Energy Informatics

RESEARCH Open Access

Automatic grid topology detection method based on Lasso algorithm and t-SNE algorithm

Sheng Huang[1], Huakun Que[1], Yingnan Zhang[2], Tenglong Xie[3]* and Jie Peng[4]

*Correspondence:
Tenglong Xie
flydragon2006@126.com
[1] Measurement Center of Guangdong Power Grid Co., Ltd., Guangzhou 511545, China
[2] China Southern Digital Power Grid Group Co., Ltd., Guangzhou 510700, China
[3] Heyuan Power Supply Bureau, Guangdong Power Grid Co., Ltd., Heyuan 517000, China
[4] Guandong Power Grid Co., Ltd., Guangzhou 510030, China

Abstract
For a long time, the low-voltage distribution network has the problems of untimely management and complex and frequently changing lines, which makes the problem of missing grid topology information increasingly serious. This study proposes an automatic grid topology detection model based on lasso algorithm and t-distributed random neighbor embedding algorithm. The model identifies the household-variable relationship through the lasso algorithm, and then identifies the grid topology of the station area through the t-distributed random neighbor embedding algorithm model. The experimental results indicated that the lasso algorithm, the constant least squares algorithm and the ridge regression algorithm had accuracies of 0.88, 0.80, and 0.71 and loss function values of 0.14, 0.20, and 0.25 for dataset sizes up to 500. Comparing the time spent on identifying household changes in different regions, in Region 1, the training time for the Lasso algorithm, the Constant Least Squares algorithm, and the Ridge Regression algorithm is 2.8 s, 3.0 s, and 3.1 s, respectively. The training time in region 2 is 2.4s, 3.6s, and 3.4s, respectively. The training time in region 3 is 7.7 s, 1.9 s, and 2.8 s, respectively. The training time in region 4 is 3.1 s, 3.6 s, and 3.3 s, respectively. The findings demonstrate that the suggested algorithmic model performs better than the other and can identify the structure of LV distribution networks.

Keywords Power grid topology, Lasso algorithm, Household variable identification, T-SNE algorithm, Big data

Introduction

The electric power system includes the low-voltage distribution network (LVDN), which is in charge of transferring electricity from substations to final consumers. Low-voltage distribution usually includes 220 V cables, switchboards, switchgear, etc., which is used to distribute the power transmitted by high-voltage transmission lines to the users, and is the last link of the power supply service, and its daily management and maintenance level directly affects the users' experience and satisfaction with the power (Mehrim 2022). However, for a long time, LVDN lines are difficult to effectively control, and there are problems such as frequent line changes. The traditional maintenance of the topological relationship of the station area requires staff to go to the site for data collection and troubleshooting, which makes the efficiency low (Ma et al. 2023). Due to the lack of a correct structure of the station topology map, the management of power lines appears to be extra difficult, such as fault detection, three-phase unbalance and line loss reduction. The LVDN has the problem of untimely management and complex and frequently changing lines, which makes the problem of missing topology information in the power grid increasingly serious. Therefore, this study proposes an automatic grid topology detection model based on Lasso algorithm and t-SNE algorithm, which identifies the household-variable relationship by Lasso algorithm, and then identifies the topology of the station grid by t-SNE algorithm model. Intended to achieve more refined management of the power grid and provide a more efficient method for its management. The study is structured into four primary sections. The first section provides a synopsis of the grid structure research issues of other academics. The second part is an overview of the algorithms mainly used in this research, and the third part is the model results obtained by applying the algorithms to the research and analyzing the results. An overview of the previous research and a plan for further research are presented in the fourth section.

Related works

LVDN is a key component of the power grid that helps link users to it. A novel transformerless interconnected cascaded multistage architecture was proposed by Hock and Batschauer to enhance the low- and medium-voltage distribution networks' ability to withstand unbalanced nonlinear loads. Based on a single star bridge cell with three star-connected inductors to interchange active power (AP), the structure was constructed. Based on experimental results, it was shown that the proposed topology, when paired with the control system, could effectively minimize load disturbances, boost regulation capabilities, and regulate grid current and DC voltage (Hock and Batschauer 2022). To make the best use of the current grid structure and maximize the integration of renewable energy sources, Miller et al. suggested a time series based planning method for high voltage distribution networks. The method determined line load indicators based on line overloads and losses, and automatically implemented grid reinforcement and expansion measures to meet security requirements. At the same time, the grid topology information was considered to realize efficient expansion. The outcomes revealed that the ability of the method has high efficiency (Miller et al. 2022). Paul et al. proposed a new transformerless microinverter topology. Simulation and experimental studies verified the effectiveness and feasibility of the scheme (Paul et al. 2023). For transmission system parameter identification, Xia et al. devised a novel multi-scale folded attention map convolutional network to get over the drawbacks of conventional deep learning techniques. The network created a NestedTask Block loss function to balance the parameters of large and small targets, and it employed U-block to sample the multi-scale data. The multi-scale folded attention map convolutional network outperforms state-of-the-art techniques in recognition, according to experimental results (Xia et al. 2022).

To overcome the scalability limitations of t-SNE when dealing with large datasets, Henrique et al. proposed a new model by embedding multidimensional data points in 3D space and optimizing the memory access strategy and utilizing acceleration techniques to improve the model. According to the results, the design showed promise for use in a multi-core processor environment and increased execution performance by up to 460% when compared to the conventional t-SNE model (Henrique et al. 2021). Liu et al. introduced the t-distributed stochastic neighborhood embedding method in order to assist the clustering analysis of groundwater geochemical data. The findings demonstrated that t-SNE is a viable auxiliary technique since it is more effective than principal component analysis at identifying the number of clusters and defining spatial bands of groundwater geochemistry (Liu et al. 2021). To differentiate type 2 active nuclei from H ii galaxies in BPT maps, Zhang et al. developed an t-SNE technique applied to localized narrow emission line galaxies. The outcomes showed that, in 2D projection maps, t-SNE can distinguish between type 2 active galactic nuclei and H ii galaxies with great clarity. It can also mathematically ascertain the demarcation line of the BPT map, which precisely aligns the theoretical expectation with the factual observation and offers a useful technique for the classification of composite galaxies (Zhang et al. 2020).

In summary, many scholars have studied the topology structure of power grids and achieved certain results, but no one has introduced the Lasso algorithm into the identification of power grid topology structure. Moreover, research may focus too much on a specific aspect of the problem, while neglecting the comprehensive consideration of the overall system. Due to limitations in the size or quality of the dataset, the proposed algorithm or model may have poor performance or low efficiency. This has led to a certain impact on the credibility of the experimental results. This study proposes an automatic grid topology detection model based on Lasso algorithm and t-SNE algorithm, which recognizes the household-variable relationship by Lasso algorithm, and then recognizes the PGT of the station area by t-SNE algorithm model.

PGT study based on Lasso algorithm and t-SNE algorithm

The study proposes a Lasso algorithm based grid household variable relationship identification model, which recognizes the household variable relationship through Lasso algorithm. Then a PGT identification model based on Lasso algorithm with t-SNE algorithm is proposed, which identifies the topology of the station grid through t-SNE algorithm model.

Household change relationship identification model based on Lasso regression modeling

Metering automation system is a system that utilizes computer technology, sensors, actuators and other equipment to automate the control and monitoring of the measurement process. Through this system, the user's information on various electricity consumption links is collected. Such as low-voltage user centralized copying, distribution substation metering monitoring, and dedicated substation load management and other links (Lambert et al. 2022). Among them, distribution transformer metering and monitoring refers to the process of power metering and monitoring of distribution transformers. The power system's distribution transformer, which converts electrical energy from high-voltage transmission lines into electrical energy appropriate for low-voltage customers, is a crucial component. Distribution transformer monitoring and metering is required to ensure accurate metering of electrical energy and reliable operation of the distribution system. Distribution substation monitoring terminals are generally connected to the main station through a wireless communication channel to interact with the data. Low-voltage centralized copying system is a kind of system used for data collection and monitoring of low-voltage power users, which is usually applied to the scope of low-voltage power grids such as cities and neighborhoods (Xu et al. 2023a, b). The architecture of this type of system mainly includes components such as hardware devices, communication networks, data acquisition and processing software, and its structure is shown in Fig. 1.

In Fig. 1, the low-voltage centralized metering system consists of three main parts, namely, the master station, the concentrator and the collector. The main station is the core management center of the low-voltage meter reading system, usually located in the power operation management department or data center. The main station is responsible for monitoring, managing, and controlling the operation of the entire low-voltage collection and reading system. Establish communication connections between the main station, concentrator, and collector, receive data from the collector, process the data, and generate corresponding reports, analysis results, or operation instructions. The main station can also perform system configuration, fault diagnosis, and remote control operations. A concentrator is an intermediate node located between the collector and the main station, responsible for receiving data from the collector and transmitting it to the main station for processing and management. Concentrators are usually deployed within the coverage range of the power grid, with strong communication and processing capabilities, able to support communication with multiple collectors, and achieve centralized data aggregation and transmission. A collector is a device installed in a low-voltage power grid, used to monitor and collect real-time data information of the power system,

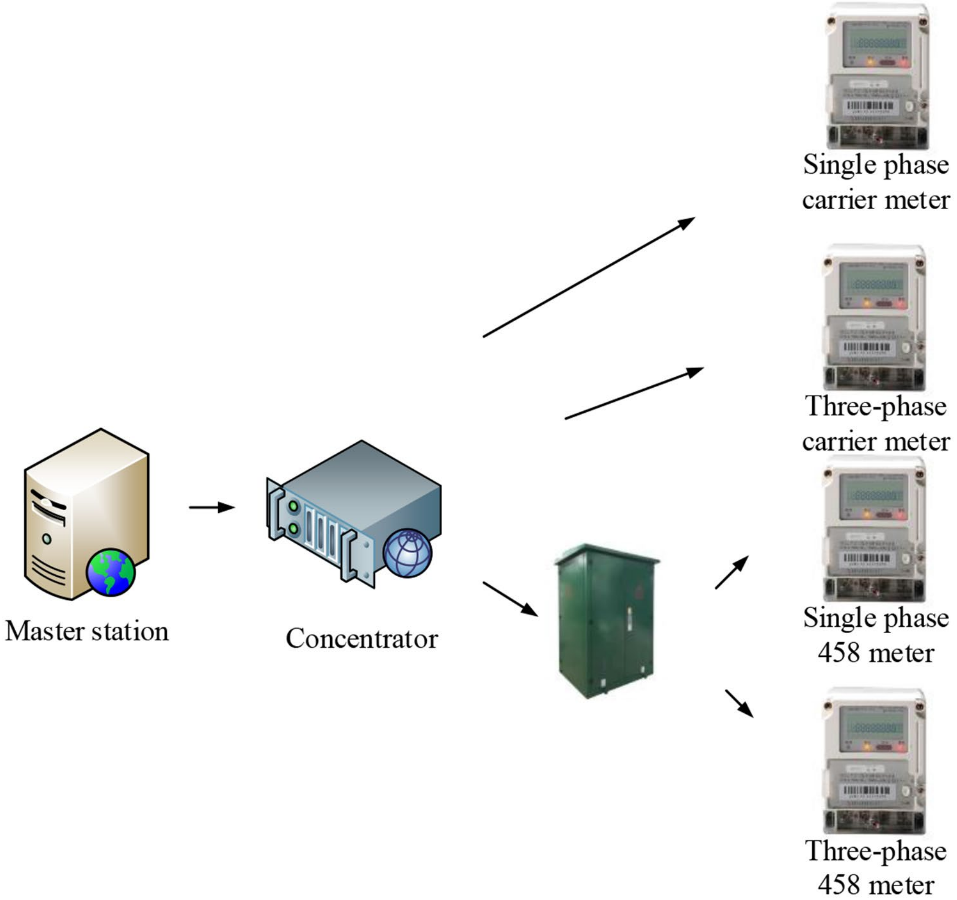

Fig. 1 Structure of low voltage collecting and reading system

such as current, voltage, power, energy consumption, etc. Collectors are usually installed on key nodes or equipment in the power grid, connected to power equipment to collect real-time data and transmit it to the concentrator or main station. The deployment and configuration of the collector can be adjusted and optimized according to specific monitoring needs and system architecture to achieve real-time monitoring and management of the power grid operation status. Existing grid companies are able to collect grid data on the transformer side and the customer side through metering automation systems [12–13]. The traditional identification algorithm based on the similarity of voltage series has misjudgment, but the current data of the transformer low voltage side and the user side have a logical relationship, however, the synchronization of the time of the data usually collected has a big problem, and it cannot reflect the current load situation at a certain moment. Equation (1) expresses the law of energy conservation's formula, which states that a station's daily power supply is equal to the total daily power consumption of all of its users.

$$y_d = \sum_{i=1}^{n} \alpha_i x_i(d) + \varepsilon_d \tag{1}$$

In Eq. (1), $x_i(d)$ is the amount of electricity on day d of the i th customer meter in the station to be identified. y denotes the total collected electricity and n denotes the total meters. m denotes the total amount of data and ε_d denotes the error variable, which mainly has meter accuracy as well as time error. α_i denotes the regression coefficient to be solved, and the relationship is shown in Eq. (2).

$$\alpha_i = \begin{cases} 1, Right \\ 0, wrong \end{cases} \tag{2}$$

When α_i in Eq. (2) equals 0, it indicates that the user does not belong to the identified station area and the household-transformer relationship is inconsistent; on the other hand, when α_i equals 1, it indicates that the user is within the identified station area for the power supply and the household-transformer relationship is accurate. Since in the real situation, the line of the station area has loss problems, the larger the electrical load and the farther the load point is from the transformer, the greater the loss to the station area (Luo et al. 2022). Therefore the household variable linear model is shown in Eq. (3).

$$Y = X\beta + \varepsilon \tag{3}$$

In Eq. (3), X denotes the user power matrix and Y denotes the column vector of the total meter power of the station. β denotes the column vector of regression coefficients and ε denotes the column vector of errors. The optimization objective function of linear regression is to minimize the mean square deviation and its expression is shown in Eq. (4).

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 \tag{4}$$

In Eq. (4), $\hat{\beta}$ denotes the estimated regression coefficient, Y is the dependent variable, and X is the independent variable. β denotes the actual regression coefficient, and $\|\cdot\|$ denotes the L2 paradigm. If the regression coefficient is found to be close to 0, it is a meter that does not belong to that station area. The algorithm is belongs to the least squares problem, and the advantage of the least squares method is that the principle is very simple and easy to understand. And under certain conditions, the least squares method can guarantee the uniqueness and existence of parameters. However, the scope of application is limited to meet the needs of actual data analysis (Xu et al. 2023a, b). Therefore, the problem can be solved by Lasso regression. Lasso algorithm is a regularization algorithm for linear regression and related problems. The Lasso algorithm can automatically select the most important features by adding L1 norm penalty to the objective function, promoting sparsity of model coefficients and compressing irrelevant or redundant feature coefficients to zero. Due to the Lasso algorithm's ability to compress the coefficients of certain features to zero, the final model is more concise and easy to interpret. This allows for a clearer understanding of the impact of the model on the prediction results, thereby providing more convincing explanations for decision-making. The Lasso algorithm can to some extent handle the problem of multicollinearity, that is, when there is strong correlation between features, the Lasso algorithm can effectively select one of the features and compress the coefficients of other related features to zero, thereby reducing the risk of overfitting in the model. The Lasso algorithm introduces L1 norm penalty during the fitting process, which makes the coefficients of the model more stable. Compared to ordinary least squares linear regression, Lasso algorithm has a certain anti-interference ability for noise and outliers in the data. The algorithm introduces the L1 regularization term on the basis of the least squares method, and achieves parameter estimation by minimizing the objective function, and its objective optimization is shown in Eq. (5).

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{5}$$

In Eq. (5), λ represents the penalty coefficient, and the size of the penalty coefficient affects the number of non-zero coefficients in the regression coefficients. The penalty term has less of an effect when λ is small, and it will have less of an impact on the coefficients in β when λ is high. Therefore, the choice of the penalty coefficient is more critical. The regression coefficients of the Lasso function are updated by the coordinate descent method, and the obtained linear model is fitted by the regularization algorithm. Equation (6) provides an expression for the prediction error, which is used to assess the correctness of the fitted model.

$$PE_{\lambda} = E\,\|Y_0 - \hat{\eta}_{\lambda}(X_0)\|_2^2 \tag{6}$$

In Eq. (6), $E(\cdot)$ represents the expectation of the corresponding sequence. Different λ correspond to different errors. The accuracy of the model is tested by K-fold cross-validation. Cross-validation is a statistical method used to estimate the performance of machine learning models. It involves partitioning the data into subsets, training the model on some subsets (training set), and validating it on others (validation set). This helps prevent overfitting and ensures the model generalizes well to new data. The expression of which is shown in Eq. (7).

$$CV_{\lambda}(K) = \frac{1}{K}\sum_{k=1}^{K}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\left(Y_i - \hat{\eta}_{\lambda}^{-k}(X_i)\right)^2\right] \tag{7}$$

In Eq. (7), the penalty coefficient λ selects the value of λ with the optimal average prediction error, and the household variable identification (HVI) process based on the total meter of the station and the user's metered power is shown in Fig. 2.

In Fig. 2, Firstly, obtain the user electricity meter data of the substation's total meter box and initialize the Lasso model. Convert the daily electricity consumption data from the unrecognized substation total table and low-voltage user table into matrix form. Then initialize the user's electricity consumption regression coefficient and set the Lasso model parameters. Divide the data into training and testing samples, and iteratively solve the regression coefficients using coordinate descent method to determine whether the regression coefficients converge. When convergence is not achieved, the regression coefficients are iteratively solved using the coordinate descent method again until convergence is determined. Then determine whether 10 cross validations have been completed. If not, continue to divide the data into training and testing samples. If completed, solve for the prediction error, obtain the regression coefficient value under the minimum prediction error, and finally identify the topological variation relationship.

PGT study based on Lasso algorithm and t-SNE algorithm

It is difficult to provide a clearer analysis of the topology structure of the power grid, as it only focuses on the relationship between household changes. The topology of the desktop distribution network is a power system structure used to realize the transmission and distribution of electric energy. The current grid forms are varied, of which radial grid lines are more common, and the more typical radial low-voltage distribution PGT is shown in Fig. 3.

As shown in Fig. 3, the structure of the radial LVDN has four main layers, which are the subscriber layer, the substation box layer, the branch line layer, and the transformer layer. In the branch line layer, multiple meter boxes are connected in parallel after each branch box, and there is no upstream and downstream physical connection relationship between meter box supports [16–17]. To keep the three-phase load of the station as balanced as possible, the single-phase energy meters in the meter boxes need to be connected to the three phases. Tree trunk LVDN structure is to lead the trunk line through the transformer, through the trunk line through the distribution box, the trunk line is divided into several branches, each branch line at the same time to the individual users, a typical tree trunk LV distribution PGT is shown in Fig. 4.

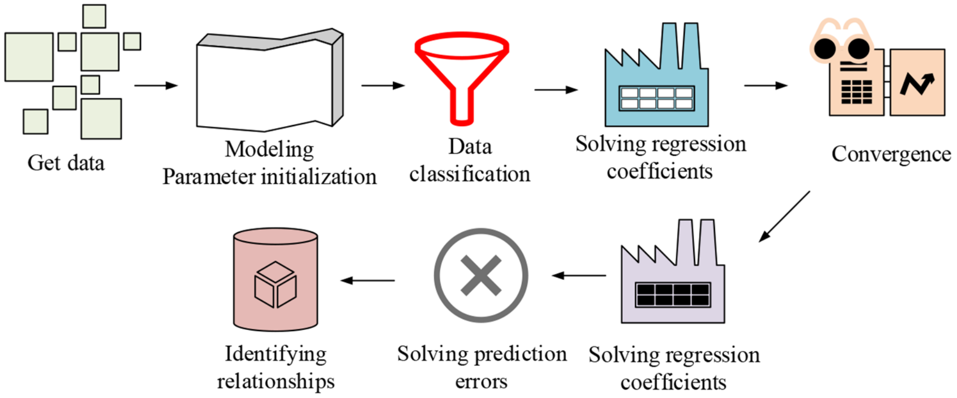
Get data
Modeling
Parameter initialization
Data classification
Solving regression coefficients
Convergence
Solving regression coefficients
Solving prediction errors
Identifying relationships

Fig. 2 Identification process of household transformation relationship

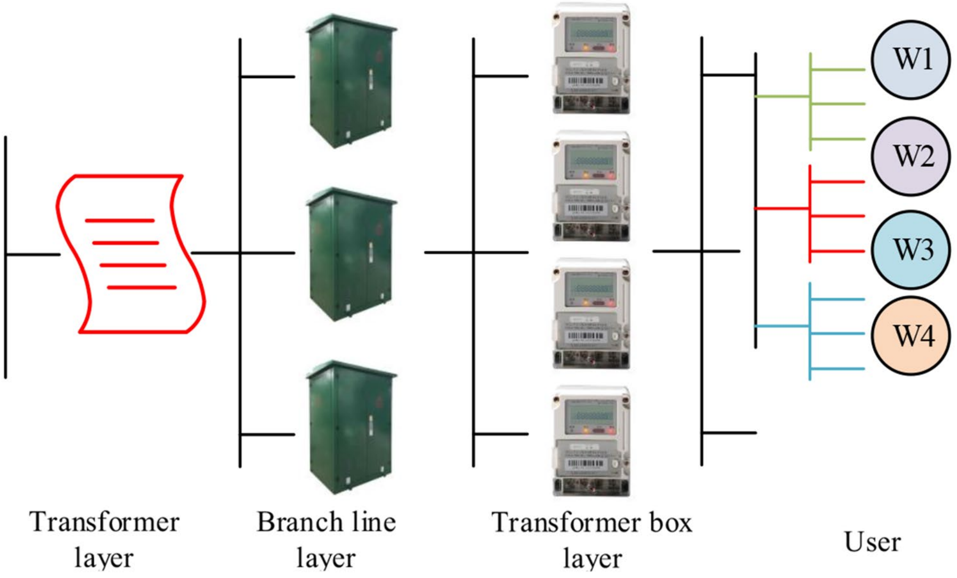

Fig. 3 Topological structure of radial LVDN

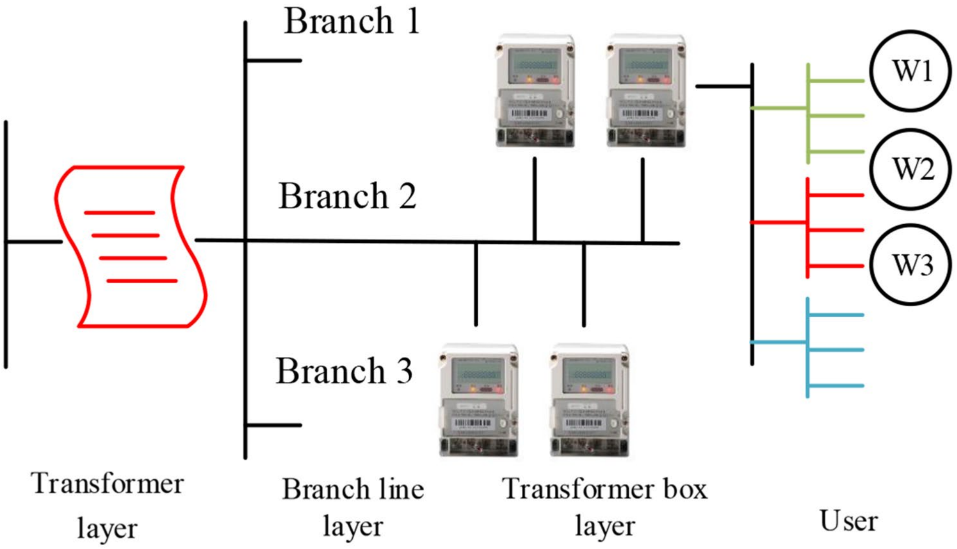

Fig. 4 Topological structure of tree trunk LVDN

In Fig. 4, the trunked LV distribution PGT has four main layers, which are the subscriber layer, the meter box layer, the branch line layer and the transformer layer. The meter box is connected to the branch line by stringing, and generally the line between two neighboring meter boxes is long, and the voltage will drop due to the long grid line [18–19]. Equation (8) illustrates the link between the voltage and the voltage at the transformer's low voltage side for each node on any line.

$$\begin{cases} U_i = U_0 - \sum_{k=1}^{i} \frac{P_{Zk}R_k + Q_{Zk}X_k}{U_{k-1}} \\ P_{Zk} = P_k + P_{loss-k} \\ Q_{Zk} = Q_k + Q_{loss-k} \end{cases} \tag{8}$$

In Eq. (8), U_i is the voltage at the position of node i , P_i is the AP at node i , and Q_i is the reactive power (RP) at node Q_i . R_i is the resistance at node i position and X_i is the reactance value at node X_i position. P_{Zi} is the transmitted AP of node i and Q_{Zi} is the transmitted RP of node i .P_{loss-k} is the power transmission AP loss at node i and Q_{loss-k} is the power transmission RP loss at node i . As the operating conditions and loads of the station are different at different moments, it leads to the voltage fluctuation following at the customer. As a result, t-distributed stochastic neighbor embedding (t-SNE) downscales the voltage data. The t-SNE algorithm is primarily used to investigate the structure of high dimensional data. It is a nonlinear dimensionality reduction (DR) and visualization tool that maps high dimensional data to low-dimensional space (LDS). By optimizing the objective function, t-SNE can preserve the local structure between data points as much as possible while reducing dimensionality. This means that after dimensionality reduction, similar data points will still maintain relatively close positional relationships in low dimensional space, which is conducive to discovering local features and clustering structures of the data. Moreover, t-SNE is suitable for dimensionality reduction of high-dimensional data. Even densely distributed data in high-dimensional space can be mapped to low dimensional space for visualization and analysis through t-SNE. Meanwhile, t-SNE is a non-linear dimensionality reduction technique that can better capture the nonlinear relationships and structures of data compared to traditional linear dimensionality reduction methods. This makes t-SNE perform better when dealing with complex datasets. For any two voltage samples, the one-sided probability between the samples is shown in Eq. (9).

$$p_{j|i} = \frac{\exp(-\|U_i - U_j\|^2/2\sigma_i^2)}{\sum\limits_{k \neq i} \exp(-\|U_i - U_k\|^2/2\sigma_i^2)} \tag{9}$$

In Eq. (9), $p_{j|i}$ denotes the conditional probability of a voltage data point appearing in another data point, and the further apart the two voltage data points are, the greater the value of the conditional probability, and the higher the similarity. σ denotes the voltage sample centered Gaussian distribution standard deviation. For solving the standard deviation, it is solved by the given perplexity parameter [20–21]. The voltage data is downscaled by t-SNE and the probability distribution between the downscaled voltage data points is shown in Eq. (10).

$$q_{ij} = \frac{(1 + \|V_i - V_j\|^2)^{-1}}{\sum\limits_{k \neq 1} (1 + \|V_i - V_j\|^2)^{-1}} \tag{10}$$

In Eq. (10), V_i and V_j represent the two voltage data points corresponding to the two voltage data points in the LDS. For the two sets of conceptual distributions in the high- and LDS, the information loss between the two probabilities can be evaluated by the calculation of the KL dispersion. The t-SNE algorithm has the final optimization objective of the minimum KL dispersion, whose expression is shown in Eq. (11).

$$\min C = KL(P \,\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{11}$$

In Eq. (11), both P and Q denote matrices, p_{ij} denotes the elements in matrix P, and q_{ij} denotes the elements in matrix Q. The t-SNE algorithm model is used to optimize the gradient of the KL dispersion of the objective function, and then solves the low-dimensional voltage data set, whose expression is shown in Eq. (12).

$$V^{(s)} = V^{(s-1)} + \eta \frac{\partial C}{\partial V} + \alpha(t)(V^{(s-1)} - V^{(s-2)}) \tag{12}$$

In Eq. (12), S denotes the iterations, $\alpha(t)$ denotes the power factor, and η denotes the learning rate. The flowchart of voltage data DR based on t-SNE is shown in Fig. 5.

In Fig. 5, the voltage data DR process based on t-SNE first obtains the total meter and single-phase user voltage data of the station area, and then sets various parameters, such as perplexity, DR, iteration number and power factor. Then calculate its one-sided probability, i.e., probability distribution, initialize the DR result of the voltage data set, calculate the low-dimensional probability distribution, solve the low-dimensional voltage data set iteratively by the gradient descent method, determine whether the iteration is completed or not, and if the iteration is not completed, then continue to calculate its status probability distribution. If the iteration is completed, the final DR result is output. The user phase difference is distinguished by unsupervised clustering algorithm, and the user characteristics are clustered by mean chain algorithm. The user is detected by the household change detection model based on Lasso regression algorithm, and then the

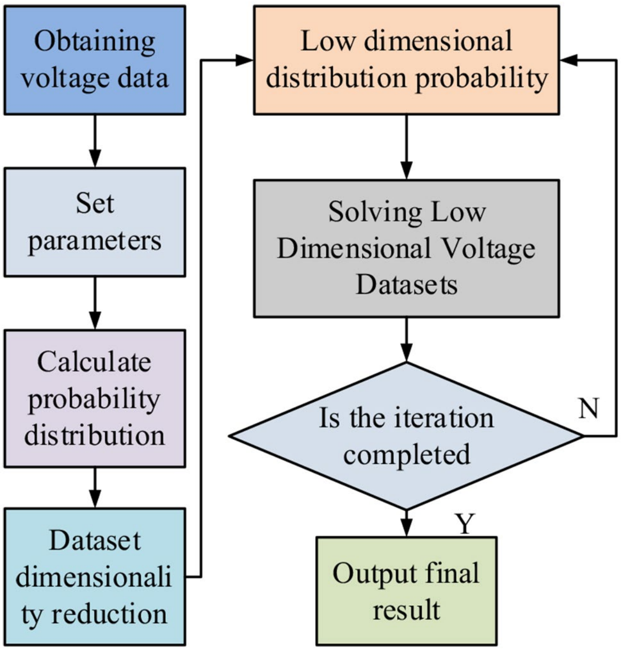
Obtaining voltage data
Set parameters
Calculate probability distribution
Dataset dimensionali ty reduction
Low dimensional distribution probability
Solving Low Dimensional Voltage Datasets
Is the iteration completed
N
Y
Output final result

Fig. 5 Flow chart of voltage data DR based on t-SNE

phase and meter box topology are recognized by t-SNE-based PGT, whose model structure is shown in Fig. 6.

In Fig. 6, the reduced-dimensional voltage dataset is first input to set the clusters for phase and meter box, and then the individual voltage samples are used as a class cluster to calculate the similarity between any two clusters. The two class clusters with the largest similarity are fused to determine whether the required clusters is reached, and if it is not reached, the similarity between the two clusters continues to be calculated repeatedly. If the number is reached, the user's phase topology is analyzed and obtained. The relationship between the user's and the transformer is recognized by Lasso algorithm to get the meter box topology and finally the topology information is output. This model identifies the household change relationship and reduces the dimensionality of the data through the Lasso algorithm, and then identifies the topology of the substation power grid through the t-SNE algorithm model.

PGT modeling based on Lasso algorithm and t-SNE algorithm

The study introduces ordinary least squares (OLS) and ridge regression (RR) to compare the performance of the model, and then introduces agglomerative clustering (AC) method and K-Means method to compare the clustering effect of the model.

Performance analysis of household change relationship identification model based on Lasso regression modeling

This study selected multiple low-pressure platforms under the jurisdiction of a certain city or region as the research objects, and collected data as the dataset by consulting relevant literature.The CPU used for the experimental hardware configuration is Intel Core i5-8750 H, the GPU is NVIDIA Geforce GTX2080Ti with 8 GB of video memory, and 16 GB of RAM. OLS and RR are introduced for comparison. And Fig. 7 displays the outcomes.

Figure 7 (a) shows the information loss rates of the three algorithms on different datasets, while Fig. 7 (b) shows the loss functions of the three algorithms on different datasets. As shown in Fig. 7 (a), as the training set increases, the information loss rates of the three algorithms also continuously decrease. When the dataset reaches 500, the information loss rates of Lasso algorithm, OSL algorithm, and RR algorithm are 0.15, 0.13, and 0.11. The proposed Lasso algorithm performs the best among the three algorithms. The Lasso algorithm, the OSL algorithm, and the RR algorithm have loss function values of 0.14, 0.20, and 0.25 when the data set reaches 500. This is in line with Fig. 7(b), which

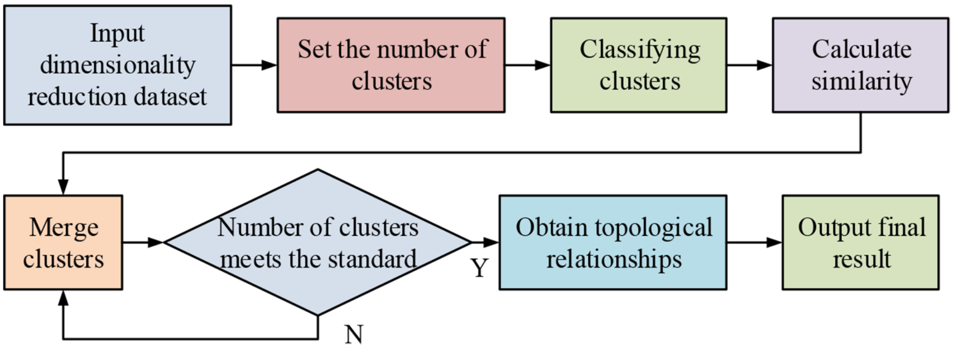
Input dimensionality reduction dataset
Set the number of clusters
Classifying clusters
Calculate similarity
Merge clusters
Number of clusters meets the standard
Y
N
Obtain topological relationships
Output final result

Fig. 6 Flow chart of phase and meter box topology identification

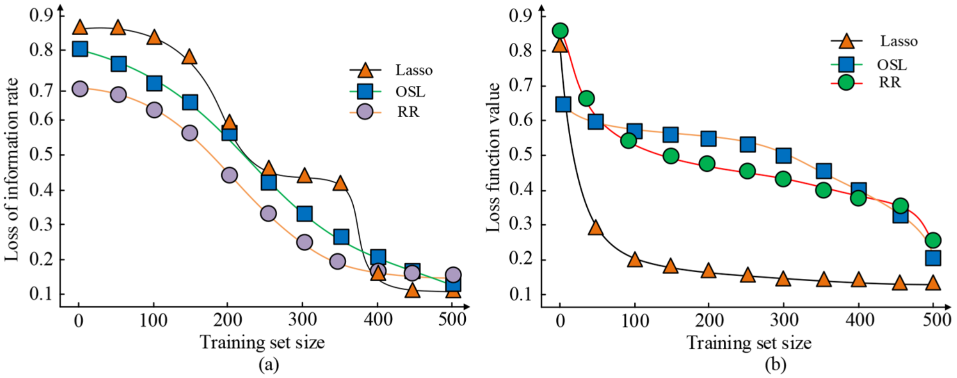

Fig. 7 Comparison of household change recognition performance among three algorithms (**a**) Accuracy of various models under different training set sizes (**b**) Loss function value of various models under different training set sizes

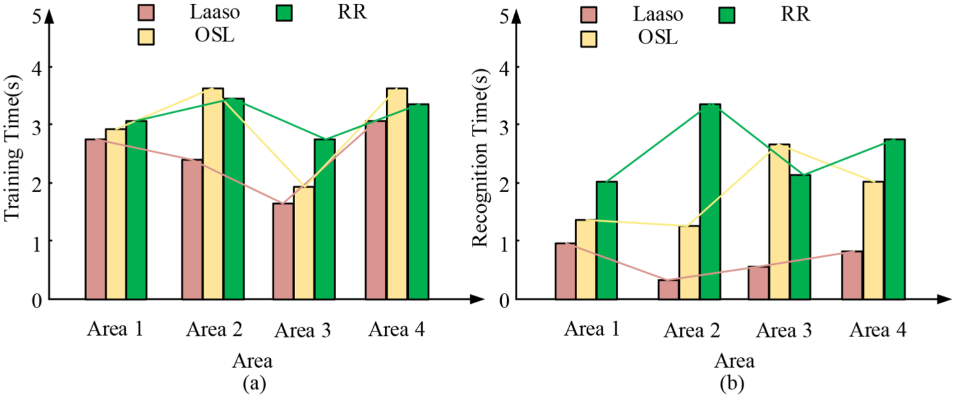

Fig. 8 Comparison of recognition time for different algorithms (**a**) Training time for different models (**b**) Recognition time for different models

shows that the loss function values of the three algorithms decrease as the training set expands. Of the three algorithms, the suggested Lasso algorithm has the smallest data function value. Out of the three algorithms, the suggested algorithm performs the best, according to the experimental data. Figure 8 displays the results of comparing the time required for HVI in various areas.

The training time of various algorithms in various regions is depicted in Fig. 8(a), and the recognition time of various algorithms in various regions is represented in Fig. 8(b). In Fig. 8(a), in region 1, the training time of Laaso algorithm, OSL algorithm and RR algorithm models are 2.8 s, 3.0 s, 3.1 s, respectively. In region 2, the training time of the three algorithms are 2.4 s, 3.6 s, 3.4 s, respectively. In region 3, the training time of the three algorithms are 7.7 s, 1.9 s, 2.8 s, respectively. In region 4, the training time of the three algorithms' training times are 3.1 s, 3.6 s, and 3.3 s, respectively, and the algorithms of each algorithm have longer algorithm training times in region 4. In Fig. 8(b), among the three algorithms, the proposed Laaso algorithm has the least recognition time among the algorithms and shows stronger performance in different regions. Table 1 displays the comparison of the three algorithms with superior performance in terms of overall performance.

Table 1 Comparison of recognition results of various algorithms

Area	USER	Data Sample Days	Recognition Accuracy (%)			Recognition Recall Rate (%)		
			Lasso	OSL	RR	Lasso	OSL	RR
Area 1	56	121	97.6	55.4	52.1	89.2	44.6	44.6
Area 2	83	145	97.6	80.4	77.3	89.2	64.5	64.6
Area 3	97	147	64.3	55.4	52.5	91.6	46.7	45.9
Area 4	134	141	70.8	30.4	27.3	91.5	87.1	76.9
Area 5	209	214	52.1	40.7	37.6	94.6	75.3	73.6
Area 6	217	314	72.6	80.4	77.3	92.9	87.1	98.7
/	Average value		75.8	49.7	46.6	91.4	69	45.4

In Table 1, among the six regions, the accuracy of region 5 is lower in each algorithmic model. The recognition accuracy of Lasso, OSL, and RR algorithm models are 52.1%, 40.7%, and 37.6%, respectively, and the recognition checking rate is 94.6%, 83.3%, and 83.3%, respectively. The accuracy for region 2 is higher, in which the recognition accuracy of Lasso, OSL, and RR algorithm models are 97.6%, 80.4%, and 77.3%, and the recognition check rate is 89.2%, 64.5%, and 64.6%, respectively. The experimental findings demonstrate that the suggested Lasso method performs better across the board for the model.

Performance analysis of PGT recognition model based on Lasso algorithm and t-SNE algorithm

Figure 9 illustrates the model clustering effect of the t-SNE technique. Based on HVI, the grid's network topology is identified using the voltage data obtained from the metering terminal.

Figure 9(a) represents the downscaling effect when the confusion degree is 2. Figure 9(b) represents the downscaling effect when the confusion degree is 10. Figure 9(c) shows the effect of DR for a confusion level of 20. Figure 9(d) represents the DR effect when the confusion degree is 40. In the figure, the DR effect is better when the value of the confusion degree is 10–20, which can retain the characteristic differences between the phases, the voltage data points between the same phase are obviously aggregated, and the separation of the voltage data points between different phases is obvious, and the differences between the individual data have been completely removed when the confusion degree is 40. Introducing K-means algorithm and Agglomerative clustering (AC), the accuracy of each algorithm was compared, and the results are shown in Fig. 10.

Figure 10 (a) shows the comparison of accuracy under different models on datasets of different sizes, while Fig. 10 (b) shows the comparison of loss functions for different models on datasets of different sizes. As shown in Fig. 10 (a), among the three algorithms, the proposed Lasso t-SNE algorithm model has the highest recognition accuracy. From Fig. 10 (b), it can be seen that among the three algorithms, the proposed Lasso t SNE algorithm model has the smallest loss function value.To rate the models that arose from this investigation, fifty power workers were chosen at random and split into five groups. The results are displayed in Table 2.

In Table 2, the five groups rated the Lasso-t-SNE algorithm model as 96.5, 95.7, 98.2, 89.3, and 88.5, the K-Means algorithm model as 84.4, 83.6, 83.4, 82.3, and 85.3, the AC algorithm as 78.9, 80, 80.6, 80.3, and 82.6, and the OLS algorithm model with ratings of 76.6, 75.9, 78.1, 74.1, and 80.7, and for the RR algorithm model with ratings of 71.2, 70.5,

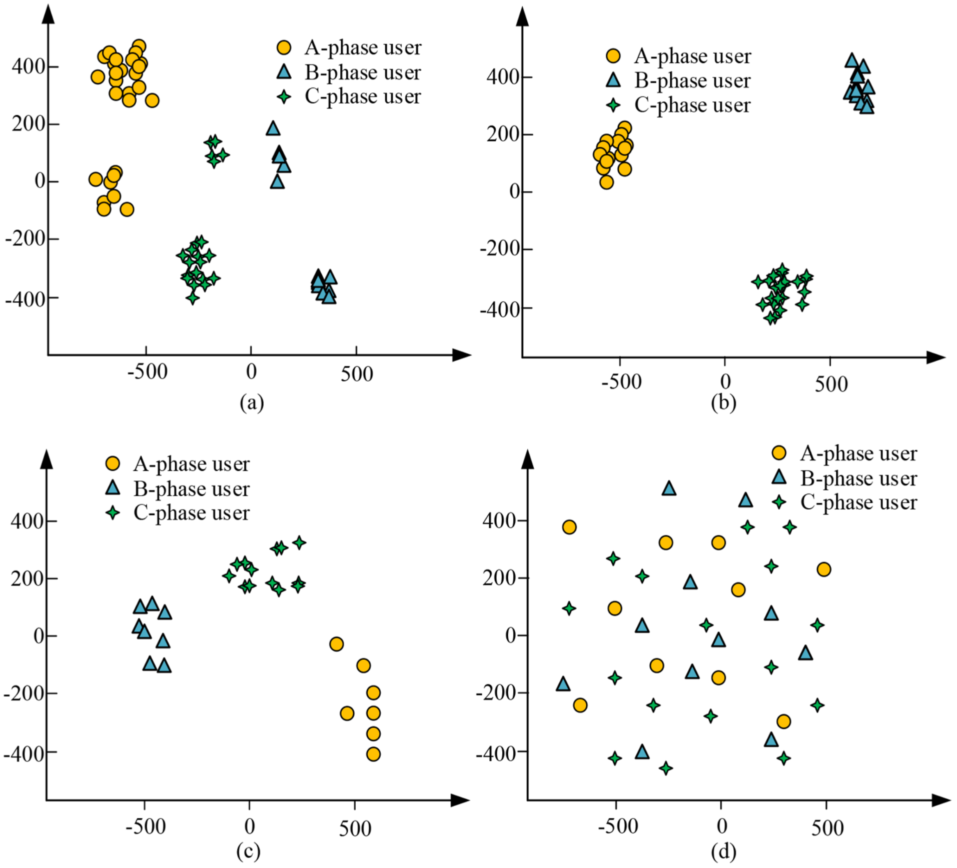

Fig. 9 DR effect of models under different levels of confusion (**a**) Perp=2 (**b**) Perp=10 (**c**) Perp=20 (**d**) Perp=40

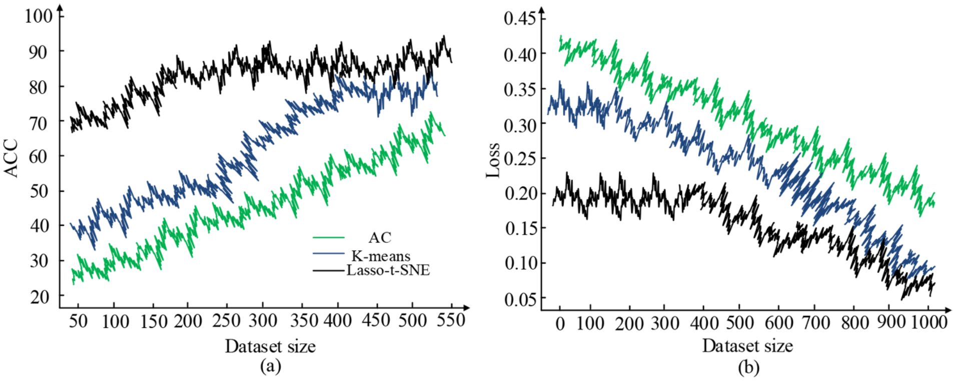

Fig. 10 Comparison of accuracy and loss function under different models (**a**) Comparison of ACC for different algorithm models (**b**) Comparison of loss function values for different algorithm models

Table 2 User evaluation form

/	Group 1	Group 2	Group 3	Group 4	Group 5
Lasso-t-SNE	96.5	95.7	98.2	89.3	88.5
K-Means	84.4	83.6	83.4	82.3	85.3
AC	78.9	80	80.6	80.3	82.6
OLS	76.6	75.9	78.1	74.1	80.7
RR	71.2	70.5	72.7	68.7	75.3

72.7, 68.7, and 75.3. The experimental results show that the proposed Lasso-t-SNE algorithm has the highest ratings among the five algorithm models.

Conclusion

In the power system, the LVDN has an extremely important role as the final link connecting the user and the high-voltage distribution network. This study proposes an automatic grid topology detection model based on Lasso algorithm and t-SNE algorithm, which identifies the household-variable relationship by Lasso algorithm, and then identifies the topology of the station grid by t-SNE algorithm model. The experimental findings showed that while the size of the training set rose, the accuracy of the Lasso, OSL, and RR algorithms increased and the value of the loss function reduced. When the data set reached 500, the accuracy rates of Lasso algorithm, OSL algorithm and RR algorithm were 0.88, 0.80 and 0.71, and the values of the loss function were 0.14, 0.20 and 0.25. In region 1, the training time of Laaso algorithm, OSL algorithm and RR algorithm models were 2.8 s, 3.0 s and 3.1 s, respectively. In region 2, the three algorithms' training times were 2.4 s, 3.6 s, and 3.4 s, respectively. In region 3, the training times of the three algorithms were 7.7 s, 1.9 s, and 2.8 s, respectively. The proposed Laaso algorithm had the least recognition time among the algorithms and showed strong performance in different regions. Fifty electricians were randomly selected and divided into five groups to rate the models that emerged from this study, and the five groups rated the Lasso-t-SNE algorithm models as 96.5, 95.7, 98.2, 89.3, and 88.5. The research results indicate that the proposed model has excellent performance for different regions and has the least recognition time among various algorithms. However, there are still shortcomings in this study, as the selected data is not comprehensive enough. If the dataset can be expanded and data types can be added, real scenario data can be collected and validated and evaluated. Through on-site testing, it is possible to more accurately understand the performance and applicability of the model in different power grid scenarios, identify potential problems, and make adjustments and optimizations. Simultaneously collect power grid datasets of different types and sources, including power grid data from different regions, scales, and operating states. By testing the model on diverse datasets, its adaptability and robustness to different situations can be evaluated, and its universality in practical applications can be verified, which can validate the performance of the model.

Author contributions

Conceptualization, methodology, writing—original draft preparation, S.H.; software, validation, formal analysis, investigation, writing—review and editing, H.Q. and Y.Z.; visualiza-tion, supervision, resources, data curation, T.X. and J.P. All authors have read and agreed to the published version of the manuscript.

Funding

Not Applicable.

Data availability

The data will be made available on request.

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 17 April 2024 / Accepted: 20 May 2024

Published online: 29 May 2024

References

Das P, Babadi B (2023) Non-asymptotic guarantees for reliable identification of granger causality via the LASSO. IEEE Trans Inf Theory 69(11):7439–7460

Gherardini S, Van Waarde HJ, Tesi P, Caruso F (2022) Topology identification of autonomous quantum dynamical networks. Phys Rev A 105(5):157–168

Henrique MB, Ramirez PAT, Nunan ZWM (2021) Improving Barnes-Hut t-SNE algorithmin modern gpu architectures with Random Forest KNN and Simulated Wide-Warp. ACM Journal on Emerging Technologies in Computing Systems (JETC), 17(4):53.1-53.26

Hock RT, Batschauer AL (2022) An interconnected single-star bridge-cell topology for grid support. International Journal of Electrical Power & Energy Systems, 135(2):107448.1-107448.8

Kumar P, Pal N, Sharma H, Kaiser MJ (2022) Optimization and techno-economic analysis of a solar photo-voltaic/biomass/diesel/battery hybrid off-grid power generation system for rural remote electrification in eastern India. Energy, 247(5):123560.1-123560.17

Lambert P, De Bodt C, Verleysen M, Lee JSQMDS (2022) A lean stochastic quartet MDS improving global structure preservation in neighbor embedding like t-SNE and UMAP. Neurocomputing 503(7):17–27

Li H, Liang W, Liang Y, Wang G, Li Z (2023) Topology identification method for residential areas in low-voltage distribution networks based on unsupervised learning and graph theory. Electr Power Syst Res 215(2):1089691–10896915

Liu H, Yang J, Ye M, James SC, Tang Z, Dong J, Xing T (2021) Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data. J Hydrol 597(1):254–267

Luo Y, Ma J, Yeo CK (2022) Identification of rumour stances by considering network topology and social media comments. J Inform Sci 48(1):118–130

Ma S, Cheng G, Li Y, Zhao R (2023) Dimension reduction method of high-dimensional fault datasets based on C_M_t-SNE under unsupervised background. Measurement, 214(21):112835.1-112835.12

Malik PK, Guha A, Seshu P (2022) Topology identification for super-stable tensegrity structure from a given number of nodes in two dimensional space. Mech Res Commun 119(6):1474–1482

Mehrim M (2022) A circuit level analysis of power distribution network on a PCB layout exposed to intentional/unintentional electromagnetic threats. Integration 89(5):25–36

Miller M, Rudion K, Nagele H, Schnaars J (2022) A grid reinforcement approach for an optimized planning of high-voltage distribution grids under consideration of line loading indicators. IET Renew Power Gener 16(9):1841–1852

Pal S, Roy A, Shivakumara P, Pal U (2023) Adapting a swin transformer for license plate number and text detection in drone images. Artif Intell Appl 1(3):145–154

Paul AR, Bhattacharya A, Chatterjee K (2023) A novel single phase grid connected transformer-less solar micro-inverter topology with power decoupling capability. IEEE Trans Ind Appl 59(2):949–958

Poliar PG, Straar M, Zupan B (2023) Embedding to reference t-SNE space addresses batch effects in single-cell classification. Mach Learn 112(2):721–740

Xia M, Wang Z, Lu M, Pan L (2022) MFAGCN: A new framework for identifying power grid branch parameters. Electric Power Systems Research, 207(7):107855.1-107855.10

Xu Q, Zhang C, Chen H, Yang H (2023a) Finite-time topology identification of stochastic delayed coupled systems on multi-weighted networks based on graph-theoretic method. J Comput Sci 69(5):10200911–102009114

Xu Z, Jiang W, Xu J, Wang D, Wang Y, Ou Z (2023b) Distribution network topology identification using asynchronous transformer monitoring data. IEEE Trans Ind Appl 59(1):323–331

Zhang XG, Feng Y, Chen H, Yuan QR (2020) Powerful t-SNE technique leading to clear separation of Type-2 AGN and H ii galaxies in BPT diagrams. Astrophys J 905(2):97–106

Zhang C, Feng Y, Li R, Yang HN (2023) Synchronisation and topology identification of stochastic delayed multi-group models with multi-dispersal and markovian switching. Int J Syst Sci 12(4):2498–2518

Publisher's Note