

Article

Graph Convolutional Spectral Clustering for Electricity Market Data Clustering

Longda Huang ¹, Maohua Shan ¹, Liguu Weng ^{2,*} and Lingyi Meng ^{2,3}

¹ China Electric Power Research Institute Co., Ltd., Nanjing 210003, China; huanglongda@epri.sgcc.com.cn (L.H.); danmaohua@epri.sgcc.com.cn (M.S.)

² Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China; bq834744@student.reading.ac.uk

³ Department of Computer Science, University of Reading, Whiteknights, Reading RG6 6DH, UK

* Correspondence: 002311@nuist.edu.cn

Abstract: As the power grid undergoes transformation and the Internet's influence grows, the electricity market is evolving towards informatization. The expanding scale of the power grid and the increasing complexity of operating conditions have generated a substantial amount of data in the power market. The traditional power marketing model is no longer suitable for the modern power market's development trend. To tackle this challenge, this study employs random forest and RBF models for processing electricity market data. Additionally, it explores the synergy of graph convolutional network and spectral clustering algorithms to enhance the accuracy and efficiency of data mining, enabling a comprehensive analysis of data features. The experimental results successfully extracted various electricity consumption features. This approach contributes to the informatization efforts of power grid enterprises, enhances power data perception capabilities, and offers reliable support for decision makers.

Keywords: power data; data mining; big data analysis; clustering analysis



Citation: Huang, L.; Shan, M.; Weng, L.; Meng, L. Graph Convolutional Spectral Clustering for Electricity Market Data Clustering. *Appl. Sci.* **2024**, *14*, 5263. <https://doi.org/10.3390/app14125263>

Academic Editor: Filipe Soares

Received: 16 April 2024

Revised: 10 June 2024

Accepted: 14 June 2024

Published: 18 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the deep integration of information and communication technology with power production and the comprehensive development of the smart grid and energy internet, there is an explosive growth trend in power data, accompanied by increasing diversity in data types and sources [1,2]. Power data are transitioning from structured, small-scale, and low-speed data to the realm of big data in the electricity sector. Faced with massive electricity consumption information, challenges include long-term storage of data, complex data structures, high computational frequency, and disorderly mapping of logical relationships [3]. Addressing these characteristics requires expanding the storage capacity of electricity information collection systems, optimizing intelligent algorithms, and improving parallel computing efficiency in distribution grid automation systems. Therefore, there is an urgent need to employ novel data processing technologies and methods to extract the operational and societal service value embedded in the data [4].

Currently, research on power big data mainly focuses on its application status, key technologies, and prospects. Boyd [5], for instance, outlined the types and collection technologies of power big data in various operational scenarios of the power grid. Weinzettel et al. [6] analyzed the application prospects of power big data in enhancing lean management levels in the power grid, optimizing energy scheduling, transmission, and serving users. Additionally, some studies have explored the value inherent in power big data. For example, power big data plays a significant role and holds value in supporting decision making in energy enterprise management, developing energy-saving and environmentally friendly products, and providing comprehensive energy data services [7]. Power big data flows horizontally and vertically across different business areas in the power grid, contributing

to precise grid governance and enhancing social service capabilities [8]. Zhao et al. [9] streamlined the data flow process to simplify intelligent distribution grid operations, proposing a data-driven value chain to improve the accuracy and adaptability of output results. Park et al. [10], based on information value theory, constructed a power service industry value chain from the supply side to the demand side and analyzed the business models of power big data. Janssen et al. [11] analyzed the current status of power big data applications and decision making, proposing critical paths and methods for power big data to support management decisions.

In the realm of integrating and mining electricity consumption information, the current mainstream approach involves leveraging big data mining and clustering methods to unearth implicit associations among electricity consumption data. C3 Energy, a company based in the United States, has pioneered the development of the C3 Energy Analytics Engine platform in the field of integrating and analyzing large-scale electricity consumption data. Within this platform, two energy analysis tools have been developed to analyze the electricity consumption behavior of industrial users, commercial users, and residential users in distribution networks. Lin et al. [12] explored the fusion of association rules and wavelet neural networks, constructing a real-time online cleansing model for big data. By conducting correlation analysis on data dimensions, the model differentiates between true data anomalies caused by equipment failures and anomalies resulting from sensor measurement errors. This is achieved through network parameter adjustments, enhancing the efficiency of data cleansing. Lv et al. [13] focused on user time-series curves, employing symbol clustering for dimensionality reduction. Subsequently, Euclidean distance was used for similarity measurement on the reduced-dimensional data, obtaining reference curves for the original data. Interpolation fusion was then performed based on the reference curves to handle anomalies in the original curves. Lin et al. [14] proposed a curve smoothing and dimensionality reduction method for power user time-series based on information entropy segment aggregation and slope event processing. Traditional Euclidean distance in spectral clustering algorithms was replaced with dual-scale similarity measurement, enabling the unsupervised clustering of power users. This contributes to the formulation of strategies for power companies in response to user electricity behavior in the electricity market. Mao et al. [15] addressed the complexity and decentralization of electricity consumption data resulting from the widespread use of smart meters. They introduced an adaptive K-means distributed clustering algorithm, locally clustering smart meter data for physically proximate power users, obtaining typical local user curves. Weighted values were assigned to each local user curve, and traditional clustering methods were employed for global secondary clustering, revealing typical user electricity behaviors. Liu et al. [16] tackled the issue of load data category imbalance and large dataset sizes in power systems. They utilized a feedback-corrected back propagation neural network on the Spark parallel computing platform for global clustering. Local clustering of load data was accomplished using the K-medoids clustering method for each training sample. To address imbalances in local samples, oversampling methods were applied, achieving data classification. The aforementioned studies illustrate an increasingly profound global understanding of the value of power big data. Research on supporting the production, operation, decision making, forecasting, management, and services of power and grid enterprises has shown initial success. However, research on the value creation process, models, and mechanisms remains relatively nascent.

Furthermore, researchers have identified that graph structures possess the capability to capture intrinsic relationships within data [17,18], often exhibiting stronger expressive power than the original data. Consequently, graph learning emerges as a fundamental method of representing data within a graph structure, widely applied in hierarchical clustering and spectral clustering algorithms. Zhang et al. [19] pioneered the use of a sparse weakly connected graph to generate initial clusters. They measured the similarity between clusters using the product of the average in-degree and the average out-degree in the k-nearest neighbor graph, followed by agglomeration in hierarchical clustering.

This approach effectively addresses density imbalance issues in high-dimensional spaces. Zhang et al. [20] introduced a novel graph structure hierarchical clustering method, utilizing path integral as the structural descriptor of clustering and incremental path integral as the similarity measure between clusters. Barton et al. [21] proposed an enhanced graph-learning-based hierarchical clustering algorithm named Chameleon2. They modified the internal clustering quality measure, added an extra step to ensure algorithm robustness, and introduced an automatic cutoff selection method for producing high-quality clusters autonomously. The graph partitioning method based on adjacency matrix eigenvectors was initially proposed by Donath and Hoffman [22]. Fiedler [23] later suggested using the second eigenvector of the Laplacian matrix for graph partitioning, demonstrating its close relationship to binary partitioning. Hagen and Kahng [24] subsequently identified the correlation between clustering, graph partitioning, and eigenvectors of similarity matrices, introducing a proportional cut model. Despite the successful applications of spectral clustering in various domains, there remain many challenges requiring further in-depth research, particularly in the clustering of high-dimensional complex feature data, which remains a current focal point in research.

In summary, as the scale of the electrical grid continues to expand and the operating environment becomes increasingly complex, the electricity usage information in the power market is showing a rapidly growing trend in both spatial and temporal dimensions, presenting a significant exponential growth trend. However, current research still primarily focuses on using traditional methods to process and analyze these data. At the same time, the presence of a large amount of redundant information in massive datasets makes it exceptionally difficult to identify effective data, thus complicating the precise capture of individualized characteristics of user electricity behavior. This further increases the challenges faced in handling high-dimensional, multi-source data, imposing higher demands on existing data processing and analysis techniques. Therefore, this study aims to effectively handle large-scale electrical data and extract relationships between user electricity usage characteristics, a problem that has not been fully resolved. This paper introduces data fusion and feature extraction techniques in the electricity market. By applying Isolation Forest algorithms and radial basis function (RBF) neural networks, data fusion is achieved from the perspectives of the spatial network distribution of electric power data and unified timestamps. Additionally, a method using graph convolutional networks (GCN) and spectral clustering for data dimensionality reduction and feature extraction is proposed. This method can analyze various types of electric load, meeting the diverse electricity needs of different types of end users. By conducting intelligent data fusion and analysis of electricity data, the accuracy and efficiency of data processing are improved, thereby supporting decision making in the power market.

2. Integration and Mining of Electricity Market Data

2.1. Electricity Information Data Fusion

Electricity consumption information big data fusion primarily involves advanced data processing methods applied on the foundational data layer [25], as shown in Figure 1. Focused on the reliability of power supply data for end-users, this process incorporates the Isolation Forest algorithm [26] and RBF neural network [27] into the handling of electricity big data to achieve data fusion at the fundamental level. The Isolation Forest algorithm primarily targets anomalous data, enabling rapid detection of anomalies within massive datasets. Concurrently, it considers the spatial network distribution of users, leveraging associated data from neighboring users sharing the same meter box. This additional step aids in further assessing whether the detected anomaly for a user is indicative of a fault or a metering storage anomaly. Based on this determination, the approach decides whether to use RBF neural network for the reconstruction of anomaly data, thereby achieving the fusion of data timestamps.

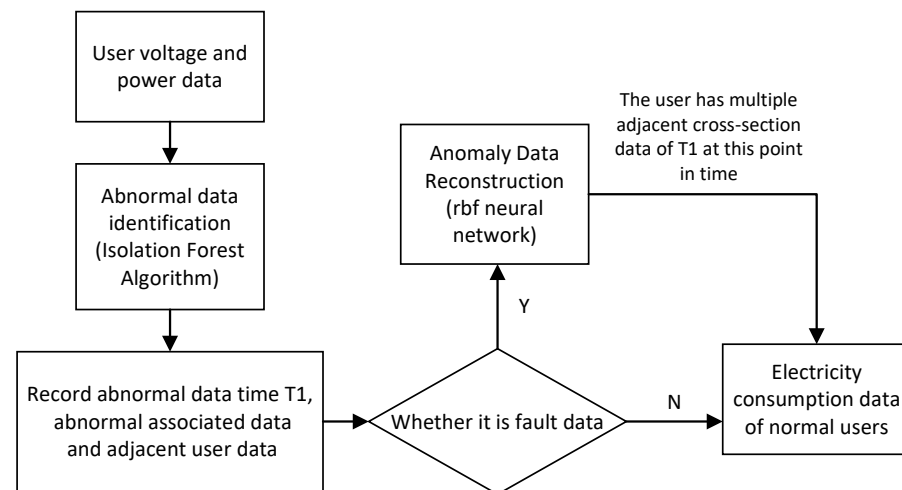


Figure 1. Electricity consumption information fusion flow chart.

2.1.1. The Isolation Forest Algorithm

Isolation Forest theory belongs to non-parametric testing and unsupervised learning and has been used for detecting anomalous data [28]. The core idea of this theory is to partition the data sample space using random planes. For each subsample space, based on the principle that there is only one type of user in each subsample space, sample attributes are continuously selected to perform subsample space partitioning. In this process, since anomalous data are outliers, the number of splits they undergo is generally less than the number of splits experienced by normal data. In the Isolation Forest algorithm, the path length of outlier values is typically short. When the data density is high, outliers require multiple splits to be found, whereas when the density is low, outliers can be found with just a few simple splits. Therefore, as a normal data point, x_i needs multiple splits to be isolated, while x_0 requires few splits.

Isolation Forest theory mainly includes two steps: training trees and scoring data points. For a training tree, the main steps include the following:

Firstly, it is assumed that the dataset $X = \{x_1, \dots, x_n\}$, where x_i represents a data point with m dimensions. η data points are randomly selected as the training subsample set γ , which is placed under the root node of the tree $T_{l,r,h}$. Here, h denotes the height of the tree, l is the left node at height h , and r is the right node at height h .

Next, a random dimension j is chosen from the m dimensions of the data points. Within the current node's dataset, a random split value p is selected for dimension j , with p ranging between the maximum and minimum values of dimension j in the subsample set γ . A hyperplane is formed using this split value p , placing data points smaller than p on the left side l of the current node and those greater than p on the right side r , creating two subsample sets $\gamma_{l,h}$ and $\gamma_{r,h}$.

Finally, the steps described above are recursively applied to the datasets of the left and right branches of the node, generating new leaf nodes until the dataset at a leaf node contains only one data point, or the tree height reaches the designated growth height, at which point the splitting stops. To illustrate, in the case of user voltage data, abnormal data timestamps are identified using isolation forest theory detection, followed by an assessment of the existence of abnormal data.

2.1.2. Exceptional Data Reconstruction Based on RBF Neural Network

The power data of the user is first preprocessed by the abnormal data identification algorithm based on isolated forest to distinguish the normal data from the abnormal data caused by harmonic interference, fault and other factors. Subsequently, these preprocessed data are used as input values and substituted into the RBF neural network for network training. The RBF neural network consists of an input layer, a hidden layer, and an output

layer, representing a forward three-layer structured network. The network topology architecture is illustrated in Figure 2.

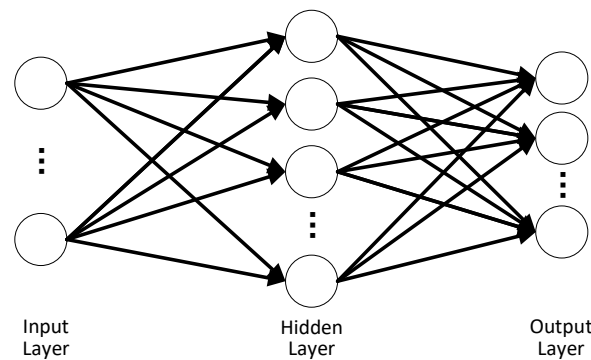


Figure 2. RBF neural network topology architecture.

The core elements of the network architecture include center vectors, the number of nodes in the hidden layer, radial basis function width, and weight matrices. The network parameters, namely the radial basis function width, weight matrix, number of hidden layer units, and centers, are explicitly determined through network training. The key to network training is specifying an appropriate number and orientation of center points. The expression for the radial basis function is denoted as Equation (1):

$$F(\|\theta - W_j\|^2) = \exp\left(-\frac{\psi}{d_{t\max}}\|\theta - W_j\|\right), \quad (1)$$

where W_j represents the center of the radial basis function and the input vector θ has the same dimensions. The symbol ψ represents the number of center points, and $\theta - W_j$ represents the distance between the input vector and the center point, with $d_{t\max}$ representing the maximum distance between center points.

The width parameter λ of the radial basis is represented as Equation (2) and the expansion constant β of the radial basis is expressed Equation (3):

$$\lambda = \frac{d_{t\max}}{\sqrt{2\psi}}, \quad (2)$$

$$\beta = \delta d_{\min}, \quad (3)$$

where $d_{\min}$ represents the distance between each center point, and δ represents the overlap index. A cost function ω is defined Equation (4) to measure the network's output y_z .

$$\omega = \frac{1}{2} \sum_{i=1}^n (\varphi_i - y_{z_i})^2, \quad (4)$$

where φ_i and y_{z_i} represent the expected output and actual output of the RBF neural network output nodes, respectively.

The adjustment process of the network weight matrix, radial basis function width parameter, and adjustment of the center points of the hidden layer units during the t time period is estimated using the gradient descent method. Assuming new data samples are input into the network and network parameters need to be calibrated, parameters can be modified through gradient descent. After a finite number of adjustments, the output error of the network for consistent data mining can be kept within an acceptable range. If the error ω is less than the allowable error, the sample does not need adjustment. Finally, the formula for calculating the output of the RBF neural network training is expressed as

Equation (5), completing the entire algorithm process and obtaining the results of consistent data mining.

$$f(x) = \hat{\eta}\vartheta, \quad (5)$$

where x represents the input data of RBF network and $f(x)$ represents the output result of RBF network, $\hat{\eta}$ represents the hidden layer output matrix of RBF network, and ϑ represents the output weights of RBF network.

The fundamental idea of network learning is based on the assumption that there is an inherent law of consistency in associated data. A corresponding network is created to mimic this pattern, thereby reconstructing abnormal data. The sample sequence of data is denoted as $\{\sigma(l), u(l)\}_{l=1}^{\infty}$, where $\sigma(l)$ is the input sample, and $u(l)$ is the expected output relative to $\sigma(l)$. At time t , the input sample $(\sigma(l), u(l))$ is used, and the calibration of network parameters is determined based on the difference between the real output of the RBF neural network and the expected value. This process completes the reconstruction and mining of abnormal data.

The fusion of electricity market data in this case uses the RBF neural network. The fusion of electricity market data involves integrating data from different sources, and the RBF neural network plays a crucial role in this context, effectively addressing possible curvature changes and volatility between the data. This segmented approximation ensures the smoothness of the interpolation function near the data points, providing strong support for accurately understanding market trends and patterns.

2.2. Electricity Information Data Mining Based on Clustering Algorithm

With the rapid development of smart grids and the establishment of advanced measurement technologies, the accumulation of power data in distribution networks has grown exponentially over time. Behind the massive power data lies deeper connections, making it crucial to uncover the hidden meanings in the data. Simultaneously, with the widespread use of smart meters, it is possible to collect electricity consumption data from users at higher densities, capturing fluctuations in load data and better reflecting user electricity consumption behavior. Therefore, employing big data deep mining techniques for mining electricity consumption behavior can provide a better understanding of user individuality and guide the deep-level expansion of business for power grid companies [29,30].

Since users' massive electricity consumption data are stored at multiple time scales, such as yearly, monthly, and daily, and the data dimensions are large with redundant information, direct application of clustering methods would not only consume a significant amount of time for calculations but also severely affect the accuracy of clustering. Therefore, this section focuses on two aspects: data dimensionality reduction and clustering for large-scale electricity consumption information data mining. On the one hand, to address the issue of large data dimensions, this section adopts an approach based on information entropy segmentation aggregation approximation. This method reduces data dimensions while ensuring a re-expression of user load fluctuation characteristics, resulting in a re-expression sequence with variable time resolution for load data. On the other hand, for clustering of re-expressed sequences, this study combines GCN and spectral clustering algorithms [31,32], which does not require setting cluster centers. The use of graph convolutional networks can process non-Euclidean data and effectively extract data features for learning, which is beneficial to capturing global information and thus better expressing the characteristics of nodes. At the same time, the feature matrix is used to further reduce the dimensionality of the data and reduce the computational complexity. Therefore, this paper uses spectral clustering algorithm to cluster end users, thereby extracting typical daily load curves and providing data support for the construction of end-user power supply demand reliability indicators.

2.2.1. Feature Extraction of End-User Load Data

In the face of the high dimensionality of terminal user load data, this section begins by describing the fluctuation of load curves using the concept of entropy from statistics. Based on the numerical values of entropy, the load curve is segmented, and then segment-wise aggregation is employed to achieve numerical re-expression of each segment. This process is aimed at reducing the dimensionality of load data and extracting features.

Analysis of Load Time-Series Fluctuations

The fluctuation degree of the load time-series curve is analyzed using information entropy, where a higher information entropy value corresponds to a greater fluctuation degree. Suppose X represents the load data of a time-series segment. By statistically analyzing the load values within this time period, let us assume there are n different load values denoted as $x_1, x_2, \dots, x_n$, and the probability of each load value occurring is represented by $p_1, p_2, \dots, p_n$. The entropy value H_n of this segment of time-series is then calculated as Equation (6):

$$H_n(p_1, p_2, \dots, p_n) = - \sum_{i=1}^n p_i \ln p_i, 0 < p_i < 1, \sum_{i=1}^n p_i = 1. \quad (6)$$

From the above formula, it can be observed that, when $p_1 = p_2 = p_3 = \dots = p_n$, $p_i = 1/n$, $H_n(p_1, p_2, \dots, p_n) = H_{max} = \ln n$ which is the maximum value for the entropy of this time-series segment. Simultaneously, the average entropy is defined as Equation (7):

$$\overline{H_n} = \sum_{i=2}^n \frac{1}{i}. \quad (7)$$

In order to comprehensively consider the entropy value, maximum entropy, and average entropy of the load data, a new concept τ_m is introduced to describe the fluctuation degree of the load curve for m loads in this time-series segment. When its value is 1, it indicates a significant fluctuation, and the data mean cannot be used as a substitute. The specific definition of τ_m is given as Equation (8):

$$\tau_m = \begin{cases} 1, & \frac{H_n(X_m)}{\ln n} > \omega \overline{H_n}(X_m) \\ 0, & \frac{H_n(X_m)}{\ln n} \leq \omega \overline{H_n}(X_m) \end{cases}. \quad (8)$$

In the formula, $m = 1, 2, \dots, M$; M represents the total number of loads; the scale factor is ω , its value range is $[1, 2]$, and in this article, $\omega = 1$.

By synthesizing the τ_m of the time section of multiple loads, the proportion of the number of loads of $\tau_m = 1$ in M loads is obtained, and if the value exceeds the threshold, it means that the curve of this section of M loads is very different, and it needs to be described in segments. The specific definitions of ρ as Equation (9):

$$\rho = \frac{1}{M} \sum_{m=1}^M \tau_m. \quad (9)$$

Load Time-Series Re-Expression

Determining the number of segments based on the fluctuation level of the load curve, we then employ segmented aggregation approximation to interpret high-dimensional time-series through low-dimensional data features. This approach achieves the reduction and re-expression of load time-series while preserving the individual characteristics of the load curve shape.

Take the load data $Q = \{q_1, q_2, \dots, q_L\}$ of a complete time-series; its time-series length is L . By segmenting the elements in it, a new data sequence is $\bar{Q} = \{\bar{q}_1, \bar{q}_2, \dots, \bar{q}_l\}$ with a length of l , where $l \mid L$ and l are divisible by L . The specific calculation process of $\bar{q}_i$ is as Equation (10):

$$\bar{q}_i = \frac{l}{L} \sum_{j=\frac{l}{L}(i-1)+1}^{\frac{l}{L}i} q_j, i \in \{1, l\}, j \in \{1, L\}. \quad (10)$$

2.2.2. Improved Spectral Clustering Algorithm Combining GCN and Spectral Filter

Spectral clustering is a widely used unsupervised learning method that proves effective in many practical applications [33]. However, it faces certain challenges, such as the inability to learn effective representations between non-Euclidean data in traditional spectral clustering. Moreover, algorithms based on spectral clustering require eigenvalue decomposition of the Laplacian matrix, which is time-consuming, especially when dealing with large-scale data. Additionally, applying the k-means algorithm for spectral clustering in large-scale data can also be time-consuming. To address these issues, this paper combines GCN and Chebyshev polynomial approximation of spectral filters to improve the spectral clustering algorithm, as shown in Figure 3. Initially, GCN are used to represent and analyze non-Euclidean data, followed by the construction of the Laplacian matrix. Simultaneously, this chapter employs Chebyshev polynomial approximation of spectral filtering techniques to avoid eigenvalue decomposition of the Laplacian matrix.

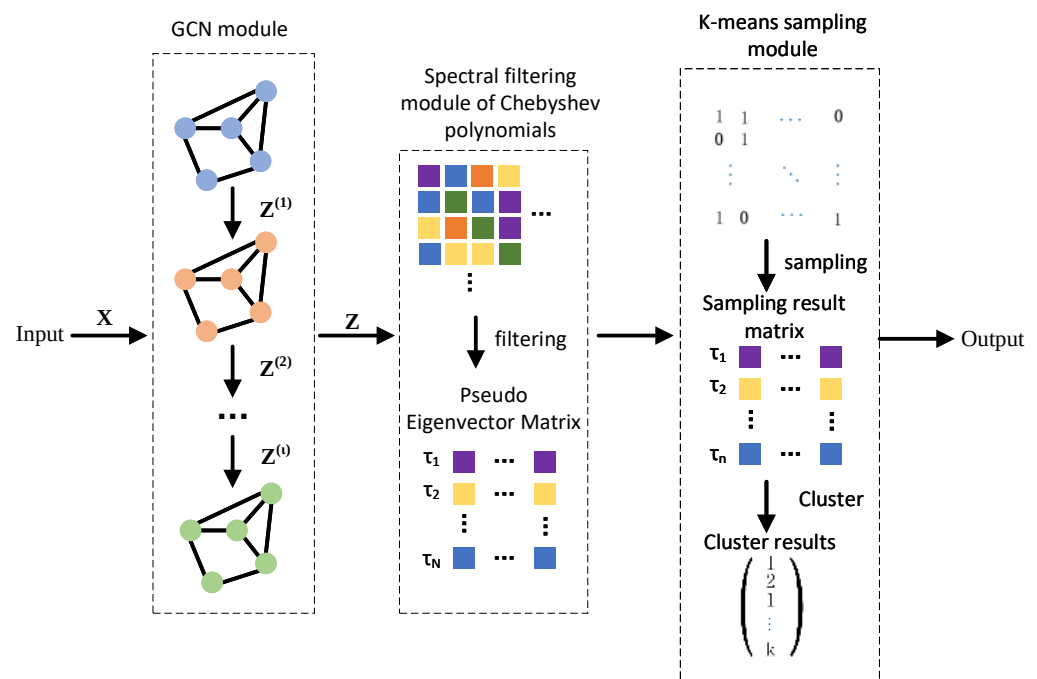


Figure 3. The GCN and spectral clustering framework diagram with graph filtering.

Graph Convolutional Network Module

Graph convolutional networks typically utilize graph structure and node feature information to iteratively aggregate the feature information of local graph neighbor nodes. Due to their advantages in combining graph structure and node content information, GCNs have been widely applied for feature extraction from non-Euclidean data.

The goal of GCN is to construct a weighted graph with the same number of nodes as the original graph. However, there may be different weights in the graph, encapsulating the topology of the original graph and node attributes. Let $A \in R^{N \times N}$ be the adjacency matrix and its node attribute matrix $X \in R^{N \times k}$, where k corresponds to the dimension of the node attribute vector.

Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes and $\mathcal{E}$ edges, let $A \in R^{N \times N}$ be the adjacency matrix, and its node attribute matrix $X \in R^{N \times k}$, where k corresponds to the dimension of the node attribute vector. GCN can be understood as a function $f(\cdot)$, and

its output is a weighted graph $\tilde{\mathcal{G}} = f(\mathcal{V}, \mathcal{E})$, represented by its adjacency matrix $\tilde{A} \in R^{N \times N}$, i.e., $\tilde{A} = f(A, X)$. For the learned representation weight matrix W in the l -th layer, Z^l can be obtained through the following convolution Equation (11):

$$Z^{(l)} = f\left(Z^{(l-1)}, A\right) = \varphi(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} Z^{(l-1)} W^{(l-1)}), \quad (11)$$

where $Z^{(l)}$ represents the l -th layer in GCN, denoted as $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. I is the identity matrix for each node's self-loop, and $\varphi(\cdot)$ is the non-linear activation function. The representation $Z^{(l-1)}$ traverses the normalized adjacency matrix $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$, resulting in a new representation Z^l . Finally, the ReLU function is applied Equation (12) in the last layer of GCN:

$$Z = f\left(Z^{(l)}, A\right) = \varphi(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} Z^{(l)} W^{(1)}), \quad (12)$$

where Z is a transformed node feature matrix, serving as both the output of GCN and the input to the spectral filtering module.

Chebyshev Polynomial Spectral Filtering Module

The Chebyshev polynomial approximation of the spectral filtering module achieves convolution in the spectral domain, similar to Fourier domain filtering of signals. The eigenvalues of the Laplacian matrix of the graph represent the frequencies of signals defined on the graph. Assuming o_i is the i -th variable of U , we obtain Equation (13).

$$o_i = |U^T \delta_i|^2, \delta_i(j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (13)$$

It satisfies all $o_i > 0$, in addition $O(i, i) = 1/o_i$ and $R = (r_1, r_2, \dots, r_s) \in R^{N \times s}$ which consists of s random signals. Let r be a vector with independent random Gaussian variables that follow a standard normal distribution. These vectors have a mean of 0 and a variance of $1/s$. Then, f_i is a feature vector, and the estimated value feature vector τ_i of each node that is very similar to f_i can be defined according to the following Equation (14):

$$\tau_i = (OH_{\lambda_k} R) \delta_i. \quad (14)$$

According to the above Equation (14), it is known that $H_{\lambda_k} = U_k U_k^T$, and it satisfies $U' = OU$. Based on the JL lemma and the U has orthogonal columns, we can obtain Equation (15).

$$|U(f_i - f_j)|^2 = |(f_i - f_j)|^2, \quad (15)$$

Let $r \in R^n$ be a vector with independent random terms and follows a standard normal distribution. Filtering r through H_{λ_k} , we obtain r_{λ_k} as shown in Equation (16). The i -th element of r_{λ_k} is Equation (17), and the mean of $(r_{\lambda_k})_i^2$ satisfies Equation (18). Therefore, a practical method to estimate $OH_{\lambda_k} R$ is to first compute $H_{\lambda_k} R$, and then normalize its rows to unit length.

$$r_{\lambda_k} = U \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k, 0, \dots, 0) U^T r = U U^T r \quad (16)$$

$$(r_{\lambda_k})_i = r_{\lambda_k}^T \delta_i = r^T U U^T \delta_i, \quad (17)$$

$$E(r_{\lambda_k})_i^2 = |U^T \delta_i|^2. \quad (18)$$

k-Means Sampling Module

The k-means sampling module implements spectral clustering on a small-scale data set, reducing the algorithm's time complexity. Let $\Omega = \{\tau_1, \tau_2, \dots, \tau_n\}$ represent the extracted n data points from the total N data points for clustering. Then, the k-means clustering algorithm is applied to this set, clustering it into k clusters. Considering that

k-means can also be applied to the original vector set $\{f_1, f_2, \dots, f_N\}$, a sampling matrix G can be formulated to satisfy $y_i^r = Gy_i$. G is an nN dimensional matrix. Ideally, G satisfies Equation (19). To recover y_i from the clustering results, a quadratic programming Equation (20) is utilized.

$$G_{ij} = \begin{cases} 1 & i = \tau_j \\ 0 & i \neq \tau_j \end{cases} \quad (19)$$

$$\min_{y_i \in R^N} \frac{1}{2} y_i^\top (G^\top G + \gamma g(L)) y_i - y_i^\top G^\top y_i^r. \quad (20)$$

Then, Equation (20) can be solved using Equation (21), where G is the sampling matrix, γ is the regularization parameter, and $g(L)$ is a positive non-decreasing polynomial:

$$(G^\top G + \gamma g(L)) y_i = G^\top y_i^r. \quad (21)$$

A method independent of the entire optimization formula is proposed for computing $g(L)$. Pseudo-eigenvectors replace first k eigenvectors, providing a high-order approximation and retaining distances between original eigenvectors. Pseudo-eigenvectors aid in reconstructing $g(L)$ by computing Euclidean distances between them. This reconstruction using pseudo-eigenvectors helps reduce the time complexity. The adjacency matrix used for reconstruction is defined as in Equation (22): The reconstructed Laplacian matrix L can be expressed as Equation (23):

$$\tilde{A}_{ij} = \exp \left(-\frac{q^2(Z_i - Z_j)}{\sigma_i \sigma_j} \right), \quad (22)$$

$$\tilde{L} = I - \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \quad (23)$$

where Z_i represents the i -th row of Z , and $q^2(Z_i - Z_j)$ represents the square of the Euclidean distance between Z_i and Z_j . σ_i represents the size of a range, denoted as $\sigma_i = q(Z_i, Z_j)$. $\tilde{D}$ retains the rows of $\tilde{A}$, and the reconstruction objective function is written as Equation (24):

$$\min_{y_i \in R^N} \|Gy_i - y_i^r\|_2^2 + \gamma y_i^\top \tilde{L} y_i. \quad (24)$$

This eliminates the step of computing $g(L)$ and reduce the time complexity of the optimization algorithm. The set of y_i obtained from Equation (24) represents complete clustering result $\{Y_1, Y_2, \dots, Y_k\}$.

3. Experiment

This study selected the power data of users in a certain region in one year as a case study to validate the effectiveness of the algorithm for integrating and deeply mining electricity information data in extracting typical load curves from different types of end-users. The dataset was obtained from the municipal-level electricity information collection system, including power data from 6000 users. The data were collected at a frequency of 72 time points per day, spanning 30 days, and totaling 180,000 load curves.

3.1. Integration of End-User Electricity Consumption Information Big Data

Firstly, the validation of the data anomaly identification method is conducted using one month of raw power data from the intelligent meters of end-users in the collection system. The validation environment includes a Win10 system, an i7 processor, and Python language version 3.8. The specific parameters for the Isolation Forest theory are designed as follows: the number of samples for training is set as $sample_size = 512$, the number of trees built for training is $n_trees = 200$, the true anomaly rate is 0.0163, and the initial value for the anomaly score is $desired_TPR = 0.81$.

To validate the Isolation Forest algorithm for identifying abnormal data and implementing data classification, typical binary classification evaluation metrics are employed [34]. This

includes constructing a confusion matrix to calculate the true positive rate (TPR), the false positive rate (FPR), the F1 score, and the average precision (avgPR). The specific results are illustrated in Figure 4.

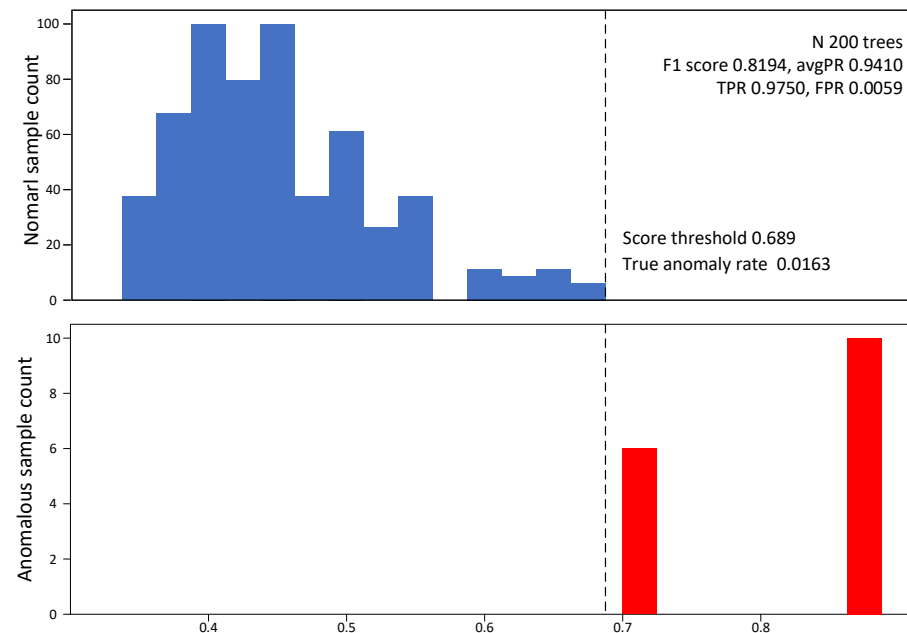


Figure 4. Abnormal value detection results.

Figure 4 displays the scoring situation of normal and abnormal data, demonstrating the effectiveness of the Isolation Forest algorithm in achieving binary classification of normal and abnormal values. The score threshold after training is 0.689. The TPR is 0.975, which correct identification rate of abnormal data, and FPR is 0.0059, indicating normal data being mistakenly classified as abnormal. The F1 score is 0.8194, and avgPR is 0.941, validating the effectiveness of the algorithm.

After identifying the anomalous data, the corresponding time coordinates can be obtained. By sampling 10 data points around each anomaly and inputting the data into the trained RBF neural network, the reconstructed normal data can be obtained. The results are shown in Figure 5a,b. Figure 5a shows the power values curve containing anomalies, and Figure 5b shows the power values curve after processing with the RBF neural network. It is evident that the anomalies have been reconstructed, achieving a smoothing effect on the user's power values curve.

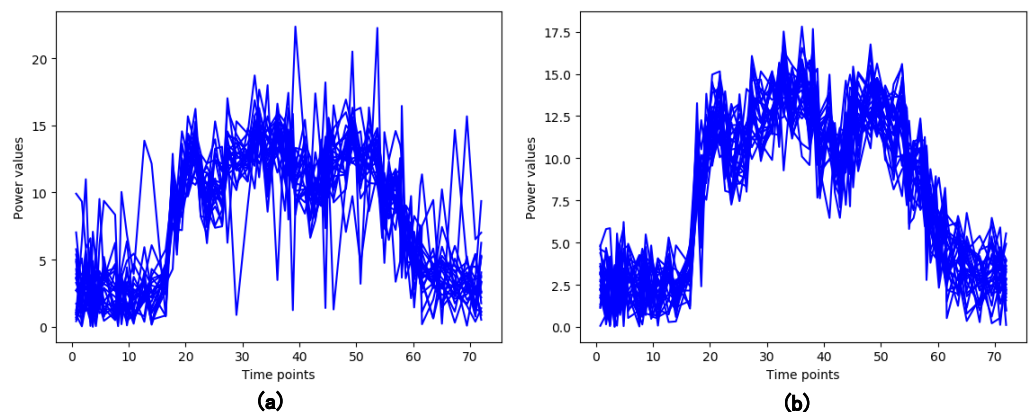


Figure 5. The results of load data processing. (a) Load data before abnormal data processing. (b) Load data after abnormal data processing.

3.2. Deep Mining of End-User Electricity Consumption Information Big Data

For residential users, the time interval for collecting load power is 20 min, resulting in 72 data points within a day's load curve. Figure 6 shows the load curve for a user on a given day. It can be seen from the figure that the user's load curve has many features and high dimension, which affects the accuracy and efficiency of clustering and extracting typical load curves for end users. To address this issue, the information entropy-based classification aggregation approximation algorithm mentioned earlier was utilized to reduce the dimensionality of load curves and achieve feature extraction.

Combining with the user power curve shown in Figure 6, the process of entropy-based segmented aggregation approximation is described. Initially, the maximum time resolution of the user time-series is 20 min, and the time-series has a period of 24 h. Based on the span of the user dataset sequence, the data dimension is determined to be 72 nodes. The initial number of segments is set to $k = 12$, with each segment containing 6 data points, and each segment's time-series span is $T = 120$ min.

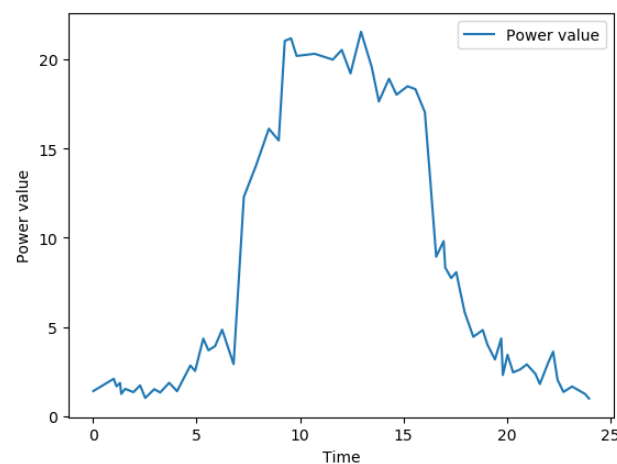


Figure 6. The load curve for a user on a given day.

Subsequently, for each segmented time-series, values are calculated, and a judgment of $\rho < \sigma$ is made. If this condition holds true, it indicates that the fluctuation level of the data sequence in that segment meets the requirements. The formula for segmented aggregation approximation is then applied to reduce the dimensionality and re-express the data in that segment. If it does not hold true, for segmented time-series where $\rho < \sigma$, if the length of the time-series can still be evenly divided by 2, a 2-segment division is performed, and the judgment is reevaluated. Otherwise, the time-series is output directly, and the result of the re-expression of each segmented time-series is saved.

Figure 7a shows the symbolic representation of the user curve using the typical segmented aggregation approximation algorithm [35], while Figure 7b presents the symbolic representation of the user curve using the proposed segmented aggregation approximation based on information entropy. By comparing the results in Figure 7a and Figure 7b, it can be observed that the proposed segmented aggregation approximation based on information entropy not only extracts features accurately but also preserves the original shape of the user load curve in detail. The algorithm uses different time resolutions for segmented aggregation approximation in regions with significant fluctuation trends, retaining central features and better re-expression the load curve.

Following the feature extraction of load curves using the information entropy-based segmented symbolic aggregation approximation method, an algorithm combining GCN and spectral clustering was employed for the unsupervised classification of end-user load curves. This process extracts typical daily load curves, and the clustering results are shown in Figure 8a–e.

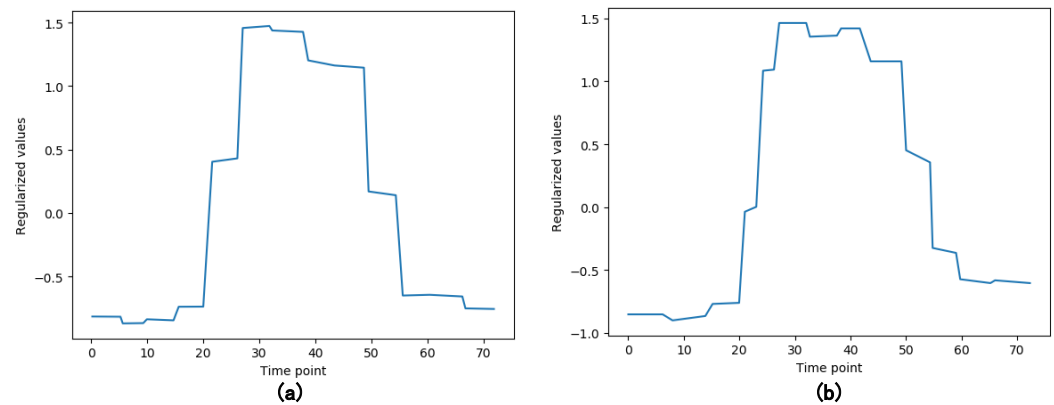


Figure 7. The results of symbolic re-expression of user curves. (a) The result of segment aggregation approximate, (b) The result of segmented aggregation approximation based on information entropy.

From Figure 8, it can be observed that the spectral clustering algorithm successfully achieves the unsupervised classification of users. In Figure 8a, the power curve for cluster one reaches its lowest point from 8:00 to 18:00 and peaks during 0:00–8:00 and 18:00–24:00, reflecting the nighttime electricity consumption characteristics of clustered users. This indicates that high-energy-consuming users adjust their production activities in response to time-of-use electricity prices. In Figure 8b, cluster two's power curve peaks from 8:00 to 18:00, while maintaining low consumption at other times, revealing the daytime production-oriented characteristics of users, primarily businesses, technology companies, and material processing factories operating during the day. Figure 8c,d show that users in clusters three and four exhibit a bimodal pattern. Cluster three reaches its peak from 19:00 to 21:00, while cluster four peaks during 9:00 to 11:00 and again in the afternoon from 15:00 to 17:00, aligning well with typical user electricity consumption patterns. In Figure 8e, cluster five shows a relatively uniform daily load curve with no distinct features, mainly including users such as plastic factories, glass processing plants, and telecommunication companies. This clustering result demonstrates the effectiveness of the spectral clustering algorithm in obtaining typical user electricity consumption behaviors, clustering users based on their spatiotemporal distribution characteristics, and deeply mining data information in various user categories in the electricity market.

To effectively evaluate the experimental results, common evaluation metrics including accuracy (ACC), normalized mutual information (NMI), and adjusted Rand index (ARI) were utilized. The proposed algorithm in this paper was compared with other algorithms using these metrics. The other algorithms include the following. SC: The main idea of spectral clustering is to treat samples as points in space, with lower weights between points that are distant and higher weights between points that are close. It involves cutting the graph formed by all samples to maximize edge weights within clusters [36]. LSC-K: Large-scale spectral clustering based on label representation [37]. CKM: Large-scale clustering using K-means algorithm [38]. LSR: Subspace clustering algorithm utilizing least squares regression [39]. The experimental results of each algorithm are listed in Table 1, and Figure 9 shows the computational time required for the corresponding algorithm on the data set. Among them, the SC algorithm has the worst experimental results because its main idea is to regard data as points in space, which is poorly adapted to the power data in the experiment. At the same time, its calculation is relatively complex and time-consuming. The CKM algorithm reduces computation time for large-scale clustering by using K-means, achieving good results in NMI but needing improvement in ACC and ARI. The LSR algorithm achieves good results across all metrics but takes a longer computation time. The proposed algorithm improves ACC and ARI by 3.69% and 2.6%, respectively, compared to the next best LSR algorithm, and NMI by 3.14% compared to CKM, demonstrating the best clustering performance. Moreover, the best result was achieved in calculation time, which

was 8 s faster than the CKM algorithm. These experimental results confirm the feasibility and superiority of the proposed method, as well as its efficient computation speed.

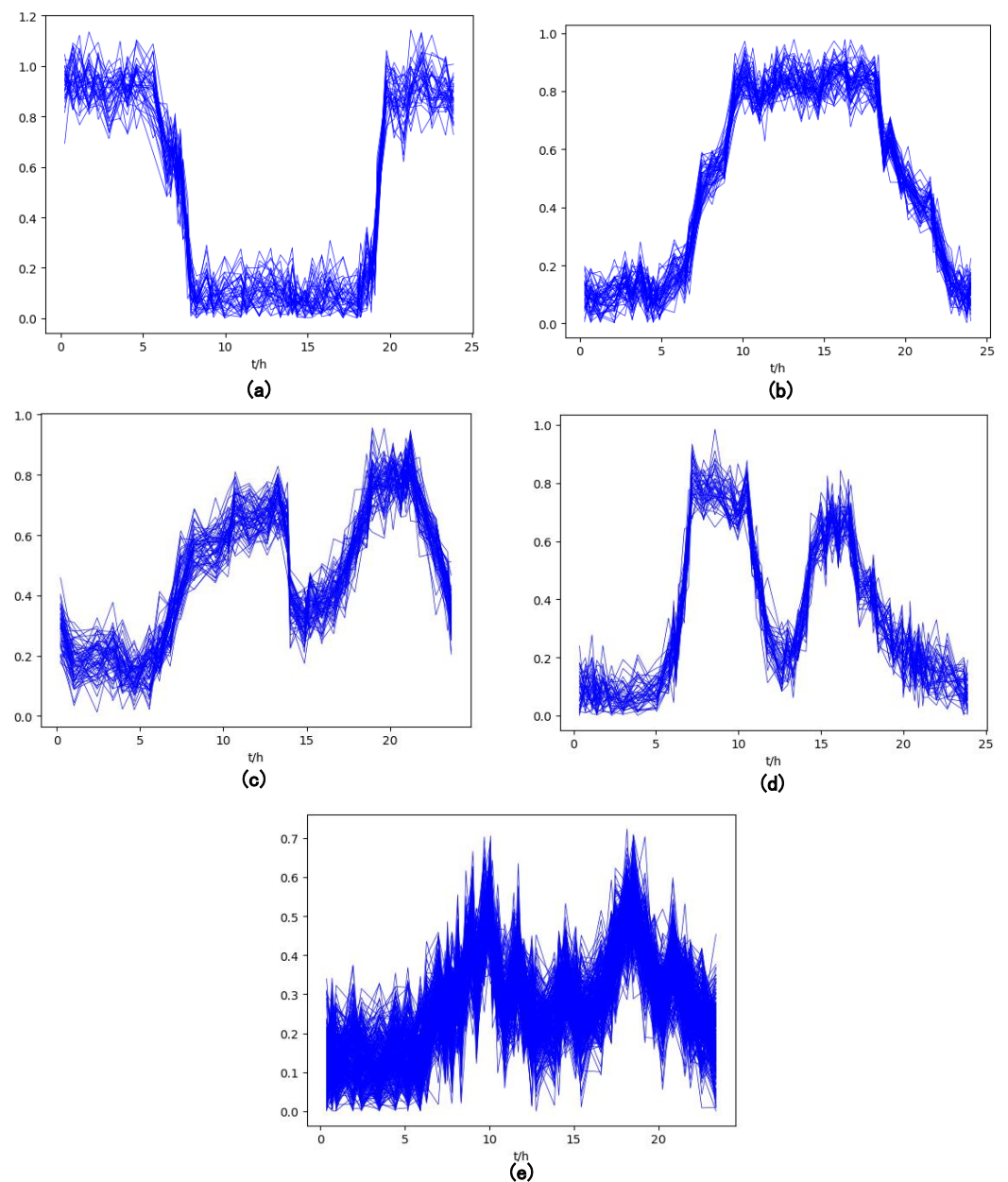


Figure 8. The results of class cluster. (a) Class cluster 1 result; (b) class cluster 2 result; (c) class cluster 3 result; (d) class cluster 4 result; (e) class cluster 5 result.

Table 1. The ACC, ARI, and NMI values of different algorithms.

Algorithm	ACC	NMI	ARI
SC	65.32	63.34	55.32
LSC-K	74.31	71.28	60.05
CKM	73.54	75.10	71.78
LSR	80.82	74.23	73.53
Ours	84.51	78.24	76.13

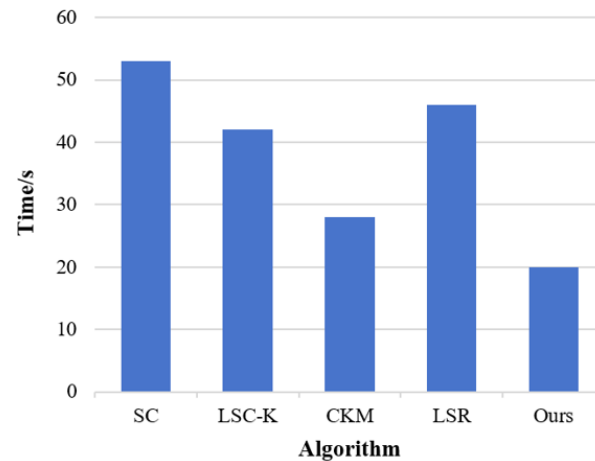


Figure 9. The computational time of different algorithms.

4. Discussion

4.1. Advantages of the Proposed Method

In our research on power data analysis, we innovatively combine the isolation forest algorithm with the radial basis function (RBF) neural network for data fusion, integrating spatial and temporal data attributes. This allows for a more comprehensive analysis of different users' electricity usage patterns. Additionally, we introduce a graph convolutional network (GCN) and spectral clustering framework for dimensionality reduction and feature extraction of power data. This method effectively captures and analyzes the non-Euclidean structured data commonly present in power distribution systems. By leveraging the unique advantages of GCN in learning data topology, combined with spectral clustering, our approach can identify subtle relationships and patterns in electricity usage that traditional analysis methods cannot detect. This enables us to more effectively handle high-dimensional, multi-source datasets.

In this study, the Isolation Forest algorithm has proven to be efficient in detecting anomalies in end-user electricity consumption data, with linear time complexity and low memory usage, and the detection results of the algorithm are quite accurate. The algorithm achieves this by constructing a tree structure to isolate abnormal data from normal data, effectively identifying potential anomalies in power consumption. Furthermore, integrating the RBF neural network during the interpolation process after anomaly detection has proven feasible, enhancing the accuracy and reliability of data reconstruction. It adeptly handles missing or abnormal data, contributing to smoother and more accurate data processing, thereby establishing a more reliable foundation for subsequent analysis.

The method involving segmented aggregation and approximate feature extraction based on information entropy achieves significant dimensionality reduction in high-dimensional data. This approach not only minimizes redundant information but also preserves key electricity consumption characteristics, offering more informative data for subsequent analysis. The GCN utilized in the method proves effective in learning the relationships between nodes in the graph structure. This capability enables a better understanding of the similarity and dissimilarity between different users in the clustering of user load curves. The spectral clustering algorithm performs well in extracting typical load curves of users. It proficiently clusters similar load curves, providing robust support for power system optimization and the quantification of user demand.

Overall, the combination of isolation forest, RBF neural network, segmented aggregation, approximate feature extraction, GCN, and spectral clustering contributes to a comprehensive and effective approach for analyzing end-user electricity consumption data, enhancing anomaly detection, and supporting power system optimization efforts.

4.2. Limitations and Future Research Directions

While the isolation forest excels in anomaly localization, it may face challenges in handling anomalies within complex data distributions and multidimensional relationships. Additionally, the RBF neural network, known for its smooth interpolation, may encounter overfitting in extreme cases, necessitating careful parameter selection for practical applications to balance accuracy and stability. The GCN algorithm within the method exhibits sensitivity to changes in the graph structure, requiring retraining or parameter adjustments when significant changes occur. The spectral clustering algorithm, though effective, presents high computational complexity when processing large-scale data, demanding efficient computing resources.

Future research avenues can concentrate on optimizing and integrating the algorithm, refining parameters of the isolation forest and RBF neural network to enhance overall applicability and robustness. The dynamic adaptability of the graph structure in GCN can be improved to better handle the dynamic changes in the power system's graph structure, ensuring real-time adaptability to the complexity of power system operations. Consideration may be given to integrating multiple algorithms to address more complex power data analysis scenarios. Attention to real-time and responsive application of algorithms is crucial, enhancing real-time power information processing capabilities for practical implementation. This improvement aims to make the technology more applicable and supportive of real-time monitoring and adjustment of power systems.

5. Conclusions

This study addresses the challenge of fluctuating end-user power supply demand within the context of electricity market reform. Leveraging the distributed Internet of Things, along with sensing, data fusion, and intelligent application functions, the study extensively explores methods for collecting end-user electricity demand.

The paper employs the isolation forest algorithm and RBF neural network algorithm for preprocessing and fusion, achieving data cleaning and mining to overcome redundancy and outliers in load curve data across time and space dimensions. Additionally, the study investigates segmented aggregation approximation based on information entropy for efficient feature extraction and mining of big data. The combination of GCN and spectral clustering is used to comprehensively analyze end-user power consumption data, enhancing anomaly detection and supporting optimization efforts within the power system. The effectiveness of the proposed data fusion and mining method for different types of end-user power data are validated through experiments.

In summary, the main contribution of this study lies in integrating distribution Internet of Things technology and utilizing advanced data processing and mining methods for in-depth analysis of end-user electricity consumption data in the electricity market. Future research directions may involve further algorithm optimization, improvements in large-scale data processing, and enhancements in real-time response capabilities to meet the increasing demands for power data analysis and adapt to the continuous evolution of the power market.

Author Contributions: Conceptualization, L.H.; methodology, L.H. and M.S.; software, L.H.; validation, L.M.; formal analysis, M.S.; investigation, M.S. and L.M.; resources, L.H.; data curation, L.H.; writing—original draft preparation, L.H. and M.S.; writing—review and editing, L.M.; visualization, M.S.; supervision, L.W.; project administration, L.W.; funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the Science and Technology Project of SGCC, named Research on Transaction Data Sharing and Integration Service Support Technology for Spatiotemporal Multiscale Analysis of the Electricity Market (5108-202355066A-1-1-ZN).

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and the code of this study are available from the corresponding author upon request.

Conflicts of Interest: Authors Longda Huang and Maohua Shan were employed by the company China Electric Power Research Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Wang, J.; Meng, K.; Cao, J.; Chen, Z.; Gao, L.; Lin, C. A review of energy internet information technology research. *Comput. Res. Dev.* **2015**, 117–134.
- Ma, R.; Zhu, D.; Xia, X.; Liu, J.; Sha, J. Energy Big Data Storage and Parallel Processing Method Based on ODPs. In Proceedings of the 2021 International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy: SPIoT-2021, online, 30 October 2021; Springer: Berlin/Heidelberg, Germany, 2022; Volume 1, pp. 543–548.
- Koot, M.; Wijnhoven, F. Usage impact on data center electricity needs: A system dynamic forecasting model. *Appl. Energy* **2021**, 291, 116798. [\[CrossRef\]](#)
- Wang, J.; Ji, Z.; Shi, M.; Huang, F.; Zhu, C.; Zhang, D. Analysis and Application Research of Big Data Demand for Intelligent Power Distribution. *Proc. CSEE* **2015**, 35, 1829–1836.
- Boyd, J. An internet-inspired electricity grid. *IEEE Spectr.* **2012**, 50, 12–14. [\[CrossRef\]](#)
- Weinzettel, J.; Havránek, M.; Ščasný, M. A consumption-based indicator of the external costs of electricity. *Ecol. Indic.* **2012**, 17, 68–76. [\[CrossRef\]](#)
- Zhang, S.; Dai, H.; Shi, Z.; Yang, A.; Yan, H. The Construction of Electric Power Big Data Analysis Platform Prospects in Smart Grid Application. In Proceedings of the 2020 IEEE/IAS Industrial and Commercial Power System Asia (I&CPS Asia), Weihai, China, 13–15 July 2020; pp. 1745–1750.
- Jiang, H.; Wang, K.; Wang, Y.; Gao, M.; Zhang, Y. Energy big data: A survey. *IEEE Access* **2016**, 4, 3844–3861. [\[CrossRef\]](#)
- Teng, Z.; Yan, Z.; Dongxia, Z. Big data application technology and prospect analysis of smart distribution network. *Power Syst. Technol.* **2014**, 38, 3305–3312.
- Park, C.; Heo, W. Review of the changing electricity industry value chain in the ICT convergence era. *J. Clean. Prod.* **2020**, 258, 120743. [\[CrossRef\]](#)
- Janssen, M.; Van Der Voort, H.; Wahyudi, A. Factors influencing big data decision-making quality. *J. Bus. Res.* **2017**, 70, 338–345. [\[CrossRef\]](#)
- Lin, J.; Sheng, G.; Yan, Y.; Zhang, Q.; Jiang, X. Online monitoring data cleaning of transformer considering time series correlation. In Proceedings of the 2018 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), Denver, CO, USA, 16–19 April 2018; pp. 1–9.
- Lv, Z.; Deng, W.; Zhang, Z.; Guo, N.; Yan, G. A data fusion and data cleaning system for smart grids big data. In Proceedings of the 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Xiamen, China, 16–18 December 2019; pp. 802–807.
- Lin, S.; Tian, E.; Fu, Y.; Tang, X.; Li, D.; Wang, Q. Power load classification method based on information entropy piecewise aggregate approximation and spectral clustering. *Proc. CSEE* **2017**, 37, 2242–2252.
- Mao, W.; Cao, X.; Yan, T.; Zhang, Y. Anomaly detection for power consumption data based on isolated forest. In Proceedings of the 2018 International Conference on Power System Technology (POWERCON), Guangzhou, China, 6–8 November 2018; pp. 4169–4174.
- Liu, Y.; Xu, L. A high performance extraction method for massive user load typical characteristics considering data class imbalance. *Proc. CSEE* **2019**, 39, 4093–4103.
- Shao, Y.; Li, H.; Gu, X.; Yin, H.; Li, Y.; Miao, X.; Zhang, W.; Cui, B.; Chen, L. Distributed graph neural network training: A survey. *ACM Comput. Surv.* **2024**, 56, 1–39. [\[CrossRef\]](#)
- Waikhom, L.; Patgiri, R. A survey of graph neural networks in various learning paradigms: Methods, applications, and challenges. *Artif. Intell. Rev.* **2023**, 56, 6295–6364. [\[CrossRef\]](#)
- Zhang, W.; Wang, X.; Zhao, D.; Tang, X. Graph degree linkage: Agglomerative clustering on a directed graph. In Proceedings of the Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Proceedings, Part I 12; Springer: Berlin/Heidelberg, Germany, 2012; pp. 428–441.
- Zhang, W.; Zhao, D.; Wang, X. Agglomerative clustering via maximum incremental path integral. *Pattern Recognit.* **2013**, 46, 3056–3065. [\[CrossRef\]](#)
- Barton, T.; Bruna, T.; Kordik, P. Chameleon 2: An improved graph-based clustering algorithm. *ACM Trans. Knowl. Discov. Data (TKDD)* **2019**, 13, 10. [\[CrossRef\]](#)
- Donath, W.E.; Hoffman, A.J. Lower bounds for the partitioning of graphs. *IBM J. Res. Dev.* **1973**, 17, 420–425. [\[CrossRef\]](#)
- Fiedler, M. Algebraic connectivity of graphs. *Czechoslov. Math. J.* **1973**, 23, 298–305. [\[CrossRef\]](#)

24. Hagen, L.; Kahng, A.B. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **1992**, *11*, 1074–1085. [\[CrossRef\]](#)
25. Darudi, A.; Bashari, M.; Javidi, M.H. Electricity price forecasting using a new data fusion algorithm. *IET Gener. Transm. Distrib.* **2015**, *9*, 1382–1390. [\[CrossRef\]](#)
26. Buschjäger, S.; Honysz, P.J.; Morik, K. Generalized isolation forest: Some theory and more applications extended abstract. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, NSW, Australia, 6–9 October 2020; pp. 793–794.
27. dos Santos Coelho, L.; Santos, A.A. A RBF neural network model with GARCH errors: Application to electricity price forecasting. *Electr. Power Syst. Res.* **2011**, *81*, 74–83. [\[CrossRef\]](#)
28. Jiang, S.; Dong, R.; Wang, J.; Xia, M. Credit card fraud detection based on unsupervised attentional anomaly detection network. *Systems* **2023**, *11*, 305. [\[CrossRef\]](#)
29. Jiang, Y.; Liu, C.C.; Diedesch, M.; Lee, E.; Srivastava, A.K. Outage management of distribution systems incorporating information from smart meters. *IEEE Trans. Power Syst.* **2015**, *31*, 4144–4154. [\[CrossRef\]](#)
30. Chakravorti, T.; Patnaik, R.K.; Dash, P.K. Detection and classification of islanding and power quality disturbances in microgrid using hybrid signal processing and data mining techniques. *IET Signal Process.* **2018**, *12*, 82–94. [\[CrossRef\]](#)
31. Zhu, X.; Zhang, S.; Li, Y.; Zhang, J.; Yang, L.; Fang, Y. Low-rank sparse subspace for spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1532–1543. [\[CrossRef\]](#)
32. Zhong, G.; Pun, C.M. Self-taught multi-view spectral clustering. *Pattern Recognit.* **2023**, *138*, 109349. [\[CrossRef\]](#)
33. Tang, C.; Li, Z.; Wang, J.; Liu, X.; Zhang, W.; Zhu, E. Unified one-step multi-view spectral clustering. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 6449–6460. [\[CrossRef\]](#)
34. Canbek, G.; Sagioglu, S.; Temizel, T.T.; Baykal, N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–8 October 2017; pp. 821–826.
35. Guo, C.; Li, H.; Pan, D. An improved piecewise aggregate approximation based on statistical features for time-series mining. In Proceedings of the Knowledge Science, Engineering and Management: 4th International Conference, KSEM 2010, Belfast, Northern Ireland, UK, 1–3 September 2010; Proceedings 4; Springer: Berlin/Heidelberg, Germany, 2010; pp. 234–244.
36. Xiong, K.; Wang, S. The online random Fourier features conjugate gradient algorithm. *IEEE Signal Process. Lett.* **2019**, *26*, 740–744. [\[CrossRef\]](#)
37. Wang, Y.; Singh, A. Provably correct algorithms for matrix column subset selection with selectively sampled data. *J. Mach. Learn. Res.* **2018**, *18*, 1–42.
38. Marques, A.G.; Segarra, S.; Leus, G.; Ribeiro, A. Sampling of graph signals with successive local aggregations. *IEEE Trans. Signal Process.* **2015**, *64*, 1832–1843. [\[CrossRef\]](#)
39. Hensman, J.; Durrande, N.; Solin, A. Variational Fourier features for Gaussian processes. *J. Mach. Learn. Res.* **2018**, *18*, 1–52.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.