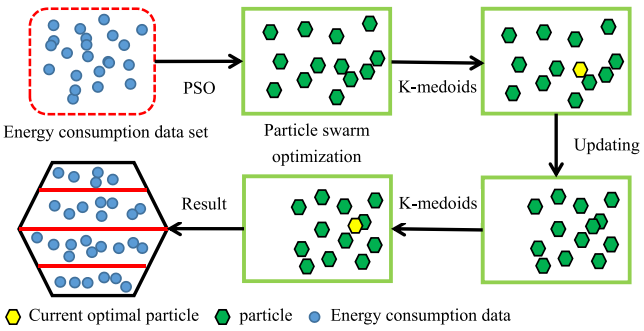


Fig. 1. Optimization process of the K-medoids algorithm by PSO.

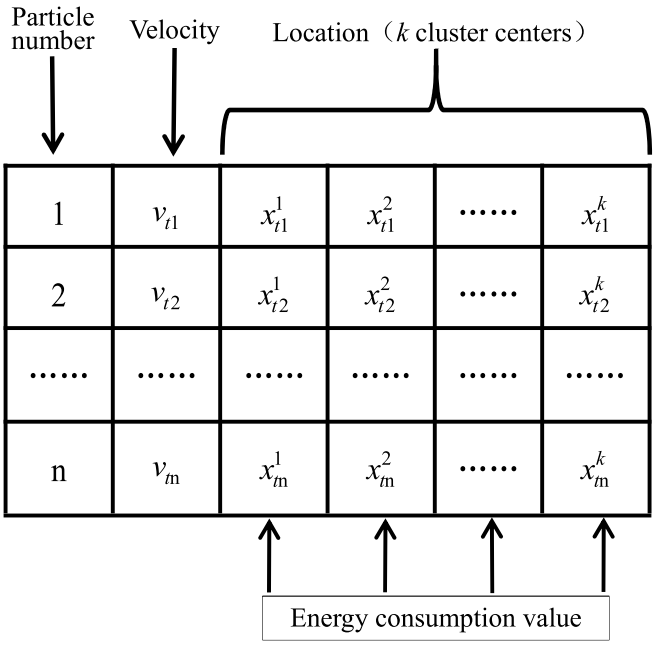


Fig. 2. Initial particle swarm.

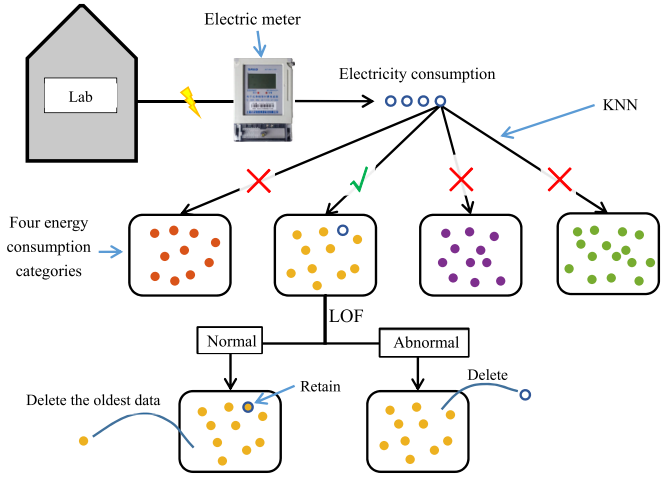


Fig. 3. Schematic diagram of point anomaly detection.

resource loss and effectively improves energy utilization.

There are many building energy consumption monitoring platforms (BECMP), internet of building energy systems (IBES), and building energy management systems (BEMS) that collect and store a large amount of energy consumption data [12]. However, due to the fact that data recording equipment and transmission channel are affected by malfunctions and weather, as well as the incomplete energy consumption monitoring technology platform [13], the detection of abnormal energy consumption data is not always accurate. To reduce resource loss, it is necessary to detect anomalies from the data and fix the faults behind to realize energy-saving [14,15]. Existing anomaly detection methods can be divided into supervised learning and unsupervised learning according to the types of data mining methods used in the analysis process [9,16].

Supervised learning refers to learning from training samples with conceptual markers. For example, decision tree, linear regression, support vector machines (SVM), and artificial neural network (ANN), etc. are used to establish the model [17]. In the current supervised learning, most FDD methods compare the predicted value with the measured value by residual analysis for anomaly detection [18–20]. Wang et al. [21] adopted a neural network model-based FDD strategy for outdoor and supply flow sensors to diagnose anomalies by quantifying the residual between the predicted value and the measured one. Residual analysis can effectively detect the anomaly of energy consumption data, but the detection may be inaccurate when the residual is small. Besides, the prediction model may be inaccurate if it does not consider all the influential factors such as weather or human factors in a real situation. There are also have classification models to find the features of abnormal energy consumption and use feature labels to describe whether the energy consumption data are abnormal or not [22–24]. Zhao et al. [25] proposed a fault detection method based on pattern recognition for chiller, which transformed the fault detection problem into a data description problem. A minimum-volume hypersphere in a high-dimensional space was used to describe the distribution of fault-free data, and faulty data is the outliers of the hypersphere. However, a major limitation of these methods is that the feasibility depends on the accuracy of feature labels and the effectiveness of classification methods.

Unsupervised learning refers to learning the data points without conceptual markers and discover the structural knowledge. This method does not need to know the label information of energy consumption data in advance, so it is more aligned with the actual needs of building managers [9]. Clustering algorithms and association rules are typical unsupervised learning algorithms [15,26]. For example, Li et al. [27] adopted these two algorithms to identify the energy consumption mode of a variable refrigerant flow (VRF) system. Experimental results show that the method can identify the energy consumption patterns in VRF systems. Ma et al. [28] proposed a variation-centered clustering analysis method to identify the typical daily heating energy consumption of 19 buildings in Norway. Unsupervised learning is independent of feature labels and is thus more suitable for the actual situation. However, due to the lack of prior knowledge of the algorithm, the clustering effect is not as perfect as that of supervised learning. At the same time, when it comes to big data processing, the detection efficiency of unsupervised learning algorithm will also be greatly reduced [9].

The previous anomaly detection of building energy consumption mostly focuses on the abnormal energy consumption mode of a certain system [13,20,25,26], or detection of static energy consumption data in the process of building operations [15,16,27,28]. There are few studies on monitoring platform for the dynamic building energy consumption data. Ma et al. [29] proposed a real-time detection method, proper orthogonal decomposition-linear stochastic estimation (POD-LSE) and fractal correlation dimension (FCD), for abnormal building energy consumption data. This method replaces the traditional residual analysis method, but it does not classify energy consumption data and does not consider the influence of energy consumption classes on the detection of point anomalies. As the static detection method of building energy consumption anomaly is prone to misjudgment in the dynamic energy consumption, Jiang et al. [30] proposed an improved anomaly detection method, which dynamically identifies the class of building energy

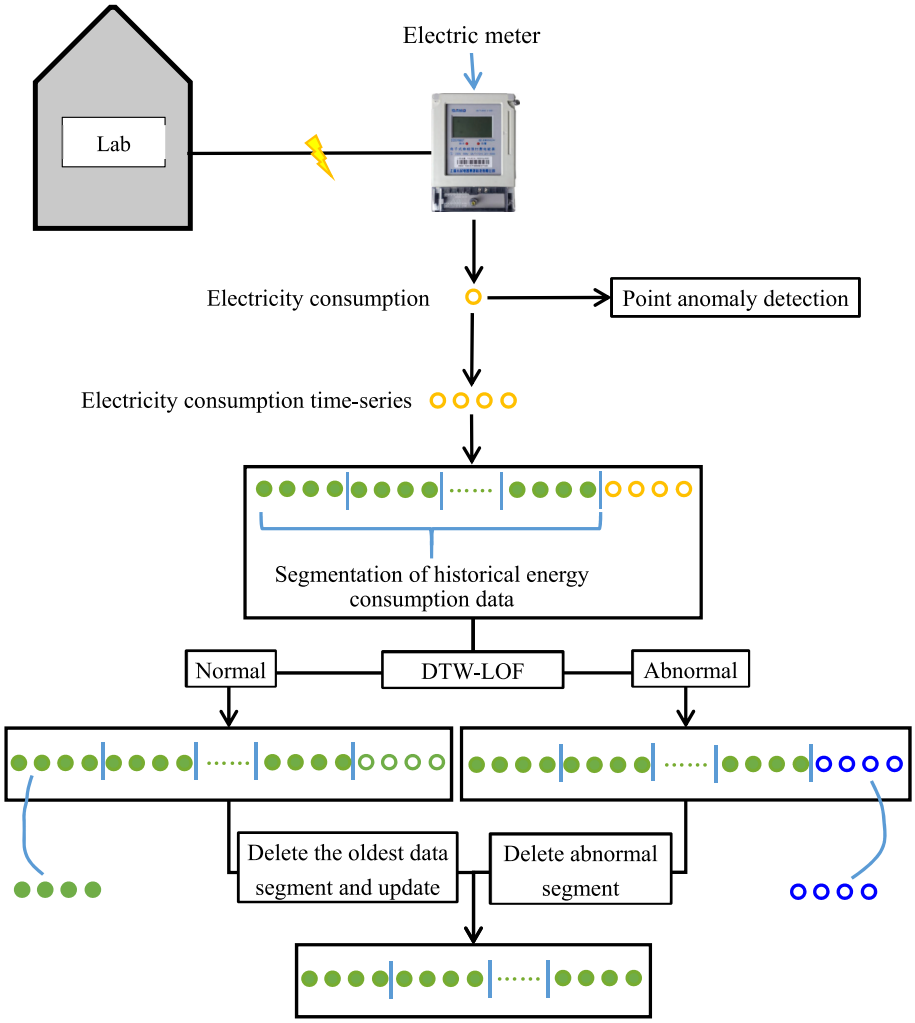


Fig. 4. Schematic diagram of a collective anomaly detection.

consumption and detects the anomaly by combining a self-adaptive density-based spatial clustering of applications with noise (SA-DBSCAN) and local outlier factor (LOF) algorithm. This method can well consider the impact of dynamic energy consumption on abnormal data detection, but the algorithm construction is complex, and the complexity also increases when the data dimension is high. The method developed in the above literature realizes the detection of point anomalies for building energy consumption data, but the possible collective anomaly of energy consumption data that has far-reaching and practical significance is not considered in the detection process. Therefore, how to simultaneously realize the point and collective anomaly detection of building energy consumption data is critical for building operators.

To improve the efficiency of dynamic analysis and calculation of energy consumption data and make the detection algorithm closer to reality, this paper proposes an algorithm combining particle swarm optimization (PSO) clustering and neighborhood detection. The algorithm is developed basing on the building energy consumption data platform coupled with supervised learning and unsupervised learning. It can dynamically identify and detect abnormal energy consumption data through local abnormal factors. By using the energy consumption data of an experimental building in a university, this investigation tests the effectiveness of the method and the performance of the building energy consumption clustering algorithm.

1.1. Research methods

The developed algorithm is divided into three steps. Firstly, to distinguish energy consumption data with different characteristics, it is necessary to cluster the input energy consumption data set. In this study, the PSO optimized clustering algorithm is used to group energy consumption data into different clusters, which have different characteristics and category labels. Secondly, since the energy consumption data generated under different conditions are not comparable, this investigation establishes a mechanism based on the clustering results to identify and match the energy consumption data input one by one in the form of time-series. Similar energy consumption data have similar characteristics such as similar periods, weather, electricity consumption, etc. The detection in the similar energy consumption data can establish an effective detection mechanism for point anomalies. Finally, with the continuous input of energy consumption data, when the time-series data reaches a certain length, the collective anomaly detection begins. The time-series data are segmented and compared to establish a collective anomaly detection mechanism for dynamic energy consumption data. Through the above three steps, a real-time dynamic anomaly detection model for building energy consumption is established.

1.2. Clustering of energy consumption data

In order to study the influence of energy consumption patterns on data anomaly detection, it is necessary to cluster energy consumption

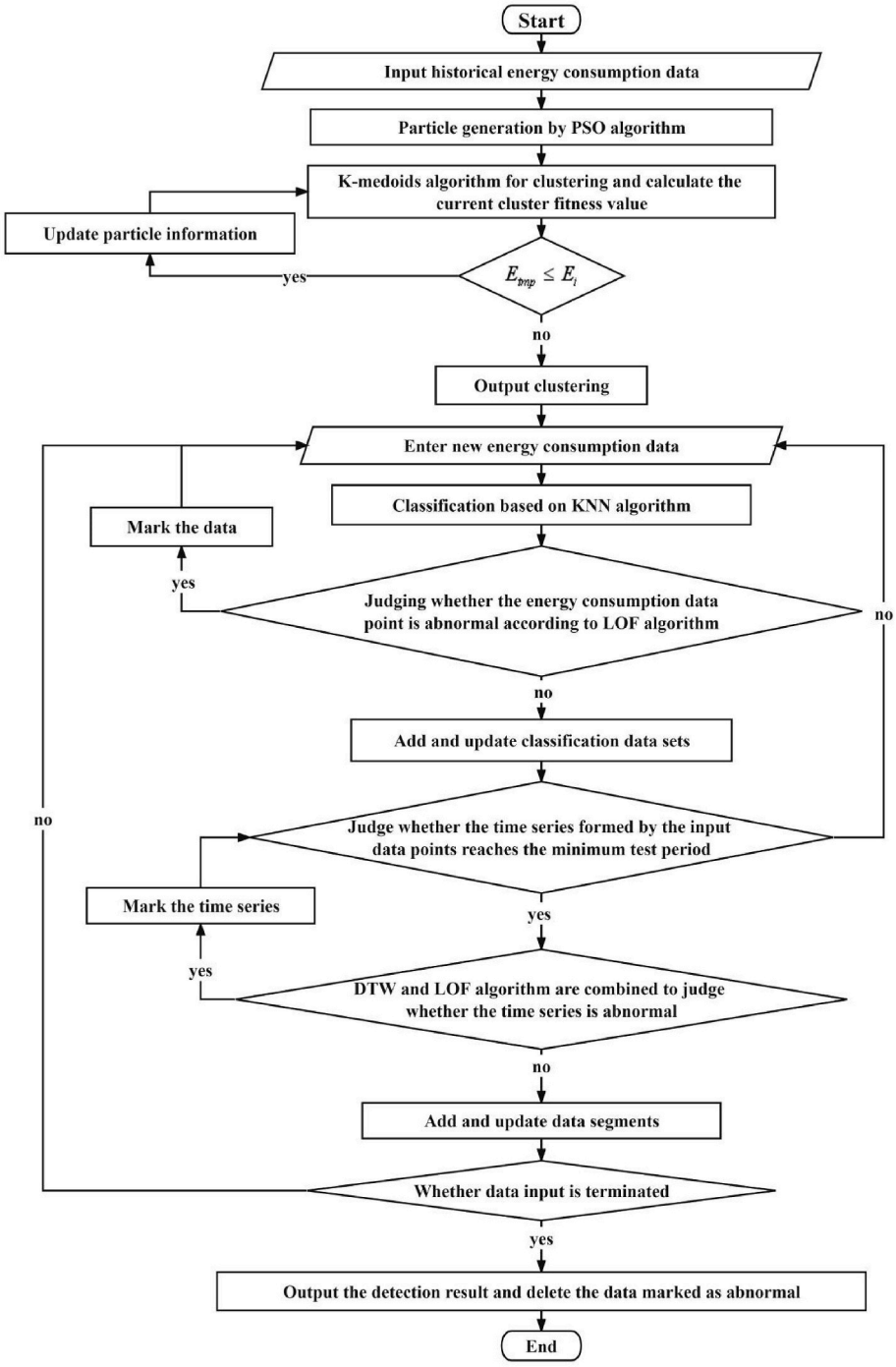


Fig. 5. Overall flow chart of anomaly detection.

data first. When the mean energy consumption is used as the clustering center, some outliers far from the center will lead to deviation between the calculated mean and the actual value. In this study, K-medoids algorithm is adopted to divide energy consumption data based on clustering centers, which need to determine the energy consumption clustering centers first.

The K-medoids algorithm has strong robustness in case of noise or outliers and generally achieves high-quality clustering results [31]. However, for a great number of energy consumption data, the computational efficiency of K-medoids algorithm will be greatly reduced. Therefore, PSO is introduced to optimize K-medoids algorithm and improve the efficiency of the anomaly detection method.

The optimization of the K-medoids algorithm by PSO is to replace the clustering center by particle swarm as shown in Fig. 1. PSO algorithm is adopted to randomly generate a certain number of initialization particles according to the energy consumption data set [32]. Each particle initialized represents a set of potential energy consumption data clustering centers, as shown in Fig. 2. Particles have different initial velocities and energy consumption values. In the process of searching, the optimal solution found by the particle itself is called the individual extremum *pbest*. The extreme value found by the particle swarm is called the global extremum *gbest*, and the corresponding particle is called the current optimal particle, representing the current optimal energy consumption cluster. Particle swarm follows the current optimal particle to search in the energy consumption data set to obtain the optimal solution. The PSO updates the cluster center and forms a new energy

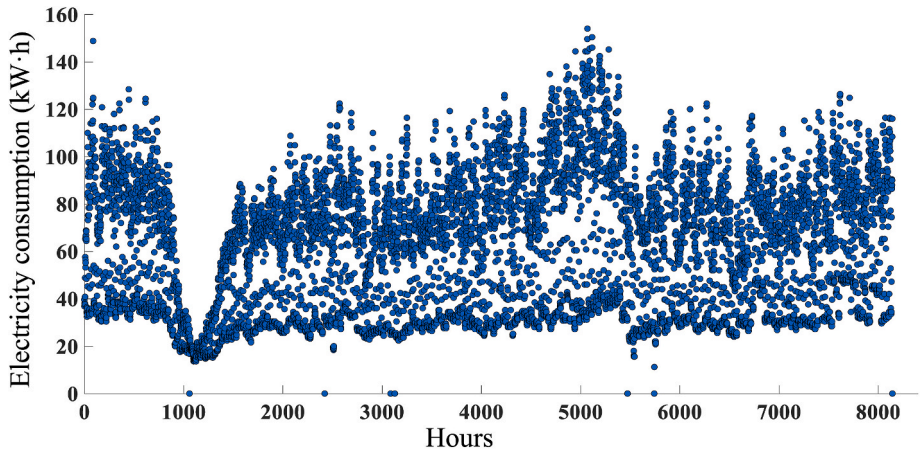


Fig. 6. Hourly electricity consumption of the experimental building.

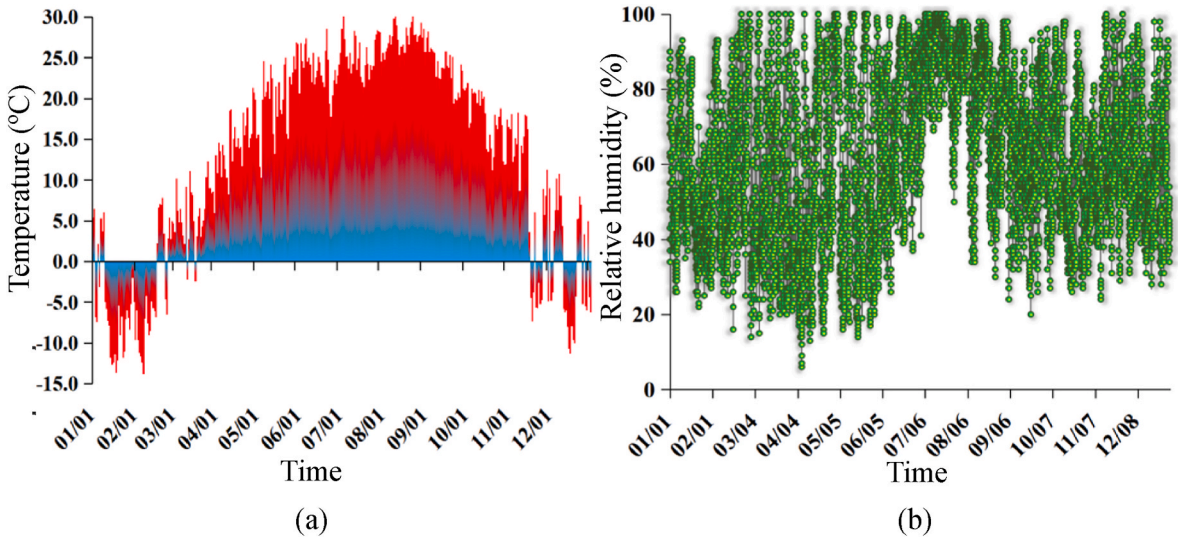


Fig. 7. Meteorological data of Dalian in 2018: (a) Hourly temperature throughout the year; (b) Hourly relative humidity.

Table 1
Characteristics of various classes of electricity consumption.

Data class	Time	Max temperature (°C)	Min Temperature (°C)	Avg Temperature (°C)	Max electricity consumption (kW·h)	Min electricity consumption (kW·h)	Avg electricity consumption (kW·h)	Number of data
Class A	10:00–11:00 14:00–17:00 19:00–20:00	18.9	−1	12.36	122.4	82	92.338	148
Class B	09:00–10:00 11:00–13:00 18:00–21:00	21.3	−0.5	11.88	81.6	61.2	71.527	219
Class C	07:00–8:00 21:00–22:00	20.7	−0.3	11.62	60.8	40.1	50.429	90
Class D	22:00–07:00	15.6	−0.3	8.92	39.6	0	29.567	263
Total								720

consumption clustering. Each clustering has a consumption cost that is calculated by k-medoids algorithm. By iteratively updating the particle speed and energy consumption value, the particle corresponding to the lowest consumption cost is finally identified.

1.3. Matching of energy consumption data

Energy consumption monitoring platforms often record energy consumption data without clear label information. The purpose of clustering energy consumption data is to obtain label information, which facilitates detection of outliers in homogeneous energy consumption data. In practice, energy consumption monitoring platform generally records energy consumption data continuously. If the data set is re-clustered every time reading in new data, the efficiency of anomaly detection will be greatly reduced. Therefore, based on the clustering results of energy consumption data in the first month of each quarter,

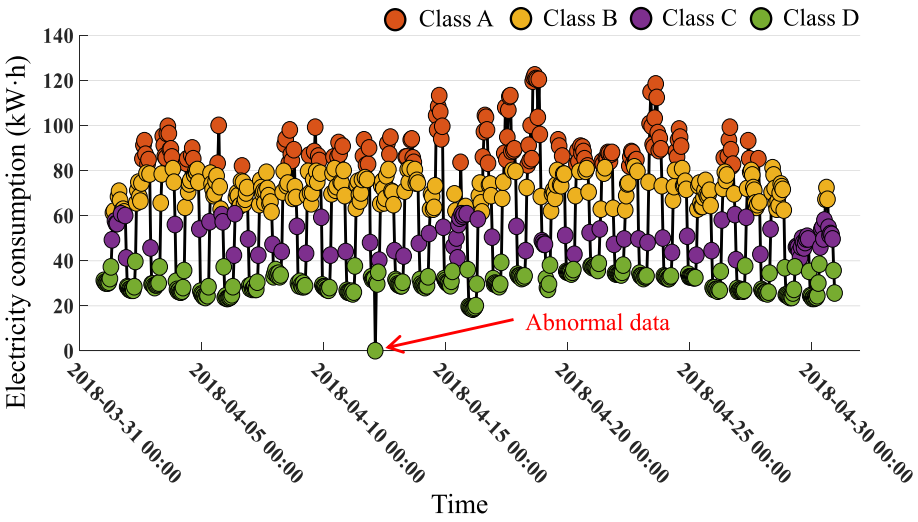


Fig. 8. Clustering results of electricity consumption in April.

Table 2
Average electricity consumption in the first month of each quarter (kW·h).

Data class	First quarter	Second quarter	Third quarter	Fourth quarter
Class A	104.628	92.338	120.927	92.518
Class B	84.812	71.527	95.471	72.714
Class C	63.449	50.429	70.788	50.016
Class D	36.988	29.567	34.363	30.592

this study adopts k-nearest neighbors (KNN) algorithm to identify the cluster that the newly read energy consumption data belongs to.

KNN algorithm is a supervised learning algorithm which matches the detection point based on distance. The advantages of KNN algorithm are high precision, insensitive to outliers, and no data input assumptions. When reading unlabeled new energy consumption data, the algorithm will find r nearest energy consumption data in the energy consumption data set. The cluster that contains most of these r data is identified. r is usually an integer not greater than 20. Energy consumption data can be matched with a cluster by Euclidean distance because it has time attribute and numerical attribute.

1.4. Anomaly detection

1.4.1. Point anomaly detection

After new energy consumption data is identified, the next step is to find abnormal energy consumption data in the cluster. LOF algorithm that is a typical high precision outlier detection method is used. For any energy consumption data in the cluster, if its neighboring energy consumption data are dense, then this energy consumption data is considered as normal. If an energy consumption data is far from its neighboring energy consumption data, the energy consumption is abnormal. The LOF algorithm calculates a LOF value for each energy consumption data, which depends on the neighborhood density of data. If the LOF value of the detected data is greater than a threshold, the energy consumption is abnormal.

First, the k-local reachability density lrd (q) is used to estimate the neighborhood density of energy consumption data q, which is defined as follows:

$$r-\mathrm{dist}(q,x)=\max(k-\mathrm{dist}(q), \tag{2}$$

$$\mathrm{lrd}(q)=\frac{1}{\sum\limits_{x\in k(q)} r-\mathrm{dist}(q,x)\Big/k}=\frac{k}{\sum\limits_{x\in k(q)} r-\mathrm{dist}(q,x)} \tag{3}$$

where $k-\mathrm{dist}(q)$ is the distance from energy consumption data q to the kth energy consumption data in the cluster and k is the number of energy consumption data falling in the kth neighborhood of energy consumption data q. The kth neighborhood of energy consumption data q refers to a circular area in the energy consumption data set, the center of which is at data q with the $k-\mathrm{dist}(q)$ as the radius. $r-\mathrm{dist}(q,x)$ is the reachable distance between the neighboring energy consumption data x and the energy consumption data q, which refers to the maximum between the $k-\mathrm{dist}(q)$ and the d (q,x). $k-\mathrm{dist}(q)$is the kth distance of energy consumption data q, and d (q,x) is the distance of the energy consumption data x to energy consumption data q. There is at least one energy consumption data located on the boundary of the kth neighborhood, but those data are counted as one data. The larger the local reachability density of data q is, the closer it is to the neighboring points and the more likely it is to belong to the same cluster, and vice versa.

According to Eqs. (2) and (3), the LOF value of energy consumption data q is defined as:

$$\mathrm{LOF}(q)=\frac{\sum\limits_{x\in k(q)}\frac{\mathrm{lrd}(x)}{\mathrm{lrd}(q)}}{k} \tag{4}$$

The numerator of Eq. (4) represents the sum of the ratio of the local reachability density of data x in the kth neighborhood of the energy consumption data q to the local reachability density of the energy consumption data q. The denominator k represents the number of data that fall in the kth neighborhood of energy consumption data q. The closer LOF(q) is to 1, the closer the point q is to its neighbor density, the more likely it is to be normal energy consumption. Referring to Jiang et al. [30] that investigated outlier detection for electricity consumption also in a campus building, the threshold for LOF value is determined to be 2. It is further confirmed in Section 4.1 by the normal data assurance rate of 95% can be reached with threshold of 2. This assurance rate is generally considered acceptable.

Once a new energy consumption data is identified for a specific cluster, anomaly detection starts. The detection process is shown in Fig. 3. If the LOF value does not exceed the threshold, the energy consumption is regarded as normal. Then the new data will replace the old data in the cluster. If it is determined to be abnormal, the dataset will not be updated. Since the energy consumption will change with time, the earlier data may not be suitable for judging the latest data. The earlier data is deleted dynamically to ensure the accuracy of anomaly detection. The updated cluster information will thus have an impact on the anomaly detection of the newly read energy consumption data. The dataset is maintained to have data points in 30 days.

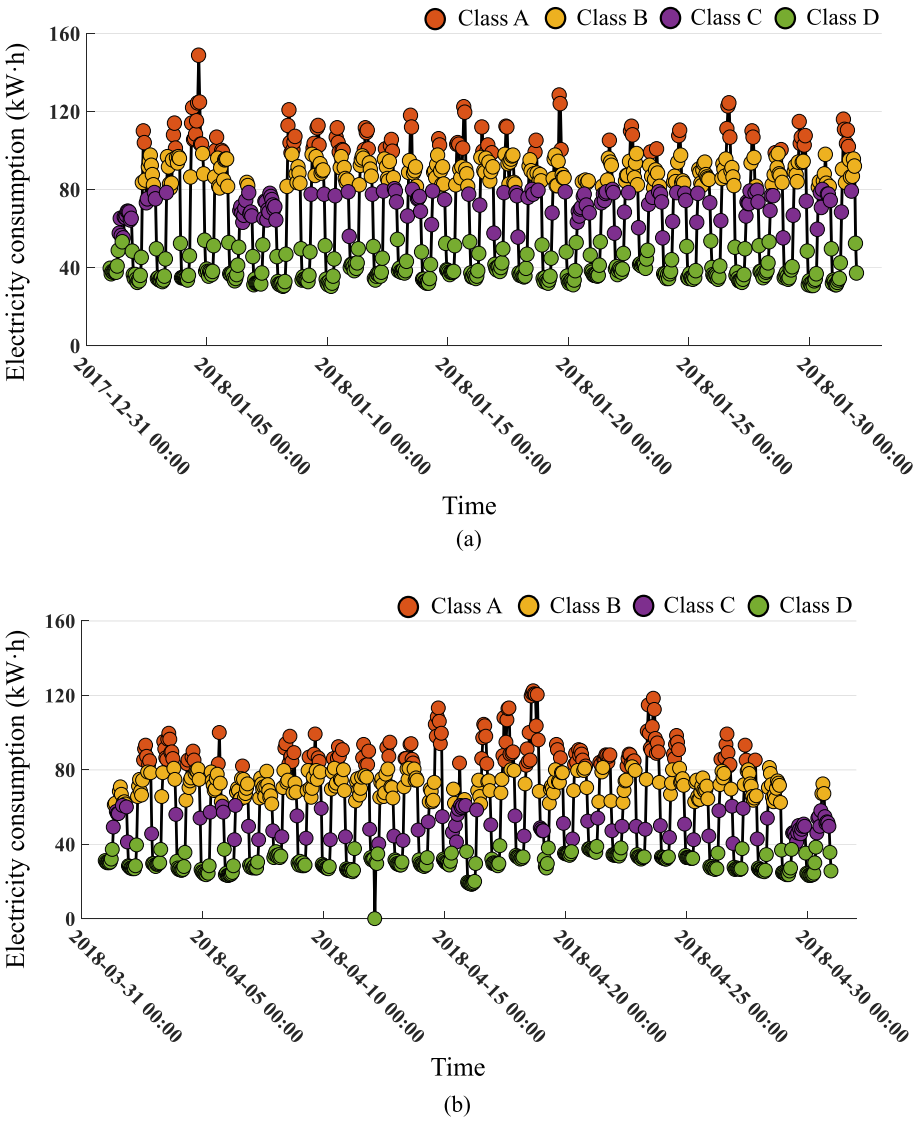


Fig. 9. Clustering results of energy consumption data in the first month of each quarter: (a) January; (b) April; (c) July; (d) October.

1.4.2. Collective anomaly detection method for energy consumption data

An energy consumption data has time attribute and numerical attribute. The energy consumption data are sorted according to the time attribute in successive periods of time. The collective anomaly is to be detected based on time-series energy consumption data instead of a single data. The energy consumption data used in this paper has obvious regularity and its minimum period is 24 h. Therefore, the energy consumption data every 24 h is used as the minimum time-series data to avoid the impact of data length.

For the single energy consumption data, Euclidean distance can be used for clustering and LOF value for anomaly detection. However, for collective anomaly detection, it is necessary to measure the similarity between time-series data to get the distance between them. The traditional Euclidean distance cannot be used to effectively solve the distance between two time-series data. In this paper, dynamic time warping (DTW) distance is used to calculate the distance between two time-series data. Then LOF algorithm is used to detect abnormal data segments.

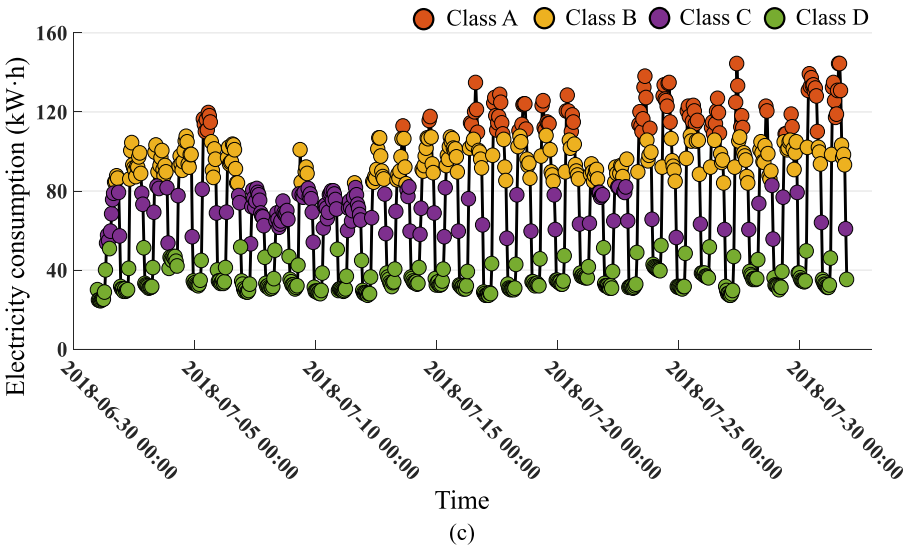
The advantage of DTW is that it allows the comparison of two time-series of unequal length and it is insensitive to the synchronization of time-series data. The DTW distance between two time-series data$\mathrm{X}\{x_1, x_2, x_3, \ldots\ldots, x_i\}$ and $\mathrm{Y}\{y_1, y_2, y_3, \ldots\ldots, y_i\}$ is defined recursively as follows:

$$D_{dtw}(\mathrm{X}, \mathrm{Y}) = d(x_1, y_1) + \min(D_{dtw}(\mathrm{X}, \mathrm{Rest}(\mathrm{Y})), \tag{5}$$

where$d(x,y) = \|x - y\|_p$,$\mathrm{Rest}(\mathrm{X}) = \{x_2, x_3, \ldots\ldots, x_i\}$,$\mathrm{Rest}(\mathrm{Y}) = \{y_2, y_3, \ldots\ldots, y_i\}$.

When energy consumption data input reaches a certain amount, collective anomaly detection is performed on the time-series data. The detection process is shown in Fig. 4. The principle is similar to that of point anomaly detection, but DTW distance instead of Euclidean distance is used for LOF calculation. Therefore, the threshold for collective anomaly detection is again 2. In calculating LOF value to determine the density of the detection point and the surrounding data points, the time-series data is treated as a special "point". If the LOF value is greater than the threshold, the time-series data is abnormal and deleted. Otherwise, it is a normal time-series data, which is retained and the earliest time-series is deleted.

The complete anomaly detection process is shown in Fig. 5. Firstly, the energy consumption data of the first month of each quarter were clustered by PSO optimized K-medoids algorithm and find energy consumption clusters. Then KNN algorithm is used to identify the cluster for energy consumption data newly read by the energy consumption



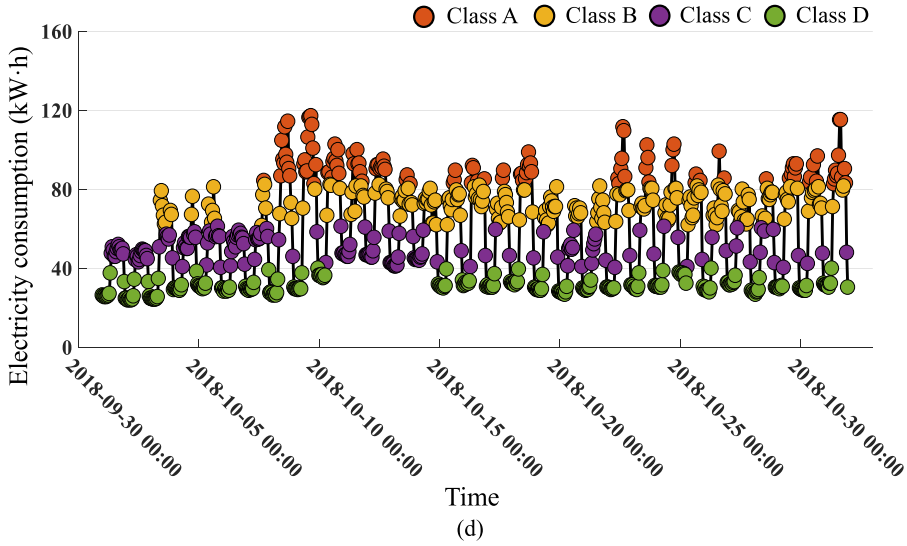


Fig. 9. (*continued*).

monitoring platform. With the cluster, the LOF algorithm is used to detect the anomaly of the new energy consumption data. If the data is determined to be normal, it is kept, and the energy consumption data of the earliest timestamp is deleted. If the data is an outlier, it is removed. With the input of energy consumption data, when the minimum time-series for collective anomaly detection is reached, it is compared with the time-series of historical energy consumption data. By using the DTW distance that measures the similarity between two time-series data, the normal time-series data is retained, and the earliest time-series data is deleted. If the time-series data is abnormal, it will be deleted. For this anomaly detection method, the real-time update of the data set is guaranteed for dynamic detection.

2. Experimental data and processing

The experimental data for building energy consumption contains two parts. One part is the cluster of data samples used to establish the real-time matching mechanism. The other part is the energy consumption data combined with the detection algorithm for anomaly detection. The processing of experimental data refers to the establishment of real-time matching mechanism of energy consumption data by cluster analysis.

2.1. Experimental data

This research collects the energy consumption data of an experimental building in a university in Dalian. The building energy consumption is mainly from heating, ventilation, air conditioning, lighting, and electrical appliances, and most of this energy consumption is in the form of electric energy. The proposed method is expected to be applicable to other sources of energy as long as there is data acquisition. For collective anomaly detection, the length of the time-series data should be identified by analyzing the characteristic of the energy data.

The experimental building is a frame-shear structure, with five floors above ground and one underground, with a total construction area of 9533 m^2 and an air conditioning area of 8580 m^2. A Variable Refrigerant Volume (VRV) system is used for cooling in summer and central heating used in winter. The electricity consumption data was recorded hourly throughout the year in 2018, from 00:00 on January 1st, 2018 to 24:00 on December 7th, 2018, as shown in Fig. 6. The electricity consumption fluctuated between 20 kW h and 160 kW h. At about 1000 h, which was Spring Festival, the university reduced the number of staff on campus and further the use of electrical equipment, making it the period with the lowest electricity consumption. Around 5000 h, which was summer, the temperature was high and the demand for cooling is large, making the

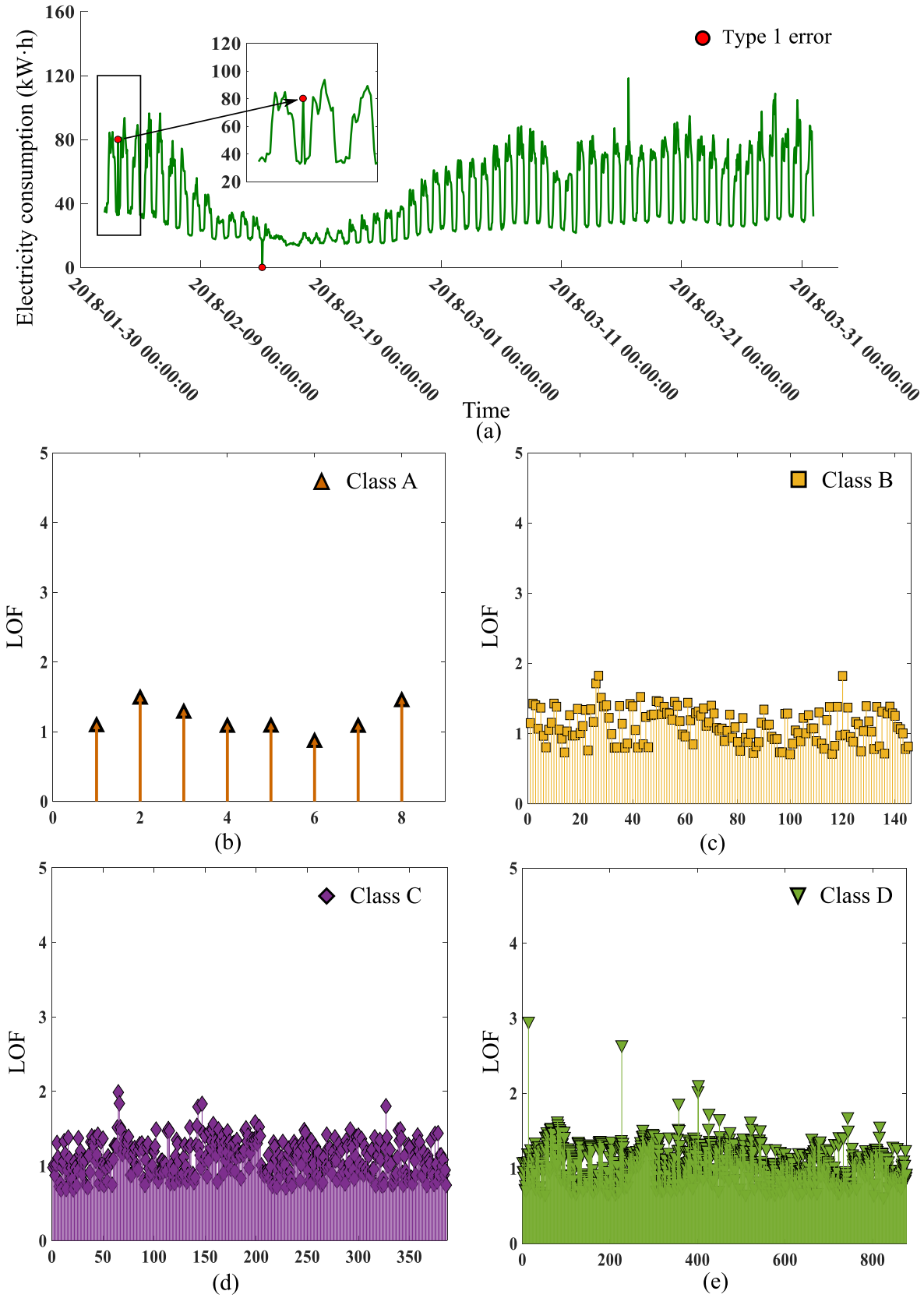


Fig. 10. Point anomaly detection results and LOF values for the first quarter: (a) Point anomaly detection results; (b) LOF values of Class A; (c) LOF values of Class B; (d) LOF values of Class C; (e) LOF values of Class D.

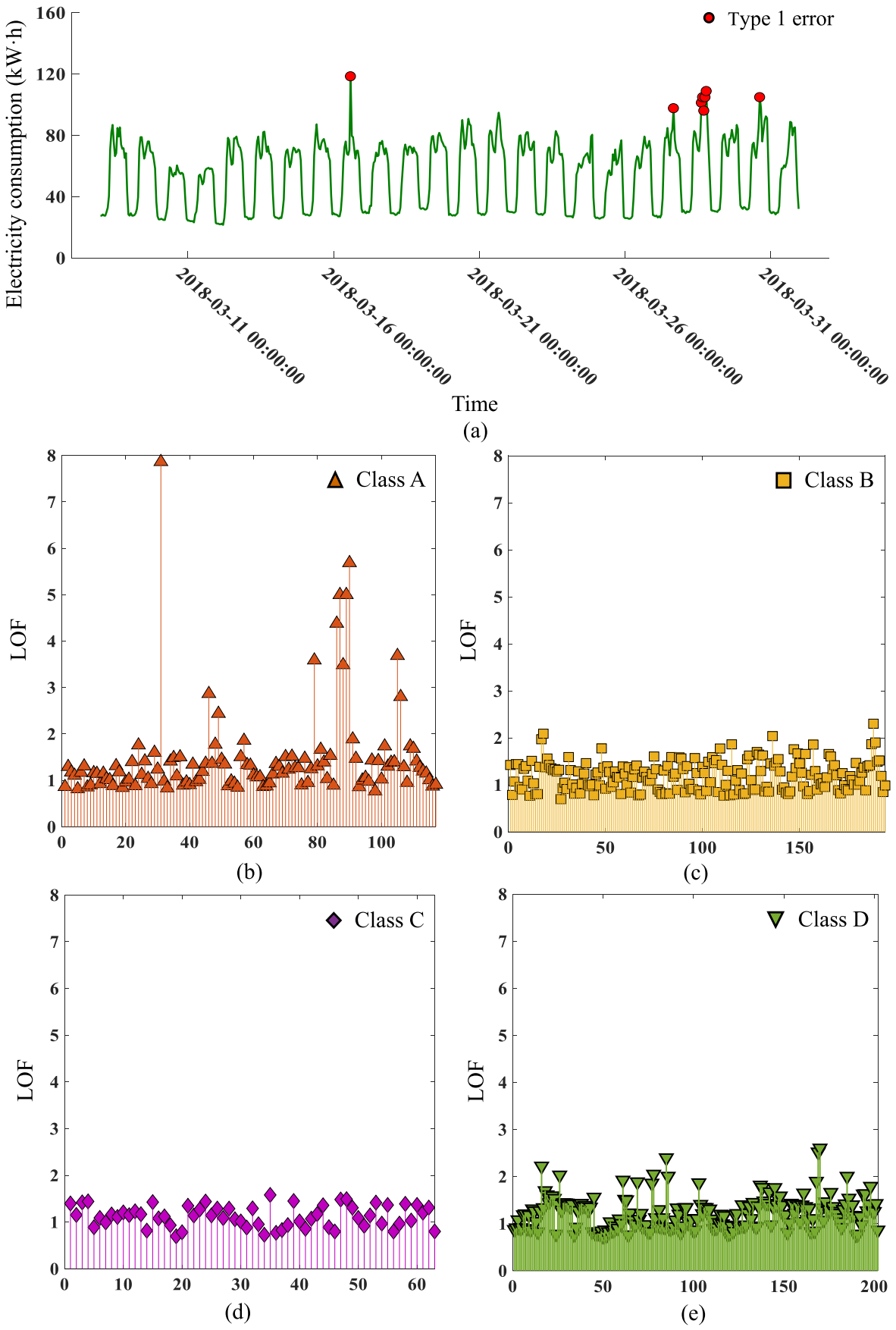


Fig. 11. The updated point anomaly detection results and LOF values for the first quarter: (a) Updated point anomaly detection results; (b) LOF values of Class A; (c) LOF values of Class B; (d) LOF values of Class C; (e) LOF values of Class D.

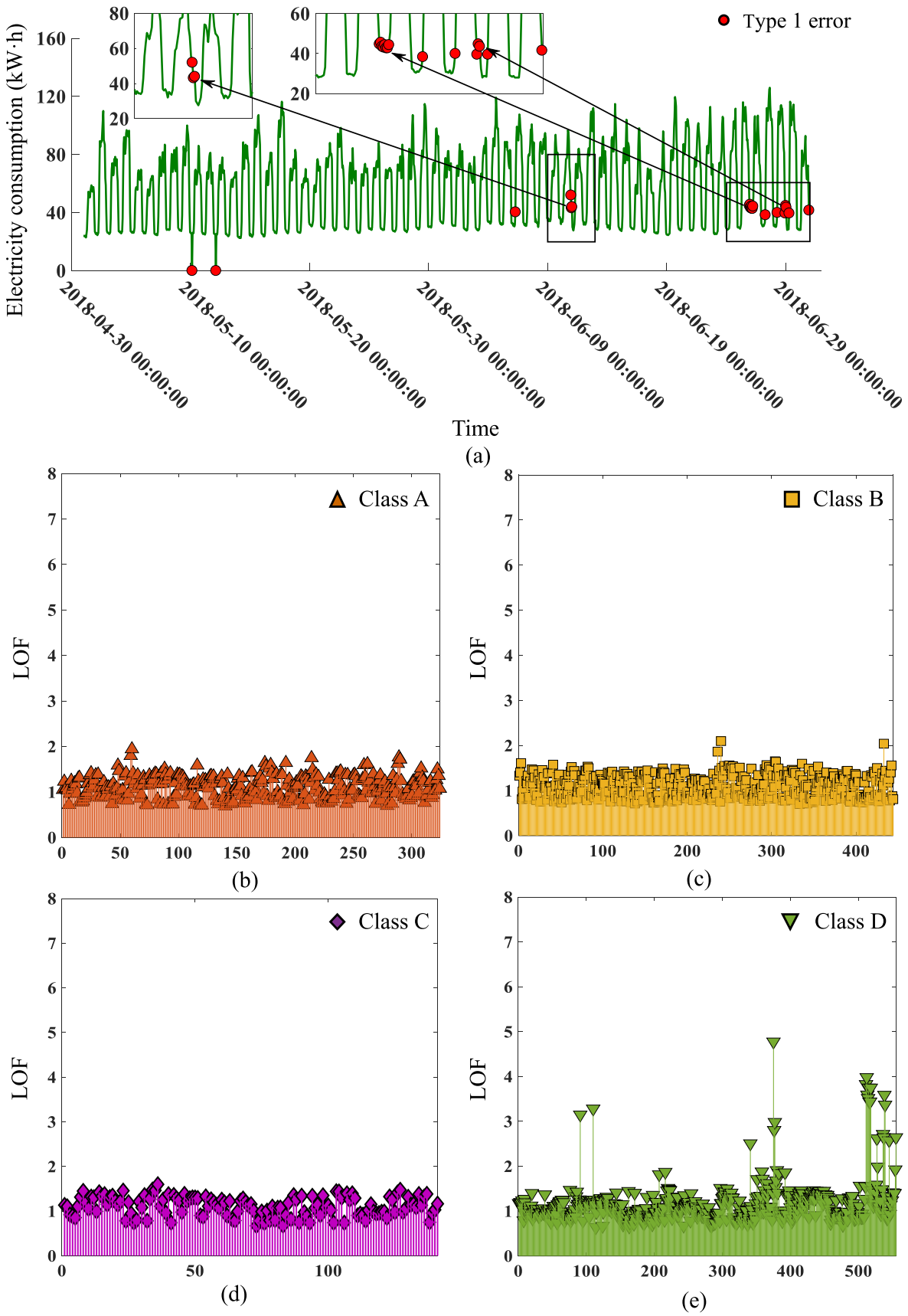


Fig. 12. Point anomaly detection results and LOF values for the second quarter: (a) Point anomaly detection results; (b) LOF values of Class A; (c) LOF values of Class B; (d) LOF values of Class C; (e) LOF values of Class D.

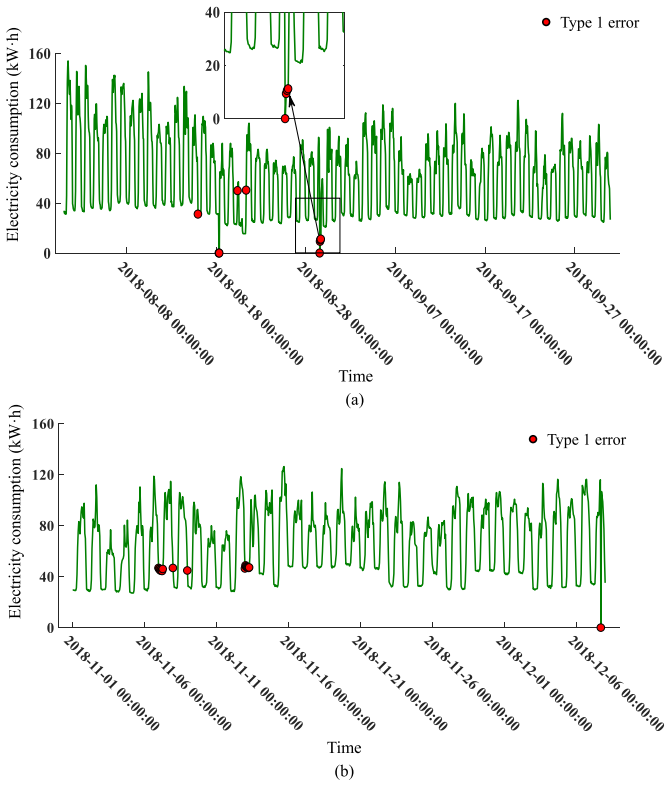


Fig. 13. Point anomaly detection results for the third and fourth quarters: (a) Detection results for the third quarter; (b) Detection results for the fourth quarter.

period with greatest electricity consumption throughout the year. Fig. 7 shows used the meteorological data in 2018.

2.2. Processing of energy consumption data

The collected data are divided into four parts. From January to March, the first quarter, the temperature and relative humidity are low, and the demand for electricity is high. However, activities in the experimental building are less because of holidays, the electricity consumption is the lowest. From April to June, the second quarter, the temperature begins to increase, the relative humidity is evenly distributed, and the electricity consumption is relatively low. From July to September, the third quarter, the temperature and relative humidity are high, and there are more experimental activities. The electricity consumption thus reaches the peak of the whole year. From October to December, the fourth quarter, the temperature and relative humidity of this quarter decrease, but they are still higher than that of the first quarter, and the electricity consumption decreases as well.

This study uses the energy consumption data of the first month of each quarter for clustering. Taking the second quarter as an example, the electricity consumption is used as the cluster sample to establish the energy consumption data matching mechanism. Energy consumption grade is established according to the level of electricity consumption [33], and the data of electricity consumption can be clustered into four classes: A, B, C, and D. The characteristics of each class are shown in Table 1. Class A is the peak of electricity consumption during a day, concentrated in the period such as 10:00–11:00 and 14:00–17:00, when the ambient temperature and occupation are high. In the night such as 19:00–21:00, due to the sudden drop of ambient temperature in Spring, the electricity consumption increases. For example, on April 9th, the temperature dropped from 10.3 °C at 14:00 to −1 °C at 20:00, and there is a peak of electricity consumption. Class B are also generated during the high-temperature period of the day, but due to low occupation in the experimental building, the peak electricity consumption still fails to meet the standard of Class A. The period of Class C overlaps with that of Class B, but in Class C, people in experimental building are even less, and electricity is consumption also lower. Class D is concentrated in the low-temperature period, roughly ranging from 22:00 to 7:00. During this period, there is no activity in the experimental building except security personnel, so the electricity consumption is the lowest in a day. It is worth noting that the lowest temperature also appears in the Class A. The clustering result is shown in Fig. 8. An abnormal data is clustered into the Class D. Because the clustering method is unsupervised and does not label abnormal energy consumption data, the algorithm automatically clusters the abnormal energy consumption data, which is in line with the anomaly detection work.

3. Results and discussions

3.1. Point anomaly detection

The energy consumption data in the first month of each quarter are clustered to calculate the average electricity consumption of various classes, as shown in Table .2. The average energy consumption in the first and third quarters is close, indicating possible similarity in the energy consumption behavior of the two quarters. It is the same for the second and fourth quarters. Fig. 9 shows the clustering results of electricity consumption in the first month of the four quarters. Except for Class D, the other three classes of energy consumption data are basically from the data during working hours. It Indicates that the electricity consumption varies greatly during working hours. Therefore, a time attribute can be added to the matching criterion of Class D energy consumption to improve the accuracy of data matching.

In order to study point anomaly detection, the hourly electricity consumption in the last two months of each quarter is detected. The detection results of the first quarter and the LOF values are shown in Fig. 10. Two abnormal electricity consumption data are detected at 3:00 a.m. on February 2nd and 3:00 a.m. on February 14th, which are data 14 and 227 in LOF values of Class D. For example, at 3:00 a.m. on February 2nd, the electricity consumption increases sharply from 34.4 kW h at 2:00 a.m. to 80 kW h as shown in Fig. 10(a). This indicates that the point anomaly detection can accurately detect abnormal energy consumption data. However, the anomaly detection in March is not satisfactory and a possible reason is that the holidays in the February lead to a sharp decrease in electricity consumption data. Many data are clustered into Class C and D and only eight electricity consumption are clustered into Class A. As the detection period covers holidays and working days, the demand for electricity changes and the demand for peak power generally decreases, so the matching mechanism needs to be reconstructed.

The electricity consumption of the first week for March is used to reconstruct the matching mechanism, and the data of the last 24 days of March are detected. The detection results and LOF values are shown in Fig. 11. The x axis of LOF results represent the amount of data. According to Fig. 11, anomalies are concentrated in Class A. A clearly abnormal at 14:00 p.m. on March 16, 2018 is successfully detected. The electricity consumption are anomalies from 15:00 to 19:00 on March 28th, 2018, which are data 86 to 89 in Fig. 11(b). The reduction of the amount of data for the matching mechanism leads to the greatly increased sensitivity of detection algorithm to data changes.

The results of the second quarter are shown in Fig. 12. It can be seen from Fig. 12(b), (c), (d) and (e) that the LOF values of three classes A, B, and C, are relatively stable. Abnormal data are concentrated in Class D, where data 91 and 110 correspond to two anomalies of 0 kW h in the detection results, 3:00 a.m. on May 10th, 2018 and 3:00 a.m. on May 12th, 2018. Also, the algorithm detects the abnormally increase in electricity consumption at night of June 10th, 2018, June 25th, 2018, and June 29th, 2018, corresponding to data 375–377, 511–518, and 536–539 of LOF values of Class D. Fig. 7(a) and (b) shows that there is no significantly change in temperature and humidity in this area, the

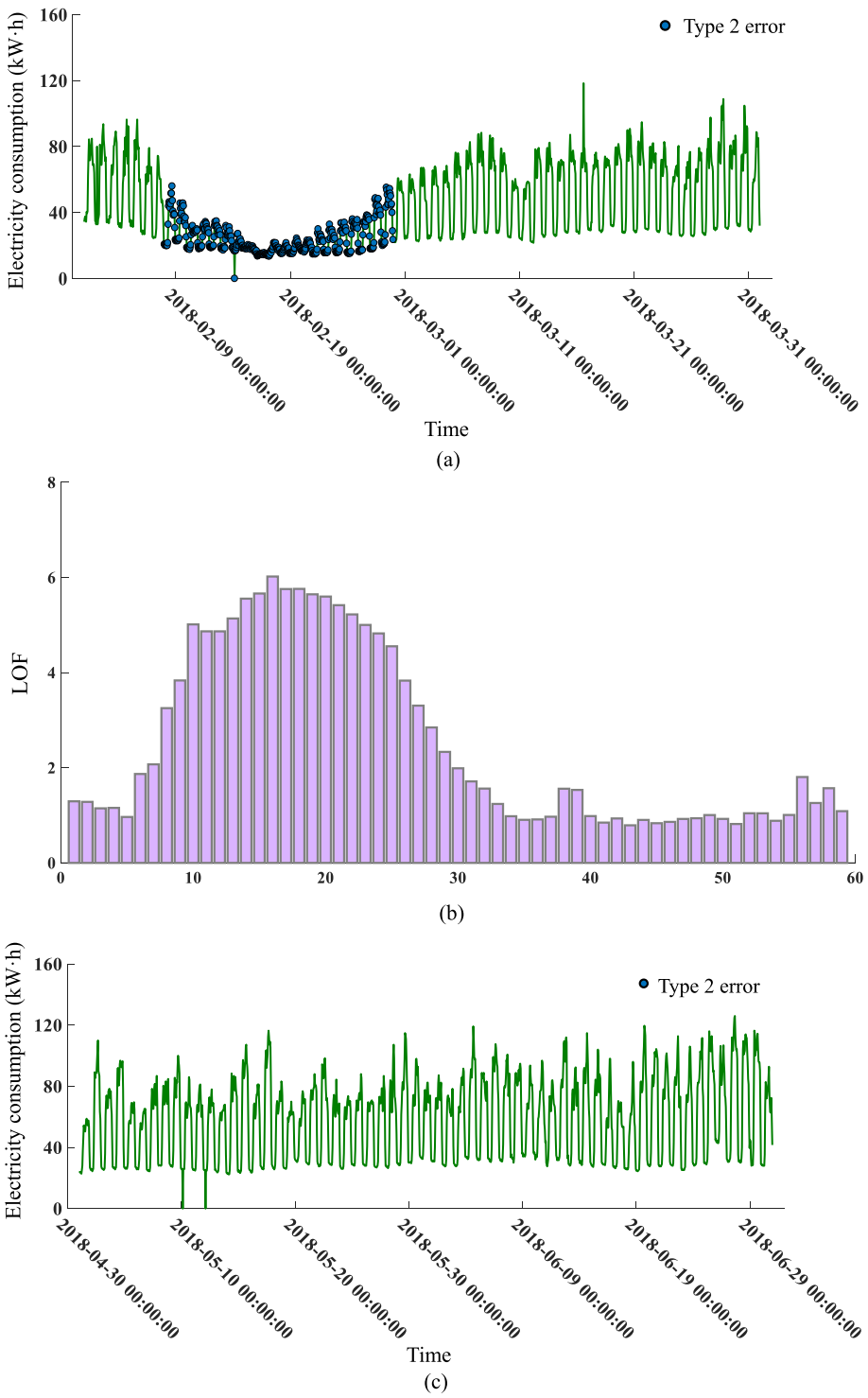


Fig. 14. Collective anomaly detection results and LOF values for the four quarters: (a) Collective anomaly detection results in the first quarter; (b) LOF values for the first quarter; (c) Collective anomaly detection results in the second quarter; (d) LOF values for the second quarter; (e) Collective anomaly detection results in the third quarter; (f) LOF values for the third quarter; (g) Collective anomaly detection results in the fourth quarter; (h) LOF values for the fourth quarter.

demand of electricity consumption should not increase significantly during this period, the anomalies are thus detected. The point anomaly detection of the third and fourth quarters are shown in Fig. 13, which also obtain expected detection results.

It is important to note that the abnormal data in the third quarter were concentrated in August. Due to the overall decline of the electricity consumption in September, indicating that electricity consumption behavior may have changed. In addition, Fig. 13(b) also shows that the detection algorithm is dynamic. Accurately identify the two drastic changes of electricity consumption on November 7th, 2018 and November 13th, 2018, updating the energy consumption of different classes to improve the accuracy of detection of subsequent energy consumption. In this paper, the point anomaly detection of electricity consumption in four quarters is carried out, and the results show that the

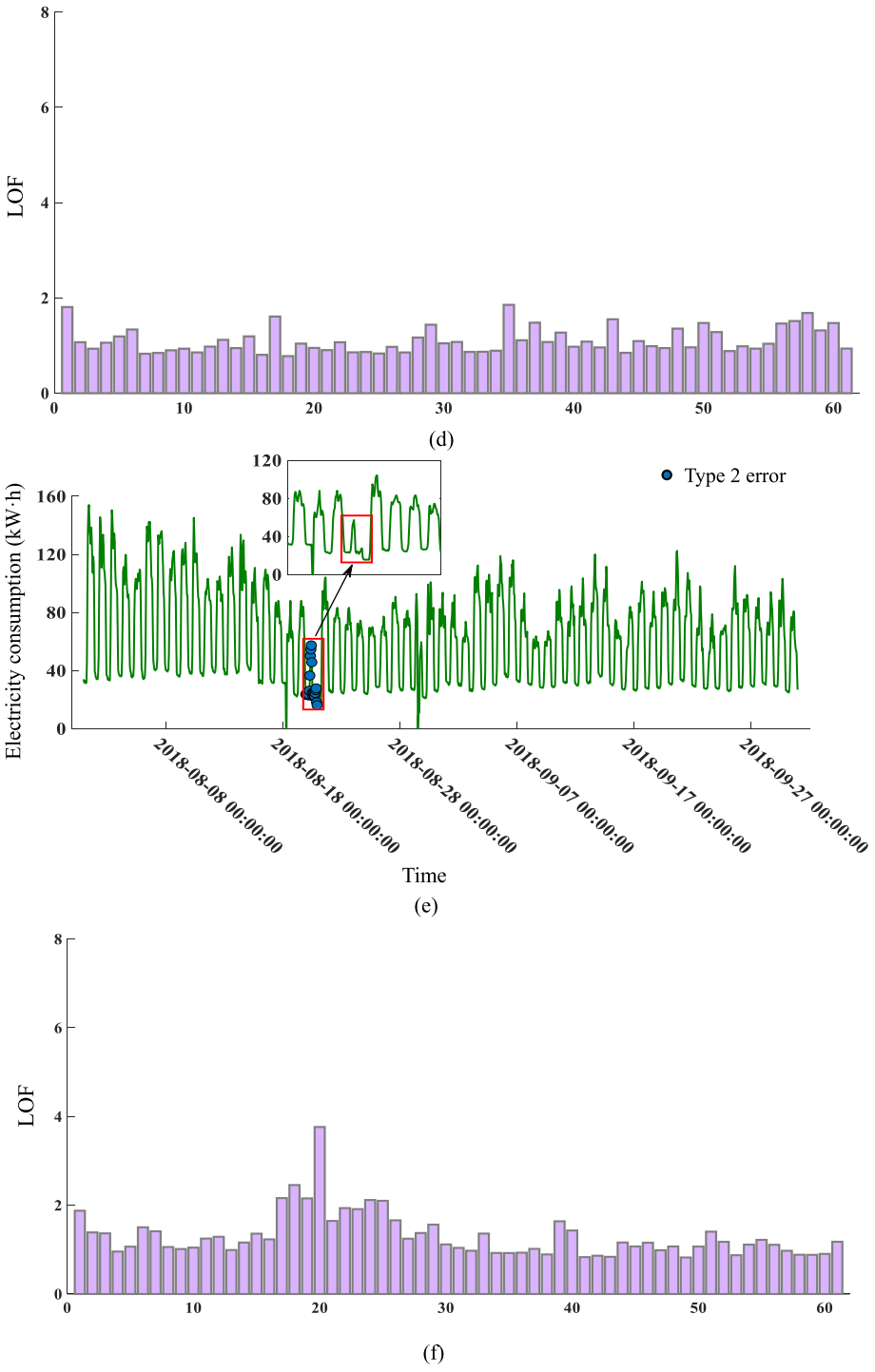


Fig. 14. (*continued*).

anomaly detection algorithm is accuracy and effectiveness for the dynamic detection of abnormal energy consumption data.

3.2. Collective anomaly detection and result analysis of energy consumption

Similarly, the data of the first month in each quarter are segmented, and the data in the last two months of each quarter are tested. The results are shown in Fig. 14. From Fig. 14(a) and (b), the electricity consumption in the first quarter shows an obvious anomaly from February 8th, 2018 to February 28th, 2018, and the electricity consumption in this data segment decreased significantly compare with the normal data segment. By comparing Fig. 7(a), it can be seen that this data segment corresponds to the period with the lowest annual temperature in Dalian, and the daily electricity consumption should be in a high state, while the actual electricity consumption is just the opposite. This indicates that there is a significant change in energy consumption behavior during this period. Because a large number of teachers and students leave school during the holiday, which results in a significant decrease in the electricity demand of the experimental building, and the daily electricity demand decreased to the lowest value on February 6th, 2018 (Spring Festival). This period corresponds to data segment 8–27 in Fig. 14(b).

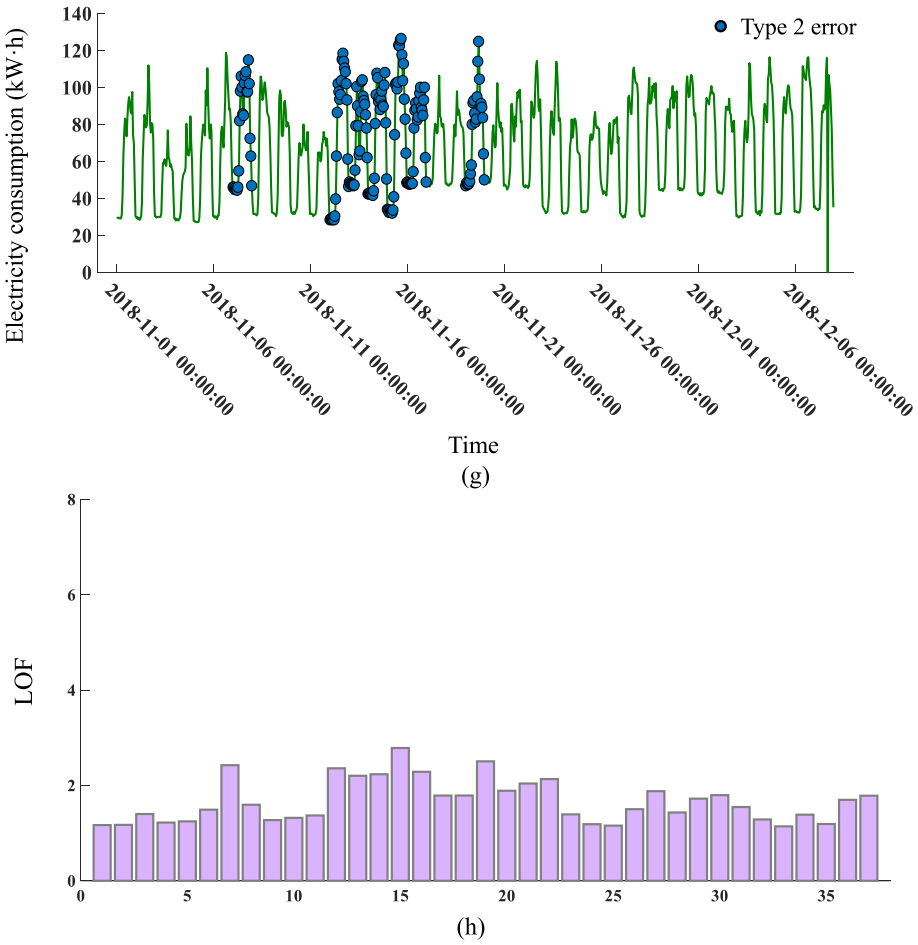


Fig. 14. (*continued*).

The LOF value of this data segment is significantly higher than that of other data segments and exceeds the threshold. Therefore, the detection algorithm determines that this data segment are collective anomalies. It is worth noting that Fig. 14(b) also shows that LOF values in data segments 38–39 and 57–59 are higher than those in nearby data segments. According to Fig. 14(a), the daily electricity consumption from March 10th, 2018 to March 11th, 2018 and from March 28th, 2018 to March 30th, 2018, are significantly different from other periods. This proves that the detection algorithm is sensitive to the fluctuation of daily electricity consumption.

The results of the second quarter show in Fig. 14(c) and (d), although LOF values of some data segments are somewhat high in this quarter, basically fluctuate around 1 and do not exceed the threshold, so it is not judged as abnormal. Fig. 14(f) shows that the data segment 17–25 of the third quarter has generally increased, corresponding to the period from August 17th, 2018 to August 25th, 2018. Fig. 14(e) shows that the daily electricity demand in this period decreased significantly compared with the previous period. The fluctuation of electricity consumption on that day is unreasonable, and it is judged as an abnormal segment.

The detection results in the fourth quarter show many abnormal segments. On November 7th, 2018 and November 12th, 2018 to November 16th, 2018 are the periods of abnormally changes in electricity consumption, corresponding to data segments 7 and 12–16 in Fig. 14(h). Since central heating is used in winter, the climate change should not cause large electricity consumption fluctuations. However, due to the use of some high-power experimental equipment, the electricity consumption fluctuated several times in this quarter, which was detected as abnormal. Then, with the dynamic update of data segments, the fluctuation of daily electricity consumption after November 21st, 2018 is no longer detected as abnormal, indicating that the dynamic update of data segments improves the accuracy of the collective anomaly detection algorithm.

3.3. Energy consumption change trend and clustering performance evaluation

3.3.1. Energy consumption change trend after detection

Complete anomaly detection of energy consumption should include two detection methods of point anomalies and collective anomalies, so as to dynamically detect the data of BECMP.

By splicing the data of the first month and the electricity consumption data after detection, the change trend of building electricity demand can be found more intuitively. The data matching results after abnormal energy consumption detection of each quarter are shown in Fig. 15. There is an obvious trenTd of overall change in electricity consumption data from Fig. 15(a) and (c). This indicates that the energy consumption behavior may change, so the matching mechanism of the detection algorithm should be updated in time.

3.3.2. Clustering performance evaluation

It can be seen from the above detection process that the basis of point anomaly detection is the clustering results of sample data. This means that the stability and accuracy of energy consumption data clustering results have a direct impact on the accuracy of point anomaly detection.

In order to explore the stability and accuracy of the energy consumption data clustering results, the data of the first month of the second quarter are taken as an example for clustering 20 times. Table 3 shows the statistical results of 20 times clustering, there are three different

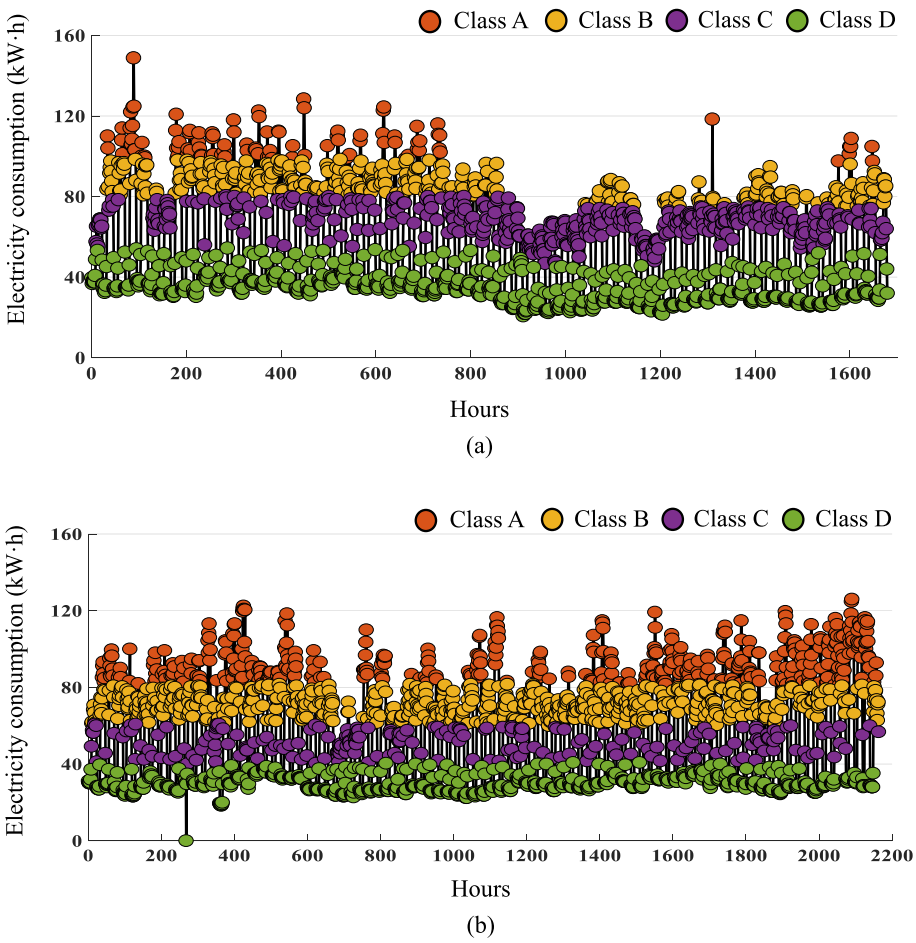


Fig. 15. Data matching results after detection: (a) First quarter; (b) Second quarter; (c) Third quarter; (d) Fourth quarter.

clustering results in total. Among them, the results of Class A data are 148 and 144, Class B data are 223 and 219, Class C data are 90 and 86, and the Class D data is stable at 263. It indicates that the electricity consumption of the experimental building fluctuates greatly in the daytime and the clustering result is prone to change, while the fluctuation is relatively small at night, so the Class D data is the most stable. At the same time, the statistical results of the 20 times clustering results also show that the error between the average values of all classes is far less than 5%, indicating that the K-medoids algorithm optimized by PSO has strong robustness to the clustering effect of the electricity consumption data.

The silhouette coefficient combines the cohesion and separation of clusters to evaluate the clustering effect. It is suitable for unlabeled datasets and has no assumptions about the distribution of the data, so it has good performance. The formula for the silhouette coefficient is defined as:

$$s = \frac{\overline{b}(x) - \overline{a}(x)}{\max(\overline{a}(x) - \overline{b}(x))} \tag{6}$$

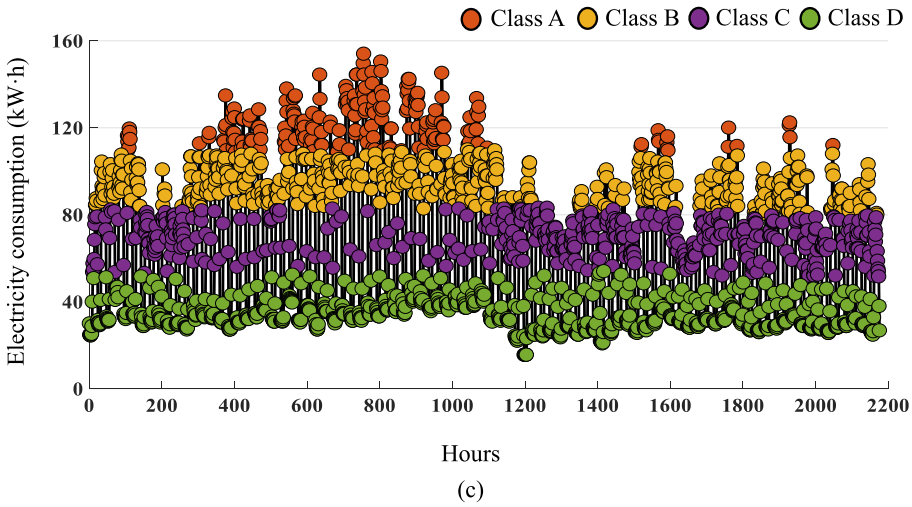
For a single sample x, $\overline{a}(x)$ in Eq. (6) is the average distance between the sample x and all other samples in the same cluster, representing the similarity within the cluster. $\overline{b}(x)$ is the average distance between sample x and all samples in the nearest neighboring cluster, indicating the similarity between clusters. s is the silhouette coefficient. According to the core idea of "small intra-cluster differences, large inter-cluster differences", $\overline{b}$ is required to be greater than $\overline{a}$, and the larger the better.

The silhouette coefficient of data ranges from −1 to 1. The closer the value is to 1, the better the data matches the cluster. Fig. 16 shows the silhouette coefficients for the clustering results. As can be seen from Fig. 16, the three clustering results are similar. The occurrence of negative values is concentrated in the boundary of Class A and Class B, except for a small number of negative values in class A, most of the other class have high silhouette coefficients, which can reach more than 80% accuracy. This proves that the K-medoids algorithm optimized by PSO is accurate in the clustering result of electricity consumption data.

In addition, this paper compared the calculation time of K-medoids clustering algorithm with and without PSO optimization under different data volume in Fig. 17. When the amount of data is less than 500, the running time of the two algorithms is very close. However, when the amount of data is greater than 500, the running time of the two algorithms differs significantly with the increase of the amount of data. Therefore, the PSO optimized K-medoids clustering algorithm avoids the problem of low calculation efficiency of the pure K-medoids clustering algorithm.

This study can tell if there is an anomaly but not the problem that leads to the anomaly. Therefore, the detection results could not be directly used for improving the building operation at the moment. In our future research, we will seek a method to further analyze the anomalies detected by the algorithm to find out the cause, to provide more direct help for improving the building operation.

The unevenly distributed occasions such as holidays would affect the energy consumption and then further the anomaly detection. To consider this, the best way is to use the historical data of the same month for anomaly detection. For example, the anomaly detection for energy data in this February should use the data from last February instead of that from this January. Due to the limitation of data, this study does not



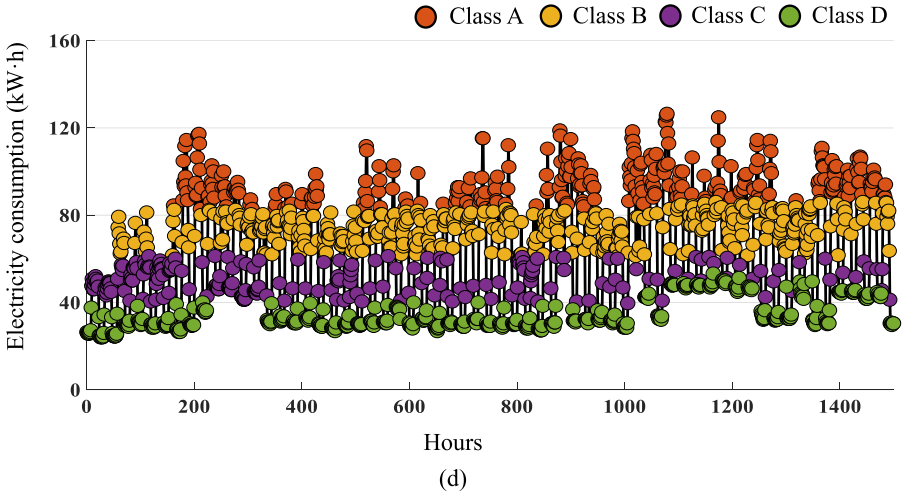


Fig. 15. (*continued*).

Table 3
Twenty clustering results of electricity consumption in the second quarter.

Order	Number of Class A	Average of Class A (kW·h)	Number of Class B	Average of Class B (kW·h)	Number of Class C	Average of Class C (kW·h)	Number of Class D	Average of Class D (kW·h)
1	144	92.625	223	71.715	90	50.429	263	29.567
2	148	92.338	219	71.527	90	50.429	263	29.567
3	148	92.338	219	71.527	90	50.429	263	29.567
4	148	92.338	223	71.335	86	49.947	263	29.567
5	148	92.338	219	71.527	90	50.429	263	29.567
6	144	92.625	223	71.715	90	50.429	263	29.567
7	148	92.338	223	71.335	86	49.947	263	29.567
8	148	92.338	223	71.335	86	49.947	263	29.567
9	148	92.338	219	71.527	90	50.429	263	29.567
10	148	92.338	223	71.715	90	50.429	263	29.567
11	148	92.338	219	71.527	90	50.429	263	29.567
12	148	92.338	223	71.335	86	49.947	263	29.567
13	148	92.338	223	71.335	86	49.947	263	29.567
14	148	92.338	223	71.335	86	49.947	263	29.567
15	148	92.338	219	71.527	90	50.429	263	29.567
16	144	92.625	223	71.715	90	50.429	263	29.567
17	144	92.625	223	71.715	90	50.429	263	29.567
18	144	92.625	223	71.715	90	50.429	263	29.567
19	148	92.338	223	71.335	86	49.947	263	29.567
20	148	92.338	219	71.527	90	50.429	263	29.567

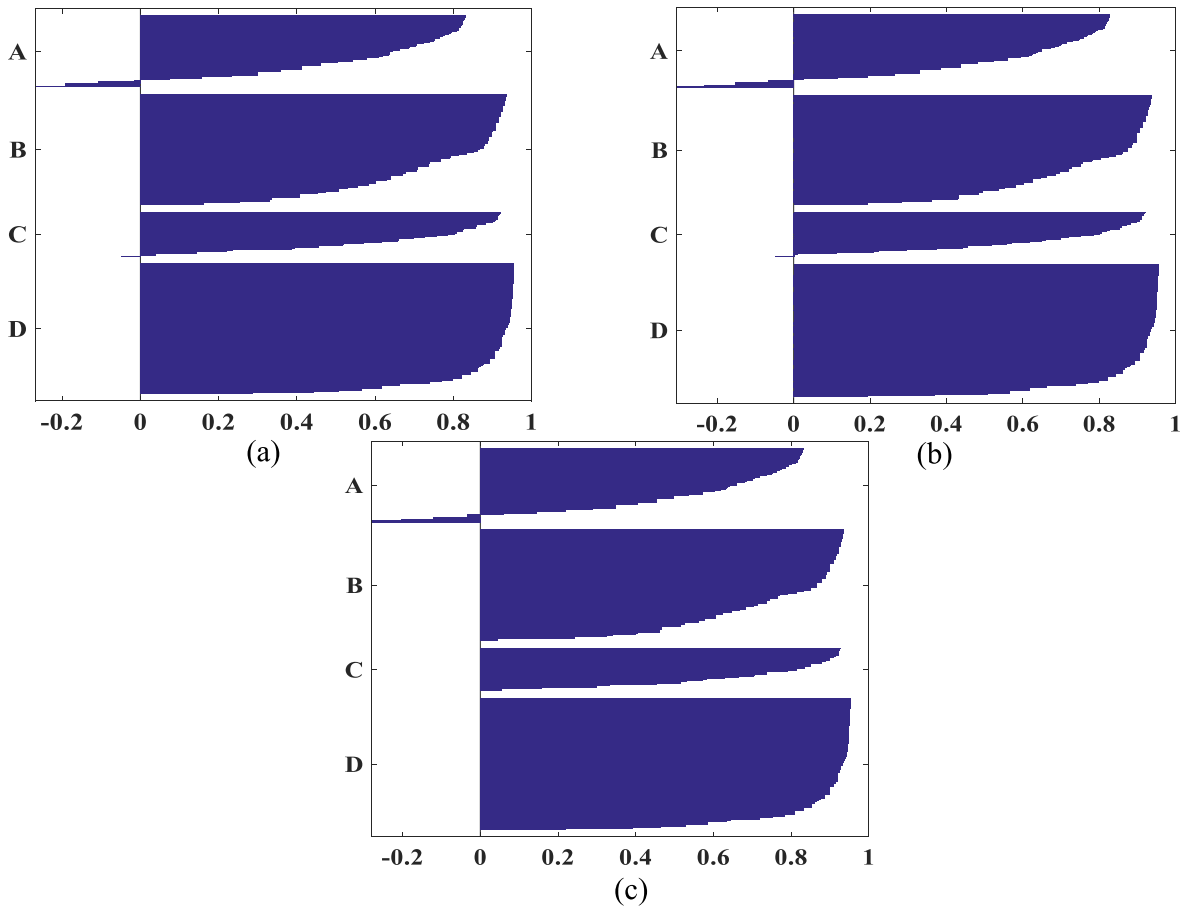


Fig. 16. Silhouette coefficients of (a) clustering result I, (b) clustering result II, and (c) clustering result III

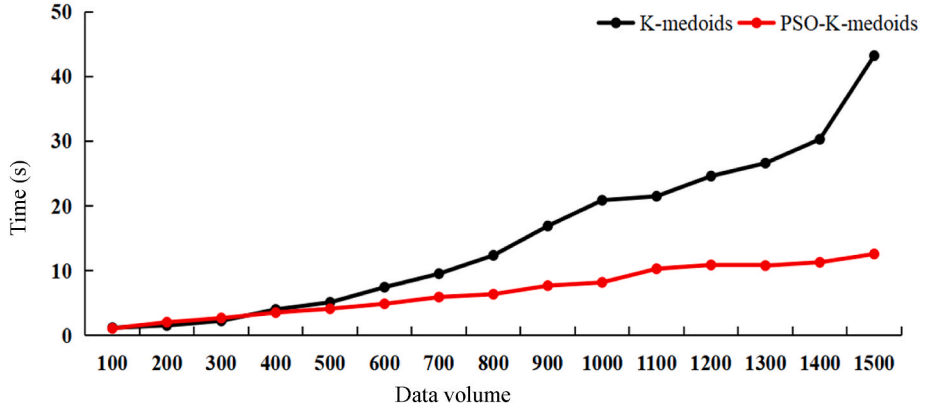


Fig. 17. The running time of K-medoids clustering algorithm with and without PSO optimization.

have the historical data to cover the occasions with a period of at least one year.

4. Conclusions

This paper proposes a dynamic anomaly detection algorithm based on BECMP. A semi-supervised algorithm is proposed, which refers to the PSO optimized K-medoids clustering algorithm combined with the KNN algorithm and DTW distance coupled with LOF algorithm. This anomaly detection algorithm can be integrated with the building operation management system to provide reliable data for building energy-saving management, operation, and maintenance. Through the detection of 8184 electricity consumption in the experimental building, the following conclusions are obtained:

(1) The detection algorithm realizes the integration of real-time energy consumption anomaly detection with data updating to ensure the dynamic anomaly detection. The detection algorithm can effectively identify the point anomalies and collective anomalies of energy consumption data.
(2) Compared with single supervised and unsupervised algorithms, the detection algorithm in this study not only reduces the interference of inaccurate label information, but also avoids the

problem of low computational efficiency of unsupervised clustering algorithm.

(3) The K-medoids clustering algorithm optimized by PSO shows excellent clustering effect, and the clustering results show that the error of the mean values of all classes are less than 5%. In addition, this study used silhouette coefficients to evaluate the clustering performance of energy consumption data, and the results show that the accuracy of each cluster reached more than 80%. It further confirms that the PSO optimized K-medoids algorithm for energy consumption data clustering has strong robustness and accuracy.

CRediT author statement

Lei Lei: Funding acquisition, Project administration, Supervision, Conceptualization, Writing - Review & Editing, Resources, Methodology. **Bing Wu:** Writing - Original Draft, Formal analysis, Investigation, Data Curation, Software, Validation.**Xin Fang**: Writing - Review & Editing, Conceptualization. **Li Chen:** Writing - Review & Editing. **Hao Wu**: Writing - Review & Editing. **Wei Liu:** Funding acquisition, Supervision, Conceptualization, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No.: 51708146 and 51808487) and Alibaba-Zhejiang University Joint Institute of Frontier Technologies (AZFT), China.

References

[1] Ma Minda, Caia Weiguang, Wu Yong. China act on the energy efficiency of civil buildings (2008): a decade review. Sci Total Environ 2019;651:42–60.
[2] Eurostat. Final energy consumption by sector. 2021. https://ec.europa.eu/eurostat/databrowser/view/ten00124/default/table?lang=en. [Accessed 6 June 2021].
[3] Monthly energy review. U.S. Energy Information Administration; June 2016.
[4] Rafe Biswas MA, Robinson Melvin D. Nelson Fumo. Prediction of residential building energy consumption: a neural network approach. Energy 2016;117:84–92.
[5] Swan LG, Ugursal VI. Modeling of end-use energy consumption in the residential sector: a review of modeling techniques. Renewable & Sustainable Energy Reviews; 2009. p. 1819–35.
[6] Perez-Lombard L, Ortiz J, Pou C. A review on buildings energy consumption information. Energy Build 2008;40:394–8.
[7] Araya Daniel B, Grolinger Katarina, ElYamany Hany F, Capretz Miriam AM, Bitsuamlak Girma. An ensemble learning framework for anomaly detection in building energy consumption. Energy Build 2017;144:191–206.
[8] Xu Chengliang, Chen Huanxin. A hybrid data mining approach for anomaly detection and evaluation in residential buildings energy data. Energy Build 2020;215:109864.
[9] Fan Cheng, Fu Xiao, Zhao Yang, Wang Jiayuan. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. Appl Energy 2018;211:1123–35.
[10] Chandola Varun, Banerjee Arindam, Kumar Vipin. Anomaly detection: a survey. ACM Comput Surv July 2009;41(3). Article 15.
[11] Xu Chengliang, Chen Huanxin. Abnormal energy consumption detection for GSHP system based on ensemble deep learning and statistical modeling method. Int J Refrig 2020;114:106–17.
[12] Capozzoli Alfonso, Lauro Fiorella, Khan Imran. Fault detection analysis using data mining techniques for a cluster of smart office buildings. Expert Syst Appl 2015;42:4324–38.
[13] Yongpan C, Xianmin M, Jili Z, et al. Development of monitoring system of building energy consumption. International Forum on Computer Science-technology & Applications. IEEE; 2010.
[14] Djuric Natasa, Novakovic Vojislav. Review of possibilities and necessities for building lifetime commissioning. Renew Sustain Energy Rev 2009;13:486–92.
[15] Imran Khana, Capozzoli Alfonso, Corgnati Stefano Paolo, Cerquitelli Tania. fault detection analysis of building energy consumption using data mining techniques. Energy Proc 2013;42:557–66.
[16] Fan Cheng, Fu Xiao, Yan Chengchu. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. Autom ConStruct 2015;50:81–90.
[17] Zhao Hai-xiang, Magoulès Frédéric. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16:3586–92.
[18] Lee Won-Yong, House John M, Kyong Nam-Ho. Subsystem level fault diagnosis of a building's air-handling unit using general regression neural networks. Appl Energy 2004;77:153–70.
[19] Liu Jiahui, Liu Jiangyan, Chen Huanxin, Huang Ronggeng, Li Zhengfei, Liu Pengfei, Guo Yabin, Shen Jiaqin, Wang Yue. Abnormal energy identification of variable refrigerant flow air-conditioning systems based on data mining techniques. Appl Therm Eng 2019;150. 398-41.
[20] Du Zhimin, Fan Bo, Jin Xinqiao, Chi Jinlei. Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. Build Environ 2014;73:1–11.
[21] Wang Shengwei, Chen Youming. Fault-tolerant control for outdoor ventilation air flow rate in buildings based on neural network. Build Environ 2002;37:691–704.
[22] Magoulès Frédéric, Zhao Hai-xiang, Elizondo David. Development of an RDP neural network for building energy consumption fault detection and diagnosis. Energy Build 2013;62:133–8.
[23] Han H, G u B, Kang J, Li ZR. Study on a hybrid SVM model for chiller FDD applications. Appl Therm Eng 2011;31:582–92.
[24] Sun Kaizheng, Li Guannan, Chen Huanxin, Liu Jiangyan, Li Jiong, Hu Wenju. A novel efficient SVM-based fault diagnosis method for multi-split air conditioning system's refrigerant charge fault amount. Appl Therm Eng 2016;108:989–98.
[25] Zhao Yang, Wang Shengwei, Fu Xiao. Pattern recognition-based chillers fault detection method using Support Vector Data Description (SVDD). Appl Energy 2013;112:1041–8.
[26] Motta Cabrera David F, Zareipour Hamidreza. Data association mining for identifying lighting energy waste patterns in educational institutes. Energy Build 2013;62:210–6.
[27] Li Guannan, Hu Yunpeng, Chen Huanxin, Li Haorong, Hu Min, Guo Yabin, Liu Jiangyan. Shaobo Sun, Miao Sun. Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. Appl Energy 2017;185:846–61.
[28] Ma Zhenjun, Yan Rui, Nord Natasa. A variation focused cluster analysis strategy to identify typical daily heating load profiles of higher education buildings. Energy 2017;134:90–102.
[29] Ma Zhongjiao, Song Jialin, Zhang Jili. A real-time detection method of abnormal building energy consumption data coupled POD-LSE and FCD. Procedia Eng 2017;205:1657–64.
[30] Jiang Hang, Lu Tun, Gu Hansu, et al. A dynamic and real-time outlier detection method for energy consumption of campus building. Comput Eng 2017;43(4):15–2027.
[31] Sheng Weiguo, Liu Xiaohui. A genetic k-medoids clustering algorithm. J Heuristics 2006;12:447–66.
[32] Kennedy J, Eberhart R. Particle swarm optimization. Icnn95-international conference on neural networks. IEEE; 2002.
[33] Lei Lei, Chen Wei, Wu Bing, Chen Chao, Liu Wei. A building energy consumption prediction model based on rough set theory and deep learning algorithms. Energy Build 2021;240:110886.