Computer Modeling in Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2022.020752

ARTICLE

A Fault Risk Warning Method of Integrated Energy Systems Based on RelieF-Softmax Algorithm

Qidai Lin[1], Ying Gong[2,*], Yizhi Shi[1], Changsen Feng[2] and Youbing Zhang[2]

[1]Pingyang County Changtai Power Industry Co., Ltd., Wenzhou, 325400, China

[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310023, China

*Corresponding Author: Ying Gong. Email: kungelita123@163.com

Received: 10 December 2021 Accepted: 30 January 2022

ABSTRACT

The integrated energy systems, usually including electric energy, natural gas and thermal energy, play a pivotal role in the energy Internet project, which could improve the accommodation of renewable energy through multi-energy complementary ways. Focusing on the regional integrated energy system composed of electrical microgrid and natural gas network, a fault risk warning method based on the improved RelieF-softmax method is proposed in this paper. The raw data-set was first clustered by the K-maxmin method to improve the preference of the random sampling process in the RelieF algorithm, and thereby achieved a hierarchical and non-repeated sampling. Then, the improved RelieF algorithm is used to identify the feature vectors, calculate the feature weights, and select the preferred feature subset according to the initially set threshold. In addition, a correlation coefficient method is applied to reduce the feature subset, and further eliminate the redundant feature vectors to obtain the optimal feature subset. Finally, the softmax classifier is used to obtain the early warnings of the integrated energy system. Case studies are conducted on an integrated energy system in the south of China to demonstrate the accuracy of fault risk warning method proposed in this paper.

KEYWORDS

Integrated energy system; RelieF-softmax method; fault characteristics; fault risk level prediction

1 Introduction

In recent years, in order to solve the problems of traditional non-renewable energy shortage of oil and coal and serious environmental pollution, China's strategic target of "carbon peak and carbon neutralization" needs to be achieved as soon as possible [1,2]. The energy structure dominated by fossil energy is gradually transforming to the energy structure dominated by renewable energy, and the new energy industry is developing rapidly [3,4]. Many scholars have successively put forward the concept of an integrated energy system. The electricity-gas integrated energy system composed of the coupling of electric power system and natural gas system is an important part of the "energy Internet" and a key link to realize the low-carbon, efficient and environmental protection utilization of energy. Taking account of the system net income and renewable energy abandonment rate is the key issue of the coordinated operation of the current comprehensive energy system [5]. The comprehensive energy

system optimizes and combines different power systems such as traditional power network, natural gas network and heat network, and uses advanced information network technology and communication technology to realize the coupling of a variety of energy systems and promote the efficient utilization of energy and equipment [6].

However, the integrated energy system is the same as the distribution grid with complex grid structure, wide coverage and miscellaneous types of equipment, the factors affecting its stable operation will increase, and the possibility of system failure will also be increased. Taking the gas network system as an example, the large-scale introduction of natural gas into the system strengthens the coupling between electricity and gas systems, but also leads to the increasingly prominent reliability problems of the system [7]. For example, random faults such as natural gas pipeline leakage and interruption of gas supply may lead to the shortage of natural gas supply, resulting leading to the rapid reduction of gas units, which threatens the safe and stable operation of the electric system [8]. Therefore, this paper studies the regional comprehensive energy system coupled by the power system and the natural gas system, fully understands the potential fault characteristics in the operation process of the regional integrated energy system, and explores how to extract the fault characteristics for early warning.

In the current research background, data mining technology is widely used in the power grid research field due to its strong computing power and adaptability. Mahiraj Singh Rawat analyzed the stability of various electrical parameters during large-scale access to the distribution grid of distributed energy such as wind power and solar photovoltaic, established a probability-based operational risk assessment model of the active distribution network, and adopted a detailed mathematical model of renewable resources based on wind and solar photovoltaic [9]. For how to analyze and control the cascade faults and prevent the power failure, the literature [10] proposed a model that uses the artificial neural power network machine learning tools to analyze the power grid trends and predict the cascade faults, so as to realize the early fault warning of the distribution network. In recent years, scholars for the regional comprehensive energy system between micro power grid and thermal power network affect the stability of fault factors are less. Literature [11] only studied isolated simulation studies of power grid and thermal network in integrated energy systems, did not study the coordinated operation of both energy potential failure factors.

For these problems, this paper proposes a modified RelieF-softmax algorithm to predict the fault of regional integrated energy system. The original data was first preprocessed, and the vacant, repetitive, and redundant data in the original feature data were removed to obtain the final initial feature set. Then, based on the new feature extraction method combined with the correlation coefficient method, we obtain the fault feature set with the strongest correlation minimum redundancy. Finally, the softmax classifier algorithm is reused and based on this algorithm, the final fault feature set is trained and tested, and the fault risk classification of the regional integrated energy system is determined according to the feature set to improve the prediction accuracy of the sample.

2 Distribution Network Structure Including an Integrated Energy System

As the physical carrier of the energy Internet, the comprehensive energy system, including electric energy, natural gas, thermal energy and other energy sources, improves the utilization rate of renewable energy through a multi-energy complementary way. According to different geographical factors and energy properties, integrated energy systems can be divided into user, regional and cross-regional levels [12], as shown in Fig. 1. In addition to the power grid system, gas network system and heat network system, it also includes various energy conversion. This paper takes the coupling system of the power

grid and the gas network system as an example to explore the fault risk level prediction of the regional comprehensive energy system.

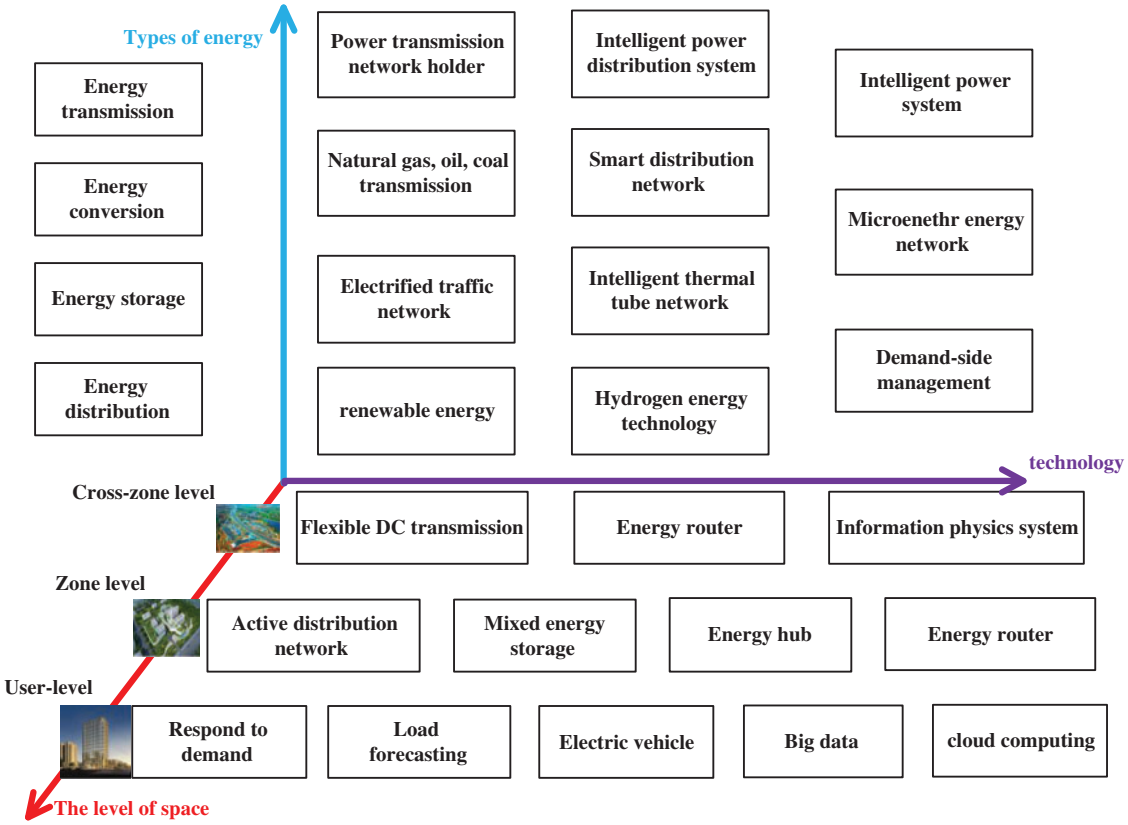

Figure 1: Comprehensive energy divide

At present, the comprehensive energy system is regarded as the initial development stage of the energy Internet, which can realize the comprehensive utilization of wind energy, solar energy, electric energy, thermal energy and other various energy sources [13]. The typical distribution network structure with an integrated energy system is shown in Fig. 2. This micro-network mainly includes wind fan, energy storage battery, electric thermal production (cold) equipment, gas unit, conventional generator set and other units, which can meet the needs of many types of electric, heat and cold loads at the same time.

In recent years, Power to Gas (P2G) technology has developed rapidly, which has built many P2G validation projects, and is operational in Europe. The technology is divided into two categories: electric hydrogen gas and electric methane. The latter methane can be directly injected into natural gas pipelines, improving the interconnection of power and natural gas systems [14,15]. As a method of energy conversion, P2G can absorb the surplus output of new energy power generation, improve the permeability of distributed power generation, and provide frequency modulation and peak modulation and other services for the power grid [16]. With the continuous maturity and commercialization of P2G technology, the electric-and gas-coupled system composed of gas turbine and P2G facilities makes the two-way flow of energy possible.

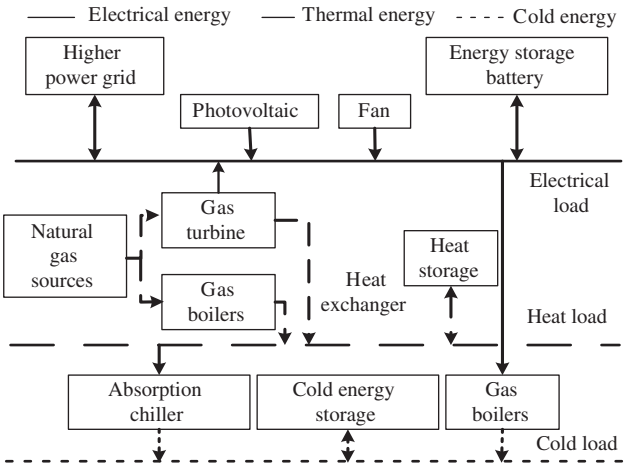

Figure 2: Microgrid structure with multi-source system

The comprehensive energy system model constructed in the paper focuses on the energy interconversion process between the power grid subsystem and the gas network subsystem, mainly through the electric gas transfer technology and gas turbine equipment realization [17]. This article takes the regional comprehensive energy system with electric-gas coupling system containing P2G as an example. The data source is 191 feeders of comprehensive energy information supervision system in southern China. This article predicts the fault risk level of the system, and obtains the results of improving the accuracy of fault risk warning through case analysis.

3 Fault Data Preprocessing

3.1 Data Preprecessing

The data in this paper are extracted from the system operation data, ledger data and historical fault information of the distribution network production management system, the comprehensive energy automation system and the electricity information collection system, and the monthly feeder data from January 2018 to June 2019 are studied.

Data preprocessing includes four steps: data cleaning, data transformation, data integration, and abnormal sample data removal [18].

(1) Data cleaning: the processing of the vacancies, repeats, outliers in the original data to ensure that the dataset is complete and reasonable during training and testing. Data vacancy value processing is to make average or median supplementary replacement or remove data in the original data. Data duplication value processing is to remove similar or duplicate data in the dataset. Data outliers processing is to identify and remove the excessively biased data based on the logical relationship of the original data.

(2) Data transformation: refers to the complex data in the original data replaced with easily analytic data, including feature construction, data classification and data quantification, the article applies to the max-minde method. For rainfall, thunderstorm, wind volume and other data, to highlight the data differences, the continuous values can be discretized and graded.

(3) Data integration: refers to performing the data statistics, integrating the data into a unified database. Given the diversity of data sources for regional integrated energy system failure features, methods for cross-validation will be necessary.

(4) Abnormal sample data removal: refers to the identification, inspection and removal of the great deviation, minimal data, wrong input data, and meaningless data in the original data set. Because the abnormal sample data has a great impact on the overall sample, it can cause large errors to the training of the data and affect the accuracy results of the test set. Therefore, Statistical, K-proximity values, and clustering methods can be used to remove the exception data, to improve the accuracy of the training results.

Finally, the final set of fault features is obtained. There are five categories of fault data, as shown in Table 1, including meteorological data, operation data, ledger data and Ran gas network data, with 24 characteristic factors.

Table 1: Initial feature set of integrated energy system failure

Type	Variable quantity	Feature field	Type	Variable quantity	Feature field
Fault data	f_1	Fault risk level	Service data	f_{13}	Maximum monthly load
	f_2	Monthly failure frequency		f_{14}	Feed verage monthly load
	f_3	Number of households affected during the power outage		f_{15}	Monthly grading
Meteorological data	f_4	Monthly cumulative rainfall	Ledger data	f_{16}	Geographical location classification
	f_5	Monthly mean temperature		f_{17}	Method of feeding line erection
	f_6	Monthly maximum temperature level		f_{18}	Power supply area classification
	f_7	Extreme weather day of the month		f_{19}	Length of the overhead section of the feeder
	f_8	Monthly average humidity		f_{20}	Feeder cable segment length
	f_9	Extreme humidity day of the month		f_{21}	Total length of feed line

(Continued)

Table 1 (continued)

Type	Variable quantity	Feature field	Type	Variable quantity	Feature field
	f_{10}	The moon wind day	Gas network data	f_{22}	Feeder gas pipe pressure value
	f_{11}	Month thunderstorm day		f_{23}	Feeder gas line length
	f_{12}	Month snow day		f_{24}	Number of feeder gas filling stations

3.2 Failure Risk Level Classification

Failure risk refers to the possibility of failure leading to the grid power failure and the degree of loss impact caused by the failure [19]. The fault risk level of the comprehensive energy system mainly includes two parts: frequency of the failure occurred (The frequency of 100 km) and effect and consequences of the failure (Fault impact range). Combined with the assessment standards of power grid companies, 191 feeders of power-gas comprehensive energy system in southern China were selected as the research objects. The fault frequency of 100 km and the number and range of households affected when the failure is taken as the basis of fault risk classification.

The 100 km fault rate is recorded as Eq. (1):

$$S_i = \frac{\sum_{j=1}^{n} f_{ij}}{L_i} \tag{1}$$

where S_i is the 100 km failure rate of the i th feeder in the region, and he f_{ij} is the identification mark for the failure j occurs on the feeder i, and the L_i is the length of the feeder i.

The impact range of the failure is based on the number of households affected by 100 km, and the calculation formula is recorded as Eq. (2):

$$C_i = \sum_{j=1}^{n_f} \sum_{k \in F_{ij}} n_{ij,k} \tag{2}$$

where C_i indicates the number of households affected by the failure on the feeder i, and n_f is the total number of failures occurring in one month in the region, and F_{ij} is a collection of transformers affected at the event of the failure j on feeder i. The $n_{ij,k}$ and the $t_{ij,k}$ indicate the number of households affected by the affected k transformer in the failure j on the feeder i, respectively.

Based on the above indicators, calculating the data on 191 feeders of power-gas integrated energy system in southern China. The comprehensive energy system fault risk level is divided into three levels: 1, 2 and 3, corresponding to three risk levels: normal, emergency and serious, and the obtained classification is shown in Table 2.

Table 2: Failure risk classification of integrated energy systems

Risk grade	Risk status	100 km fault rate	The fault affects the number of households
1	Normal	=0	0
2	Emergency	(0, 0.208]	[0, 270]
3	Serious	≥0.208	≥270

4 Selection and Extraction of the Fault Features

Characteristic selection is to select part of the features from the original dataset according to certain evaluation criteria to construct an optimal feature subset, enabling it to describe the original sample space [20]. In order to improve the accuracy of the fault risk warning of regional comprehensive energy systems, it is necessary to extract from a large number of data to the optimal fault features with relatively high fault matching with the system, and eliminate the features with low correlation. The RelieF algorithm is a filter selection algorithm widely used for feature extraction, which features screening features according to the ability to distinguish sample distance.

4.1 RelieF Algorithm

RelieF is a multivariate-filtered feature selection algorithm proposed by Kira [21] in 1992 for binary classification. Specific definition: An initial weight value $W=0$ was given to each feature in sample set D, randomly random sample s from the sample at each time. Then use the Euclidean distance formula to calculate the distance between sample s and other samples. In similar samples, the nearest neighbors H closest to sample s were found; in non-similar samples, the non-homogeneous nearest neighbors M closest to samples was found. The distance function is calculated as Eq. (3):

$$d(a,\mathbf{X},\mathbf{Y})\begin{cases}\dfrac{\mathbf{X}(a)-\mathbf{Y}(a)}{\max(a)-\min(a)}, & a \text{ is continuous quantity}\\ 0, & a \text{ is discrete quantity } \mathbf{X}(a)\neq\mathbf{Y}(a)\\ 1, & a \text{ is discrete quantity } \mathbf{X}(a)=\mathbf{Y}(a)\end{cases} \tag{3}$$

where $d\,(a,\mathbf{X},\mathbf{Y})$ shows the difference between the sample X and the sample Y.

The weight value of the fault feature is updated as Eq. (4):

$$W_a = W_a - \sum_{i=1}^{k} d(a,\mathbf{H}_i,s)/tk + \sum_{\mathbf{M}\notin class(s)}\left[\frac{N(\mathbf{M})}{1-N(class(s))}\sum_{i=1}^{k} d(a,\mathbf{M}_i,s)\right]/tk \tag{4}$$

where W_a is the weight value of features a, k is the number of nearest neighbor samples, t is the number of samples, H_i,M_i represents similar and non-similar nearest neighbors to samples, respectively. And *class* is the ratio function of the number of samples accounting for the total number of samples.

Although the RelieF algorithm has no restrictions on the data type and can give the weight values for each feature, it cannot identify the redundant fault features in the classification problem. So the article proposes an improved RelieF algorithm.

4.2 Improved the RelieF Algorithm

Some problems that exist with the RelieF algorithm:

(1) Since its initial random sampling is a put-back sampling, there may be repeated sampling limited to small category samples. Repeated-sampled samples do not provide new information to the classification and belong to invalid input, which will have an impact on the accuracy of the model results.

(2) The algorithm cannot identify the redundancy in the features, resulting in a large noise of the classification prediction input.

This paper, on the basis of the original RelieF algorithm, puts forward improvement measures: increases the hierarchical sampling of the clustering algorithm, combined with the correlation coefficient method to extract the features, so as to identify and select the redundant features.

4.2.1 Strayer Sampling Based on the Clustering Algorithm

The K-maxmin clustering algorithm [21] was used for hierarchical sampling. The K-maxmin clustering method, also known as the maximum and minimum distance method, is a clustering algorithm in the field of pattern recognition [22]. His main idea is based on the European distance formula, as far as possible to select the distant data points as the clustering center, so as to avoid the possible excessively dense clustering centers of the K-means algorithm [23].

The initial feature dataset was clustered using the K-maxmin clustering algorithm. The clustered data were then sampled stratified by comparison columns, selecting the proportion of each category to the total sample in the initial sample to determine the number of sampling assigned to each category, namely the total number of samples for each category was m. This can effectively avoid the problem of insufficient local probability of random sampling. Moreover, this sampling is not put back to sampling, ensuring that each sampling sample is the new weight value given by the feature vector, so as to exert the optimal characteristics of the hierarchical sampling of this clustering.

4.2.2 Combined with the Pearson Correlation Coefficient Method

The *Pearson correlation* coefficient was used to screen and eliminate two large correlation features. The Pearson coefficient is widely used in statistics, which reflects the degree of linear correlation between variables. Between $(-1, 1)$, the negative number is negative correlation, the positive number, and the larger the absolute value, the higher the degree of correlation. The greater the absolute value, the higher the degree of correlation indicates. The correlation ρ_{ij} expressed by the Pearson coefficient is calculated as Eq. (5):

$$\rho_{ij} = \frac{\operatorname{cov}\left(D_i, D_j\right)}{\sqrt{\sigma_{D_i}\sigma_{D_j}}} \tag{5}$$

where $cov(D_i, D_j)$ is the covariance between samples i and j, and D_i, D_j is the variance of sample i and j, respectively.

The correlation coefficient matrix was obtained by calculating the correlation coefficient between the input eigenvalues. It is generally considered that the value of the correlation coefficient greater than 0.7 is a higher degree of correlation, and its corresponding features should be attributed to the redundant eigenvector. The feature weight values were calculated combined with the RelieF algorithm, and the features with smaller weight values in the redundant feature vector with higher correlation were removed. Only one feature with higher weight values is retained in the redundant feature.

4.3 Extract the Optimal Fault Feature Variables

According to the method described above, the raw data we collected was processed and the final feature set was extracted. To improve the accuracy of failure risk level prediction for integrated energy systems, we need to eliminate the redundant feature vector already preprocessed to screen the most representative minimal redundant feature set. The flow chart of the method for extracting feature sets based on the improved RelieF algorithm is shown in Fig. 3.

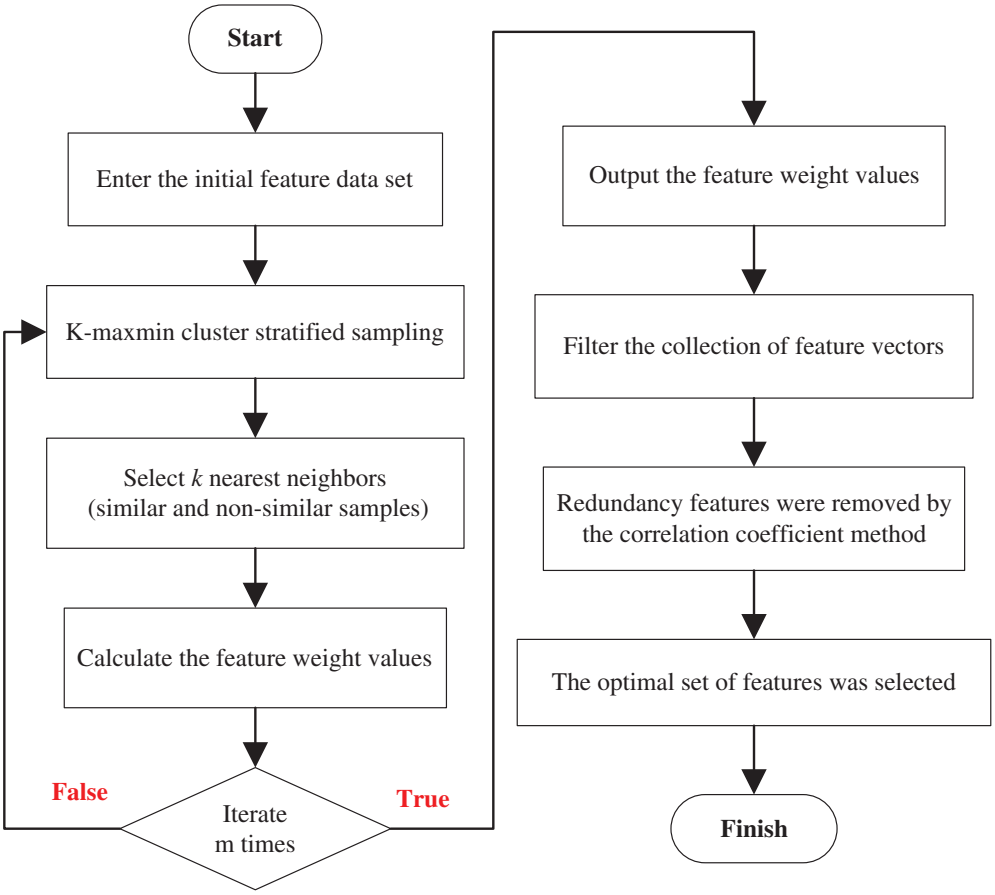

Figure 3: Extracted based on the improved RelieF algorithm flow chart of the feature method

5 Failure Risk Level Early Warning Based on the Softmax Algorithm

5.1 Based on the Softmax Algorithm

The underlying algorithm for the softmax classification is the softmax regression. It is a generalization of the logistic regression model on the multi-classification problem. Relative to the second classification problem solved by the logistic regression, in the softmax regression, we solve the multi-classification problem, and the category with the largest output probability is the category predicted by the sample set.

Enter the m data for the $\{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$, where x is the input vector and y is the category vector of its corresponding output [24]. With K categories, the output vector y_i belongs to $\{1, 2, \ldots, K\}$. Softmax regression primarily estimates the probability that the input data belongs to

each category, and for any input vector, the prediction function expression is calculated as Eqs. (6) and (7):

$$h_\theta(x) = P(y_i = 1 \mid x, \theta) = 1/\left(1 + e^{-\theta^T x}\right) \tag{6}$$

$$h_\theta(x_i) = \begin{bmatrix} P(y_i = 1 \mid x_i, \theta) \\ P(y_i = 2 \mid x_i, \theta) \\ \vdots \\ P(y_i = K \mid x_i, \theta) \end{bmatrix} = \sum_{j=1}^{K} e^{\theta_j^T x_i} \begin{bmatrix} e^{\theta_1^T x_i} \\ e^{\theta_2^T x_i} \\ \vdots \\ e^{\theta_K^T x_i} \end{bmatrix} \tag{7}$$

where P(∗) indicates the probability of events in parentheses. $\theta = [\theta_1, \theta_2, \ldots \theta_k]$ is the weight vector of n×k and n is the number of failure features of the sample. $\sum_{j=1}^{K} e^{\theta_j^T x_i}$ is a normalized parameter that guarantees that the sum of the probability of θ is 1. Data were first normalized before softmax classification.

The softmax loss function can be expressed as calculated as Eq. (8):

$$J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\sum_{j=1}^{K}\left[\ln d(y_i = j)\log\left(e^{\theta_j^T x_i}/\sum_{j=1}^{K} e^{\theta_j^T x_i}\right)\right]\right\} \tag{8}$$

where $\ln d(y_i = j)$ indicates the function for 0–1. If the value of $(y_i = j)$ is true, then $\ln d(y_i = j)$ takes 1, otherwise it is 0. The physical significance of the softmax loss function is to fully increase the proportion of sample correctly classified samples, but there is no difference between the default correct classification of various categories for balanced sample data.

Combined with Eq. (6), we can transform the fault function classification problem to predict the function parameters by solving the minimum of the softmax loss function to determine the probability of the categories of the current sample. However, there are special requirements in some research subjects. Such as the failure warning problem of the comprehensive energy system studied in the article, incorrectly predicting high-risk failure level as low risk is far higher than reducing risk rating as high-risk. Therefore, this study needs to be more accurate sample prediction of high-risk grade, and the sample data of low risk grade taken in this experiment is more than the high-risk grade samples, and the sample distribution is unbalanced. Therefore, the softmax loss function is optimized when performing the softmax prediction classification of the samples. Based on Eq. (8), this paper adds a regularization function to meet the classification requirements required in this paper, e.g., Eq. (9):

$$J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\sum_{j=1}^{K}\left[\alpha_j \ln d(y_i = j)\log\left(e^{\theta_j^T x_i}/\sum_{j=1}^{K} e^{\theta_j^T x_i}\right)\right]\right\} + \lambda\sum_{t=1}^{n}\sum_{j=1}^{K}\theta_{ij}^2 \tag{9}$$

Among them, the first item balances the classification error of the samples to adjust for the problem of unbalanced sample distribution. The second term is the regularization function, also called the L2-norm. In order to seek the global optimal solution and avoid the overfitting case of the training model, Here the optimal solution is applied to the gradient descent method, by which the softmax classifier is trained.

Seek partial derivatives for the above promoted loss function, e.g., Eq. (10):

$$\frac{\partial}{\partial\theta_j}J(\theta) = -\frac{1}{m}\left\{\sum_{i=1}^{m}\alpha_j\left[\ln d(y_i = j) - h_\theta(x_i)\right]\right\} + \frac{2}{\lambda}\theta_j \tag{10}$$

Each update iteration is performed as Eq. (11):

$$\theta_j := \theta_j - \delta \frac{\partial}{\partial \theta_j} J(\theta) \tag{11}$$

5.2 Based on the Softmax Algorithm

The fault risk warning method of comprehensive energy system based on RelieF-softmax algorithm is divided into three parts: data preprocessing, fault feature extractionand risk level prediction. Its research ideas and risk level prediction process are as follows:

1. Data preprocessing. First, the rock book data information collected from the distribution network production management system and the integrated energy automation system. Data cleaning, transformation, integration and purposeful identification elimination are then performed according to the method described above. The initial sample data matrix is finally determined, the rows of the matrix correspond to the number of samples, the columns correspond to the fault characteristics, and the last column is the risk level.
2. Extraction of the fault features. Failure characteristics need to meet two requirements, with low correlation between characteristics and the impact of fault characteristics on risk level. In this paper, we improve based on the RelieF algorithm and combine the *Pearson* correlation coefficient method to finally obtain the optimal set of features for softmax classification with low small correlation redundancy.
3. Risk level prediction. This paper trains the sample data based on softmax classifier to obtain the relationship between fault characteristics and fault risk level of integrated energy system. The prediction model was then obtained by training the samples, making the fault risk level prediction of the test samples.

The specific flowchart of this paper is shown in Fig. 4.

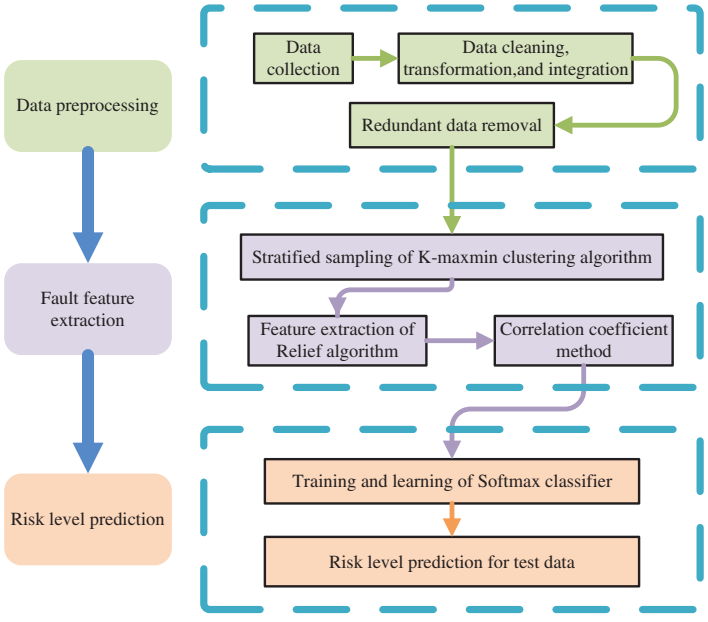

Figure 4: Integrated energy system failure warning flow chart

6 Example Analysis

This article studied 191 feeders of electricity-gas integrated energy system in Southern China. Monthly feeder fault risk levels from January to June 2020 were predicted, using the feeder data from January 2019 to December 2019 as a training sample.

6.1 Data Preprocessing

We performed the collected sample data for an example analysis. First, the sample data was preprocessed, through data cleaning, data transformation, data integration, and data redundancy and correlation. The resulting initial sample data totaled 2700, including 2538 Class I, 1111 Class II and 51 Class III.

6.2 Fault Feature Extraction

The initial sample dataset was sampled by the K-maxmin clustering algorithm, and the proportion of various samples was determined by the proportion of each type of samples in the total sample. Then the sampled samples were extracted based on the improved RelieF algorithm, calculate the weight value of each fault feature, and draw the histogram, and the results are shown in Fig. 5. This algorithm sampled a number of 30, nearest neighbor 8, and 20 iterations. In the histogram of the feature weight values, the average of the calculated weights is 0.125. We set the weight mean to the threshold, and we can see from the figure that 14 of the 20 failure features exceeded the mean, thus eliminating the other fault features below the threshold.

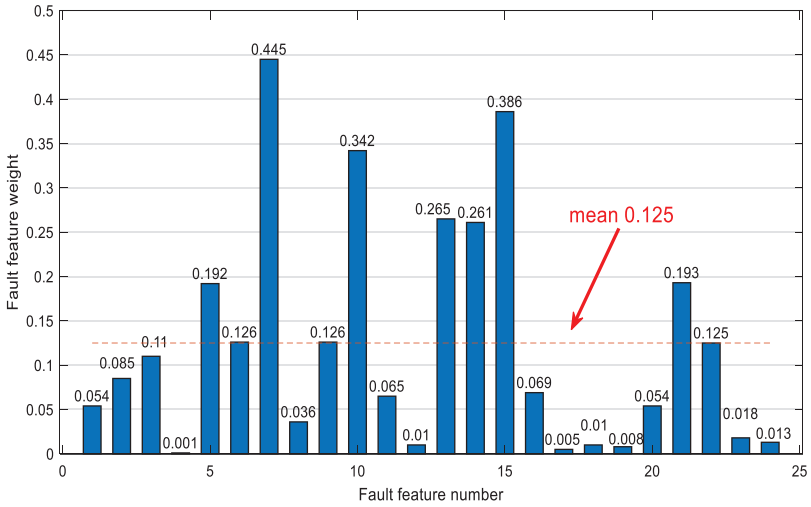

Figure 5: The fault feature weight value histogram

As can be seen from Fig. 5, 11 fault feature weight values are above threshold, and the other 13 fault features below threshold are proposed. According to the Pearson correlation coefficient method, the correlation between weight values is calculated, excluding the fault features 11, 9 and 6, so that the remaining fault features 5, 7, 10, 13, 14, 15, 21, and 22 are the final optimal set of features, as shown in Table 3.

Table 3: Optimal subset of features

Variable quantity	Weighted value	Feature field
f_5	0.192	Monthly mean temperature
f_7	0.445	Month extreme weather day
f_{10}	0342	Moonly wind day
f_{13}	0.265	Maximum monthly load
f_{14}	0.261	Feed verage monthly load
f_{15}	0.386	Monthly grading
f_{21}	0.193	Total length of feed line
f_{22}	0.125	Feeder gas pipe pressure value

We then adjusted the number of sample data sets again to 80 samples with nearest neighbor 10. We observe the remaining set of fault features in the final result, almost consistent with the most characterized set obtained for the first time. This shows that the more accurate extraction of the fault features based on the RelieF algorithm can enter the next step of data classification training learning.

6.3 Simulation Analysis of Fault Risk Level

Based on the set of 8 fault features extracted above, the sample dataset was reorganized for the training and testing of the softmax classification algorithm. For the softmax classifier, the weight decay parameter λ was 0.002, corresponding class values of 1, 2, 3, the gradient descent learning rate of 0.1 and the number of 500 iterations.

The model prediction results are shown in Table 4. Behavioral actual categories, listed as predicted categories, and the data in the table indicate the number of samples where the actual category was correctly or incorrectly predicted. For example, 1238 shows the number of samples correctly predicted as category 1 and data 11 indicates that the number of samples with the actual category 1 being predicted to be class 2 has 11, and the last column is the prediction accuracy of each category.

Table 4: The Model predicts the results

Actual category	Forecast category		
	1	2	3
1	1238	11	2
2	3	75	0
3	0	0	18
Accuracy rate	99.76%	87.21%	90.00%

As can be seen from Table 4, the rank-category prediction accuracy of the test 1347 sample data was 99.76%, 87.21%, and 90.00%, respectively. The accuracy of the overall sample was 92.32%. It can be seen that the fault risk prediction accuracy of level 1 is high. This is because the first category has a large sample base, a corresponding low fault risk level, and a good learning performance. However, the total number of samples in the third category is small, and the corresponding risk level is relatively high, which has a greater impact on the results. We can see from Table 4 that the bias in the predictions

is the Class I sample was mispredicted as in Class II and Class III, while the predictions of Class II samples are more accurate. This shows that the improved softmax classifier is better for high-level fault prediction, avoiding high-level risk prediction into low-level risk.

To verify the superiority of the modified RelieF-softmax fault risk assessment based on integrated RelieF-softmax, this paper contrasts with the traditional RelieF-softmax prediction method. The accuracy of their test results is shown in Table 5.

Table 5: Results comparison among two forecasting methods

Method	Forecast results			
	Category 1	Category 2	Category 3	Comprehensive
Improved RelieF-softmax	99.76%	87.21%	90.00%	92.32%
Conventional RelieF-softmax	89.03%	86.91%	85.78%	87.24%

As can be seen from the results of Table 5, the improved RelieF-softmax method adopted in this paper outperforms the conventional RelieF-softmax prediction methods.

7 Conclusion

This paper proposes the improved RelieF-softmax algorithm to study the fault risk level prediction of integrated energy systems with the following conclusions:

(1) Characteristic extraction based on the improved RelieF algorithm, data preprocessing on the data, including data cleaning, transformation, integration, and purposeful identification and elimination, to determine the initial sample data matrix. The proposed method combined with Pearson's correlation coefficient method, effectively overcome the shortcomings that traditional RelieF algorithm cannot remove redundancy. It can realize the dimensionality reduction of feature quantity, improve the classification performance, and finally obtain the optimal set of features for softmax classification with low correlation redundancy.

(2) Predicting the fault risk level of the integrated energy system based on the improved softmax classifier, the prediction model can effectively avoid the serious consequences caused by high-risk misclassification, reduce the cost of prediction error, and verify the effectiveness and scientificity of the method proposed herein this paper.

(3) Based on the regional comprehensive energy system composed of microgrid and natural gas network, this paper predicts the fault risk level by the 191 feeders of electricity-gas integrated energy system in a region in Southern China. The analysis found that the fault characteristic quantity of meteorological data, operation data, ledger data and gas network data all have some impact on the stable operation of electricity-gas integrated energy system.

Due to the different information acquisition channels and processing methods in various regions, the gap in various aspects should be comprehensively considered in the specific analysis process. The fault features selected in the gas network data are not complete, and the electrical comprehensive

energy system model is still relatively simple. Further research will be conducted to improve the accuracy and efficiency of fault early warning.

The accuracy of the softmax algorithm used in the fault risk level prediction of the integrated energy system still needs to be improved, and the algorithm is not optimized enough. Article [25] introduced the use of stacked bidirectional LSMRNN to assess the service life of supercapacitors [26]. Article [27] used the recurrence least squares (RLS) method and the Kalman filter (KF) online identification model to estimate the state of charging (SOC) of supercapacitors and lithium batteries in EV hybrid energy storage systems. Later, I will study the research methods in these literatures to use in the failure risk level prediction of integrated energy systems to improve the accuracy of the prediction.

Funding Statement: Project Supported by National Natural Science Foundation of China (No. 51777193).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Ma, J., Li, Q., Kühn, M., Nakaten, N. (2018). Power-to-gas based subsurface energy storage: A review. *Renewable and Sustainable Energy Reviews, 97,* 478–496. DOI 10.1016/j.rser.2018.08.056.
2. Chooruang, K., Meekul, K. (2018). Design of an IoT energy monitoring system. *Sixteenth International Conference on ICT and Knowledge Engineering(ICTKE)*, pp. 1–4. Bangkok, Thailand. DOI 10.1109/ICTKE.2018.8612412.
3. Feng, X., Li, Q., Wang, K. (2021). Waste plastic triboelectric nanogenerators using recycled plastic bags for power generation. *ACS Applied Materials & Interfaces, 13(1),* 400–410. DOI 10.1021/acsami.0c16489.
4. Feng, X., Zhang, Y., Kang, L., Wang, L., Duan, C. et al. (2021). Integrated energy storage system based on triboelectric nanogenerator in electronic devices. *Frontiers of Chemical Science and Engineering, 15(2),* 238–250. DOI 10.1007/s11705-020-1956-3.
5. Zhang, X., Bai, Y., Zhang, Y. (2022). Collaborative optimization for a multi-energy system considering carbon capture system and power to gas technology. *Sustainable Energy Technologies and Assessments, 49,* 101765.
6. Wang, Y., Wang, Y., Huang, Y., Li, F., Zeng, M. et al. (2019). Planning and operation method of the regional integrated energy system considering economy and environment. *Energy, 171,* 731–750. DOI 10.1016/j.energy.2019.01.036.
7. Liu, H. (2021). Reliability evaluation of regional energy internet considering electricity–gas coupling and coordination between energy stations. *IET Energy Systems Integration, 3(3),* 238–249. DOI 10.1049/esi2.12014.
8. Yang, T., Zhao, L., Li, W., Albert, Y. (2021). Zomaya dynamic energy dispatch strategy for integrated energy system based on improved deep reinforcement learning. *Energy, 235,* 121377. DOI 10.1016/j.energy.2021.121377.
9. Rawat, M. S., Vadhera, S. (2019). Maximum penetration level evaluation of hybrid renewable DGs of radial distribution networks considering voltage stability. *Journal of Control, Automation and Electrical Systems, 30(5),* 780–793. DOI 10.1007/s40313-019-00477-8.
10. Sudha, G., Faruk, K., Sushama, W., Ruta, K. (2015). Neural network based early warning system for an emerging blackout in smart grid power networks. *Intelligent Distributed Computing,* pp. 173–183. DOI 10.1007/978-3-319-11227-5_16.

11. S-Haghighi, A., Seifi, A. R. (2016). An integrated steady-state operation assessment of electrical natural gas and district heating networks. *IEEE Transactions on Power Systems, 31(5),* 3636–3647. DOI 10.1109/TPWRS.2015.2486819.
12. Zhang, Y., Huang, Z., Zheng, F., Zhou, R., Le, J. et al. (2020). Cooperative optimization scheduling of the electricity-gas coupled system considering wind power uncertainty via a decomposition-coordination framework. *Energy, 194,* 116827. DOI 10.1016/j.energy.2019.116827.
13. Wei, H., Zhang, Y., Wang, Y., Hua, W., Jing, R. et al. (2022). Planning integrated energy systems coupling V2G as a flexible storage. *Energy, 239(Part B),* 122215. DOI 10.1016/j.energy.2021.122215.
14. Mazza, A., Bompard, E., Chicco, G. (2018). Applications of power to gas technologies in emerging electrical systems. *Renewable and Sustainable Energy Reviews, 92,* 794–806. DOI 10.1016/j.rser.2018.04.072.
15. Lehner, M., Tichler, R., Steinmuller, H., Koppe, M. (2014). The power-to-gas concept. *Power-to-Gas: Technology and Business Models,* 7–17. DOI 10.1007/978-3-319-03995-4_2.
16. Guandalini, G., Campanari, S., Romano, M. C. (2015). Power-to-gas plants and gas turbines for improved wind energy dispatchability, energy economic assessment. *Applied Energy, 147,* 117–130. DOI 10.1016/j.apenergy.2015.02.055.
17. Liu, W., Chen, Y., Wang, L., Liu, N., Xu, H. et al. (2020). An integrated planning approach for distributed generation interconnection in cyber physical active distribution systems. *IEEE Transactions on Smart Grid, 11(1),* 541–554. DOI 10.1109/TSG.5165411.
18. Wang, H., Wang, P., Deng, S., Li, H. (2021). Improved relief weight feature selection algorithm based on relief and mutual information. *Information, 12(6),* 228. DOI 10.3390/info12060228.
19. Diao, Y., Huang, R., Wang, C., Jia, D. (2018). Fault risk prevention model of distribution network based on hidden markov. *25th International Conference on Information Science and Control Engineering (ICISCE)*, pp. 1123–1127. USA.
20. Liu, H., Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17(4),* 491–502. DOI 10.1109/TKDE.2005.66.
21. Kira, K., Rendell, L. A. (1992). The feature selection problem, traditional methods and a new algorithm. *Proceedings of 10th National Conference on Artificial Intelligence*, pp. 129–134. San Jose, CA, USA.
22. Zhen, Z., Liu, J., Zhang, Z., Wang, F., Chai, H. (2020). Deep learning based surface irradiance mapping model for solar PV power forecasting using sky image. *IEEE Transactions on Industry Applications, 56(4),* 3386–3396. DOI 10.1109/TIA.28.
23. Covões, T. F., Hruschka, E. R., Ghosh, J. (2013). A study of K-means-based algorithms for constrained clustering. *Intelligent Data Analysis, 17(3),* 485–505. DOI 10.3233/IDA-130590.
24. Hua, Y., Wang, N., Zhao, K. (2021). Simultaneous unknown input and state estimation for the linear system with a rank-deficient distribution matrix. *Mathematical Problems in Engineering, 2021,* 6693690. DOI 10.1155/2021/6693690.
25. Liu, C., Zhang, Y., Sun, J., Cui, Z., Wang, K. (2021). Stacked bidirectional LSTM RNN to evaluate the remaining useful life of supercapacitor. *International Journal of Energy Research, 46(3)*, 3034–3043.
26. Liu, C., Li, Q., Wang, K. (2021). State-of-charge estimation and remaining useful life prediction of supercapacitors. *Renewable and Sustainable Energy Reviews, 150,* 111408. DOI 10.1016/j.rser.2021.111408.
27. Wang, K., Liu, C., Sun, J., Zhao, K., Wang, L. et al. (2021). State of charge estimation of composite energy storage systems with supercapacitors and lithium batteries. *Complexity, 2021,* 1–15. DOI 10.1155/2021/8816250.