Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Electricity Theft Behavior Detection Method Based on *K*-means and Global Artificial Bee Colony Search Algorithm

Erfa Shang[1], Xiaobin Chen[2], Yuejie Xu[2,*]
[1]SPIC Xilin Gol League Huolinhe New Energy Co., Ltd, Inner Mongolia, 026099, China
[2]School of Electrical and Electronic Engineering North China Electric Power University, Beijing 102206, China
*Corresponding Author.

Abstract:
Electric theft behavior detection method has been a focal research topic in recent years. The traditional electricity theft detection method (ETDM) cannot detect electricity theft well because of the complicated means of stealing electricity and data analysis becomes a superior method. In this paper, a novel ETDM based on *K*-means and global artificial bee colony search algorithm (GABC-*K*-means) is proposed to address the above concerns. Artificial bee colony search algorithm (ABC) conquers local optimal resolution of *K*-means algorithm effectively. Global operator can adjust the dynamic balance between colony's convergence and individual's diversity and optimize global search performance. By defining the abnormal degree of load curve, we can identify the action of peccancy in using electricity. The effectiveness of the proposed method is validated in simulation section.

Keywords*: Global artificial bee colony search algorithm, clustering algorithm, electricity-theft detection.*

I. INTRODUCTION

The demand for electricity consumption is raising nowadays. The losses of electricity can be generally divided into two categories: technical losses and nontechnical losses [1]. Electricity theft is one of the most common nontechnical losses. Numerous illegal companies or individuals seek higher returns by stealing electricity in various ways. It not only results in electric shock, which may cause heavy casualties and even threaten public safety, but also brings tremendous economic losses to the state. For instance, it is reported that the loss caused by electricity theft is at least 20 billion yuan every year in China. Electric power utility places great emphasis on electric theft behavior detection. Therefore, it is necessary to carry out

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

research on electricity theft detection method (ETDM).

Methods of anti-electricity theft can be generally categorized into management means and technical means. Management means include manual inspection of defective meters and promotion of legal knowledge, which are not only extremely time-consuming but also very costly. Traditional technical means are realized by adding anti-electric stealing device into the metering device, which increases the cost and the difficulty of implementation.

The rise and development of smart grids can bring more opportunities in anti-electricity theft. The communication network system covering all aspects of the power grid can collect the power consumption and status information of users, which realizes the two-way flow of information and energy. Therefore, data mining technology is a useful solution for electricity theft behavior detection in the information flow. No matter what kind of power stealing means, its power consumption data can find abnormalities. Normal power users' consumption follows certain statistical patterns. Conversely, irregular usage pattern can be a sign of peccancy in electricity, which can be detected by data mining and machine learning technologies. Due to smart meter data is readily available, the costs are relatively moderate [2].

Since one-dimensional electricity data can not reflect the periodicity of electricity, Ma studied a wide and deep convolutional neural networks (CNN) model for electricity theft behavior detection [3]. Zhang used a two-stage unsupervised clustering method based on K-means method [4]. Considering the high dimensionality of load curve, Jiang studied a load curve fusion clustering algorithm based on wavelet transform [5]. Yang analyzed the load distribution based on the density with noise and spatial clustering, and the user's fixed and useful electricity mode is explored from the user's historical load data. After getting the corresponding load characteristics, a hybrid integer nonlinear programming for retail price of electric power is proposed [6], which is feasible. Li conducted an investigation on a statistical fuzzy technique for load curve clustering [7]. A load data classification method based on information entropy, piecewise aggregation approximation and spectral clustering (SC) is proposed [8]. This method is an improved SC clustering method based on distance and shape similarity for typical daily load data, which has good effect in data dimensionality reduction, reasonable section selection and classification. Zheng and Chen combined two new data mining methods. One method is the maximum information coefficient (MIC), which can be used to accurately detect electricity theft by shape. The other method is to realize clustering by fast searching and finding density peak (CFSFDP) [9]. Zhang studied a new energy stealing detector based on consumption mode, based on the predictability of normal and malicious consumption mode of users. A support vector machine (SVM) is used as classifier and then distribution transformer table is used to monitor the abnormal situation of consumption mode

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

[10]. The original electrical pattern ant colony clustering algorithm (EPACC) is illustrated with centroids evolution during the iterative process until stabilization. The results show that the algorithm is better than the traditional clustering algorithm [11].

Nevertheless, many former investigations have the following shortcomings:

1) Some of them need devices, which increases the difficulty of implementation.

2) Some of them requires domain knowledge which means that features need to be extracted manually.

3) The detection accuracy of non-optimized machine learning methods is not accurate enough.

Therefore, it is imperative to conduct the investigation of a novel ETDM. Based on the comprehensive analysis of previous studies, optimizing a single clustering algorithm can improve the effect of electricity theft detection. Therefore, we propose *K*-means and global artificial bee colony search algorithm to identify electricity theft behavior in this paper. *K*-means clustering algorithm is a common clustering algorithm which is fast and effective. However, the initial clustering center of the traditional *K*-means method is randomly selected and the results are greatly affected by the initial value. Besides, easy local convergence is also a non-negligible problem. Artificial bee colony search algorithm (ABC) conquers local optimal resolution of *K*-means algorithm effectively. Global operator adjusts the dynamic balance between colony's convergence and individual's diversity, which can optimize global search performance. Therefore, the global artificial bee colony search algorithm (GABC) is utilized to optimize the *K*-means clustering algorithm. The simulation section shows that global search performance of the optimized clustering algorithm is better.

II. INTRODUCTION OF ALGORITHM

2.1 *K*-means Clustering Algorithm

K-means clustering is a famous clustering algorithm. It is a widely used clustering algorithm due to its efficiency and simplicity. This algorithm is an unsupervised clustering algorithm and therefore it does not need to attain the category of data. *K*-means is different from FCM, which belongs to soft clustering algorithm and its membership degree is a number of [0,1] interval. However, the membership of *K*-means is either 0 or 1. The algorithm searches for convergence conditions through continuous iteration. The algorithm flow is as follows [12-15]:

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

a) Given the clustering number *k*, take each sample as the initial clustering center of each class.

b) Calculate the distance between each data and the clustering centers. Divide the data into the corresponding categories of the clustering centers with the smallest distance in accordance with the results.

c) The average value of each class is calculated, and take it as the new clustering center of this class. For a dataset containing $\{x_1, x_2, \ldots, x_c\}$, its cluster center calculation method is expressed as Equation (1).

$$C_i = \frac{\sum_{i=1}^{c} x_i}{c} \tag{1}$$

If the clustering center no longer changes or the convergence condition is satisfied, output the clustering results, otherwise go to step b.

2.2 Artificial Bee Colony (ABC) Algorithm

Artificial Bee Colony algorithm is a new intelligent optimization method. Compared with traditional optimization algorithms, such as particle swarm algorithm and ant colony algorithm, it has better optimization performance and ability. ABC is inspired by bee's behavior of finding food sourcessources, which is composed of employed foragers, onlookers, scouts. When an employed forager finds the food sources, it will return to the vicinity of the hive to share the food sources information with onlookers through the "swing dance", and attract a certain number of onlookers according to the amount of nectar (fitness). If the exploitation of food sources is completed, the employed forager will be transformed into a scout to search for food sources again. There is usually only one onlooker in a food source, and the number of onlookers is generally equal to the number of employed foragers. The flow of ABC is as follows [16-18]:

a) Initialize bee colony and position of all bees: define the total number of bees *N*, the maximum number of iterations *maxc*, and the maximum number of times to harvest the same food sources *lim*.

b) Based on the fitness, the bee with larger fitness becomes a employed forager, and the bee with smaller fitness becomes an onlooker.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

c) The employed forager continues to search for new food sources near the original one, and calculate the fitness value of new food sources. Based on the greedy rule, when the fitness value of the new food source is higher than the old one, the old one will be replaced by the new one.

d) Onlookers choose suitable food sources according to the information shared by employed foragers. The larger the amount of nectar, the higher the probability of the food sources being selected.

e) If when the harvest number of a food source exceeds the maximum number of times *lim*, no food sources with higher fitness have been found, then the employed foragers should stop collecting honey from it. Furthermore, randomly generate a new food source.

f) When the number of iterations exceeds *maxc* or the convergence condition is satisfied, output the results, otherwise go to step b.

2.3 Global Artificial Bee Colony Search Algorithm (GABC)

In the ABC algorithm, if an onlooker finds a better food source, it can replace the employed forager. The remaining onlookers continue to search the neighborhood until they find the best food source. Therefore, the local search ability of the artificial bee colony algorithm is superior, but the global search ability of the algorithm is insufficient, and it is easy to fall into the local optimal solution. It is reflected in the following two points:

a) In the process of bee colony updating and searching, the greedy rule is used, which accelerates the loss of bee colony diversity and causes local convergence.

b) When the bee colony is renewed, the chance of winning is very low for onlookers compared with employed foragers. It is equivalent to that randomizing bee colony is the only strategy in the artificial bee colony algorithm. It is easy to fall into local extremum due to the single strategy of bee colony renewal.

Therefore, it is necessary to increase bee colony diversity and improve the global search ability to solve the problem of local convergence. In the artificial bee colony algorithm,the way of neighborhood search is expressed as Equation (2):

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

$$x'_{ij} = x_{ij} + a(x_{ij} - x_{kj}) \quad (2)$$

where x_{ij} presents the j_{th} element of the i_{th} data, α is the neighborhood search coefficient, α is randomly selected in [-1,1], x'_{ij} represents the new data.

In order to enhance the global optimization ability of the algorithm, the third item global operator is added on the basis of Equation (2). The improved one is expressed as in (3).

$$x'_{ij} = x_{ij} + a(x_{ij} - x_{kj}) + \beta(x_{\max} - x_{ij}) \quad (3)$$

where β is a random value between 0 and 1.5, x_{max} is the global optimal fitness.

2.4 *K*-means Clustering Algorithm based on Global Artificial Bee Colony Search Algorithm

By default, the bee colony is randomly distributed on initialization. In order to accelerate the convergence of the algorithm, SOM is used to initialize the bee colony in this paper. Half of the data in SOM clustering results are randomly selected to take the average as the clustering center of GABC.

a) SSE index is generally selected as the fitness function, but this index only considers the within-class distance, without considering the between-class distance, which may affect the clustering results. However, DBI index considers both the within-class distance and the between-class distance. The smaller DBI index is, the higher the within-class distance is, the greater the between-class distance is, and the better the clustering effect is. Therefore, this paper takes DBI index as fitness function. *K*-means clustering initialize the parameters of the artificial bee colony algorithm. The total number of bees is *N*, the number of employed foragers is N_1 , the number of onlookers is also N_1, the number of scouts is between N_1 and $2N_1$, the maximum number of iterations is *maxc*, the maximum number of searches is *lim*, the number of clusters is k。

b) After the bee colony is initialized, SOM clustering algorithm is used to randomly generate *N* bees. According to the clustering center of each bee, the data set is divided into its corresponding classes. Based on its fitness, bees with high fitness are employed foragers, and bees with low fitness are onlookers.

c) The employed foragers search the neighborhood according to Equation (3), and calculate the fitness value of the new location. If the fitness value of the new position is greater than the original position, the new position will replace the old one.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

d) According to the greedy rule, the onlookers select the employed foragers to follow, and the specific selection method is expressed as in (4).

$$p_i = \frac{fit_i}{\sum_{j=1}^{n} fit_j} \tag{4}$$

where, fit_i is the fitness value of the i_{th} employed forager, P_i is the probability that the i_{th} bee is followed. The larger the fitness value is, the higher the probability of being followed is. The onlookers followed the employed foragers to search the neighborhood. If the fitness value of the new location is larger than the old one, then take the center point of the new location as the new clustering center.

e) The location of scouts is randomly assigned due to SOM clustering results.

f) After all positions are updated, if the number of searches in a food source is greater than *lim* and no new position with higher fitness is found, then the employed forager becomes a scout and randomly generates a new cluster center. Record the optimal fitness and the location of the optimal fitness. If the new optimal value is larger than the previous optimal value, the optimal value is regarded as the global optimal value, and the location of the global optimal value is updated.

g) If the number of iterations is greater than *maxc*, output the clustering result and the global optimal value, otherwise go to step b.

III. DETECTION METHOD BASED ON CLUSTERING RESULTS AND SIMILARITY

After the user's classification results are obtained by the above clustering method, all kinds of clustering centers are obtained as the typical daily load curve (TDLC) of users, which reflects the changing regularity of load under normal power consumption. Therefore, calculating the daily load curve of the users to be tested and the matching degree between them and the TDLC can screen out the suspected users. Considering the user's daily load curve is a 96 dimensional eigenvector consisted of its electricity consumption data every 15 minutes, the similarity based on vector such as Euclidean distance, Pearson correlation coefficient and cosine distance [19] can be used to calculate the matching degree between the user's daily load curve and the TDLC. These algorithms are different, the vector difference can be attributed to two kinds: the difference of absolute value of vector and the difference of vector direction. For daily load curve, the difference of vector direction is reflected in the shape of daily load curve

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

[20]. Considering a certain type of index alone can't reflect the difference between users, the two algorithms are weighted and then added up. Euclidean distance is selected for absolute value difference, which can be calculated as Equation (5). Besides cosine distance is selected for the index of the difference of curve shape, which can be calculated as Equation (6). The abnormal degree P is expressed as Equation (7).

$$\mathrm{d}(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2} \tag{5}$$

$$l(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{6}$$

$$P = w_1 \lg d + w_2 (1 - l) \tag{7}$$

where, d is the Euclidean distance between the tested users and the TDLC, l is the cosine distance between the users to be tested and the TDLC, w_1 and w_2 is the weighting coefficient. When the difference degree is greater than the threshold value, the user can be detected as a suspected power stealing user. For the two weighting coefficients, considering that the ultimate purpose of each power stealing behavior is to make the metering device record less electric energy, so the power stealing behavior is mainly reflected in the difference of absolute value. Therefore, this paper takes a larger value for the weight of European distance.

In this paper, the values of w_1=0.6, w_2=0.4 and threshold M are set to be 0.9 times of the P_{max} of various user curves. Compared with other algorithms, the mixed method based on Euclidean distance and cosine distance of load curve is simple and effective. The flow chart of abnormal curve detection method is shown in Figure 1.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

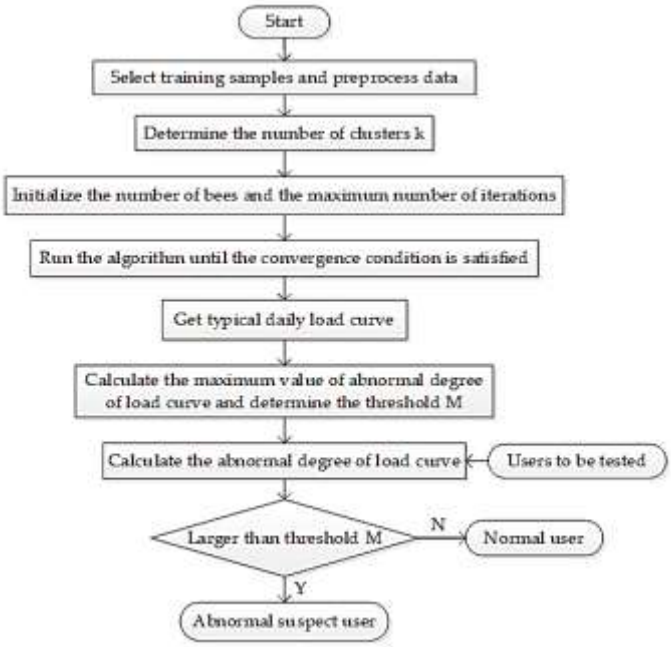


Fig 1: Detection process.

IV. ANALYSIS OF SIMULATION RESULTS

4.1 Data Selection and Preprocessing

The load data of this paper comes from 276 normal users recorded by the power supply department, which are from cement products and similar products manufacturing industry, plastic parts manufacturing and other manufacturing industry. Most users include data from January 2019 to July 2019. The data is recorded every 15 minute and the daily load curve can be obtained by the records. Firstly, delete the continuous missing data. Secondly, a few missing data are interpolated by Hermite polynomials [21]. After data preprocessing, maximum normalization is used for normalization, and then the processed data is used for clustering.

4.2 The Choice of the Best Clustering Number

The best clustering number should be determined in advance, and Elbow methods is utilized in this paper. Elbow method is based on SSE index, which can get SSE values of different clustering numbers [22]. When clustering number increases, due to the more precise division, the distance within clusters decreases, resulting in the decrease of SSE. When k is smaller than the best clustering number, the increase of *k* will lead to SSE substantially reduces.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

When *k* is greater than the best clustering number, the decrease of SSE will be significantly smaller. The whole SSE curve is an elbow shape, and clustering number corresponding to the elbow inflection point is the best. The result of Elbow curve is shown in Figure 2, which indicates the inflection point is near 4.

This paper calculates the contour coefficients of different clustering numbers to further determine the optimal clustering number. When the contour coefficient is close to 1, it means that the classification effect is better. The contour coefficients of clusters from 2 to 10 are calculated, and the results are shown in Figure 3. The contour coefficient is the largest when clusters number is 3, which means the classification effect is the best. The clustering number of this paper is set to 3 accordingly.

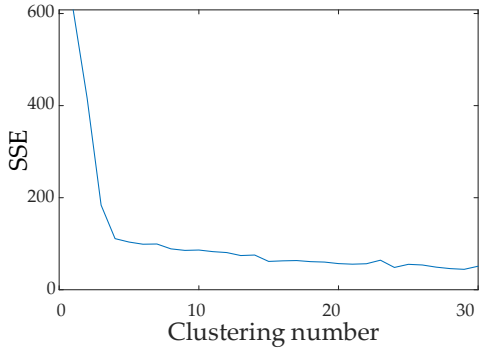


Fig 2: Elbow curve.

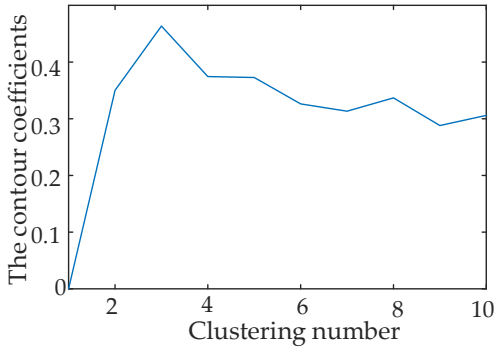


Fig 3: The contour coefficients.

4.3 Algorithm Comparison and Analysis of Simulation Results

After selecting the number of classification, *K*-means optimized by GABC is utilized to cluster the preprocessed data. The parameters of the artificial bee colony are set as follows: the maximum iterations number *maxc* is 300, the total bees number *N* is 50, the employed foragers number is 22, the onlookers number is 22, and the scouts number is 6.

To reflect the advantages of GABC-*K*-means, GABC-*K*-means and *K*-means are utilized to cluster data 20 times, and the DBI value of each clustering is calculated. The results are shown in Table 1.

TABLE I. Clustering 20 times by two algorithms

Clustering algorithm	The maximum DBI	The minimum DBI	Standard deviation

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

K-means	8.1694	7.8096	0.1217
GABC-K-means	5.9204	5.9061	0.0109

As is shown in the table above that both the maximum and minimum value of DBI of GABC-*K*-means are smaller than those of *K*-means, which means that the effect of GABC-*K*-means is better than that of *K*-means. Besides, the standard deviation of GABC-*K*-means is smaller than *K*-means, which means that GABC-*K*-means is more stable than *K*-means.

In order to reflect the difference of the convergence speed of fitness between artificial bee colony search algorithm (ABC-*K*-means) and GABC-*K*-means and the advantage of GABC-*K*-means in the global search ability, this paper uses ABC-*K*-means and GABC-*K*-means to cluster the data, and the fitness can be calculated by (8).

$$f(i) = \frac{1}{(1 + DB(i))} \tag{8}$$

where, i is the i_{th} data and $DB(i)$ is the DBI value of the i_{th} data. The fitness function value is between 0 and 1. The larger the fitness value is, the better the clustering effect is. In GABC-*K*-means, the learning step length is generally between 0 and 1.5. Considering that the data is between 0 and 1 after normalization, the learning step b=0.2.

Then two methods are used to cluster the data and record the fitness of the two algorithms in 300 iterations. The results are shown in Figure 4.

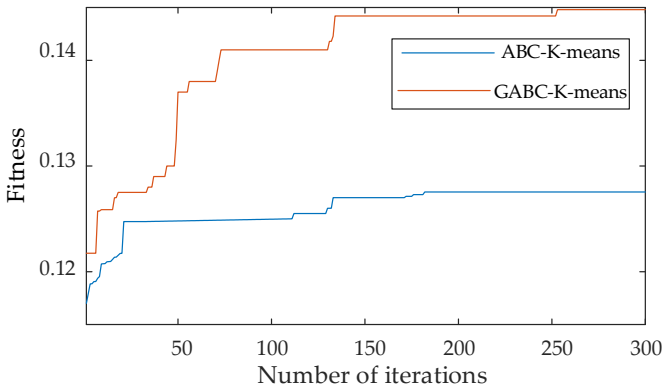


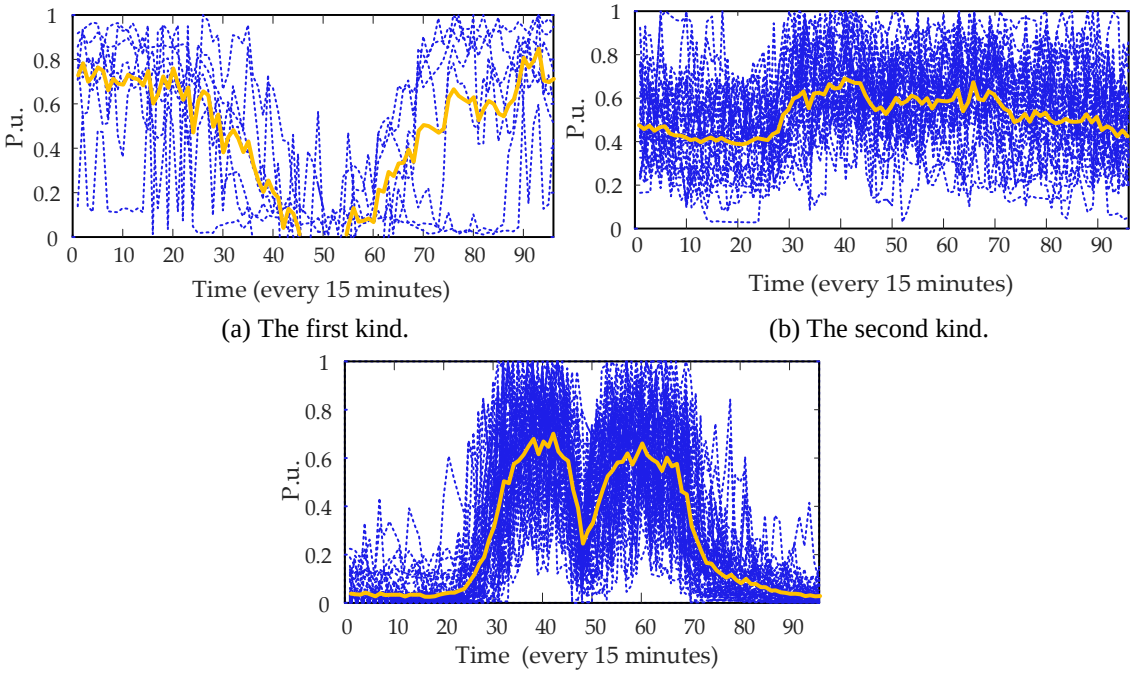
Fig 4: Fitness of the two algorithms.

As is shown in Figure 4 that the fitness value of GABC-*K*-means converged from 0.122 to 0.145, and the fitness increased by 0.023 in 300 iterations. The fitness of ABC-*K*-means

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

converged from 0.117 to 0.128, and the fitness increased by 0.011. The fitness value and convergence speed of GABC-*K*-means are better than that of ABC-*K*-means, so the classification performance of GABC-*K*-means is better.

Then GABC-*K*-means is utilized to divide the data into three categories. The clustering center are shown in Figure 5.



(a) The first kind.

(b) The second kind.

(c) The third kind.

Fig 5: Typical load curve.

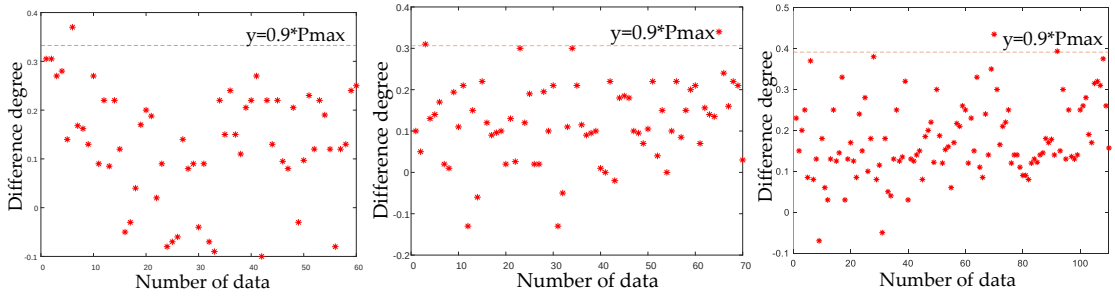
It can be seen from the above figure that the first type of users' power consumption is always high before 5 a.m., while in the daytime, the power consumption is significantly reduced and kept low, and after 11 p.m., the power consumption is significantly increased. Therefore, this type of users uses more electricity at night and less electricity during the day. Considering that the current electricity charge system in China, which divides a day into three periods: peak, average and valley, and the price difference between the peak and the low time tariff is very large. This kind of charge system encourages users to use electricity at valley time, so as to achieve the purpose of peak cutting and valley filling. The standard of time division in China is: 7:30-11:30 a.m. and 17:00-21:00 p.m. are the peak periods, 22:00-5:00 a.m. are the low periods, and the rest 9 hours are the normal periods. Therefore, the power consumption of the first type of users is typical peak avoiding, which can reduce the cost of electricity by concentrating the production in the low time period. The power consumption of the second type

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

of users changes relatively gently, with a small increase in working hours and a double peak curve rule, but it is not obvious. This type of users may be in production and operation state all day due to production needs, and the power consumption has been kept at a high level with little change. The power consumption of the third type of users is very small in the early hours of the morning, which starts to increase around 7 a.m. and reaches the peak at noon. There is a significant short-term decline in the power consumption around 12 p.m. and the power consumption will be normal in the afternoon and the power consumption drops to a low level at night. This type of users is a typical double peak curve. Its power consumption characteristics is consistent with the routine of office workers, that is, more electricity is used during working hours and less electricity is used during noon break and off-duty time.

4.4 Abnormal Daily Load Curve Detection

All kinds of clustering centers are obtained, which are taken as the TDLC of this kind. When the difference between the user to be tested and the TDLC of its category is greater than the threshold value, it can be considered that the user has abnormal suspicion. By calculating the difference degree between various users and their typical daily load curve, it is found that the difference degree of most users is less than 0.9 times of the maximum value of the difference degree, so the threshold is set as 0.9 times of the maximum value of difference degree. The distribution of difference degree of various users is shown in Figure 6.



(a) The first kind. (b) The second kind. (c) The third kind.

Fig 6: Difference degree.

The threshold setting and analysis of various kinds of users are shown in Table 2

TABLE II. Setting of various thresholds

User category	The first kind	The second kind	The third kind
Maximum of difference degree	0.3654	0.3499	0.4298

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

Threshold	0.3289	0.3149	0.3869
Number of users larger than threshold	1	1	2

After the threshold is set, if the user's difference degree is less than the threshold, it can be considered as a suspect user. If the difference degree of a user is larger than the threshold value, it can be considered as a suspect user.

V. CONCLUSIONS

Electrical theft destroys the normal order of using electricity and causes huge economic losses. The object of this paper is to provide the basis for detecting the action of peccancy in using electricity more precisely. A new method for electrical theft detection GABC-*K*-means is proposed. ABC-K-means overcomes the easy local convergence defect of *K*-means algorithm. The global operator further adjusts the dynamic balance between colony convergence and individual diversity, and the global search ability is better. The simulation results based on the recorded data show that GABC-*K*-means achieves a better performance compared with *K*-means and ABC-*K*-means for electrical theft detection in smart grids. Overall, the proposed GABC-*K*-means facilitates the monitoring of the power consuming behavior, and has great social benefit and economical benefit.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude to Mr. Yang Xue, Mrs. Yining Yang of China Electric Power Research Institute Co. Ltd., Shandong, China, for their technical support in data processing.

REFERENCES

[1] Du Z, Su S, Lliu Z, Xue Y, Yang Y, Liu S (2021) Second inspection method for electricity theft detection with low false alarm rate based on identification of production and operation status. Automation of Electric Power Systems 45(2):97-104.
[2] Wang A, Han L, Zhou Y (2021) Power theft recognition method based on power consumption data and machine learning. Information Technology 5:116-121.
[3] Ma X, Xue X, Luo H, Liu T, Yuan P (2021) Electricity theft detection based on t-LeNet and time series classification. Journal of East China Normal University(Natural Science) 5: 104-114.
[4] Zhang T, Zhang J (2018) A two-stage detection method of electricity abnormally used based on k-means. Electric Power Science and Engineering 34(12): 25-31.
[5] Jiang Z, Lin R, Yang F, Wu B (2018) A Fused Load Curve Clustering Algorithm Based on Wavelet Transform. IEEE Transactions on Industrial Informatics 14: 1856-1865.
[6] Yang J, Zhao FW, Dong Z (2019) A Model of Customizing Electricity Retail Prices Based on Load Profile Clustering Analysis. IEEE Transactions on Smart Grid 10: 3374-3386.

Forest Chemicals Revew
www.forestchemicalsreview.com
ISSN: 1520-0191
July-August 2021 Page No. 1767-1781
Article History: Received: 12 May 2021 Revised: 25 June 2021 Accepted: 22 July 2021 Publication: 31 August 2021

[7] Li W, Zhou Z, Xiong X, Lu J (2007) A Statistic-Fuzzy Technique for Clustering Load Curves. IEEE Transactions on Power System 22: 290-291.
[8] Lin S, Li F, Tian E, Fu Y, Li D (2019) Clustering Load Profiles for Demand Response Applications. IEEE Transactions on Smart Grid 10: 1599-1607.
[9] Zheng K, Chen Q, Wang E (2019) A Novel Combined Data-Driven Approach for Electricity Theft Detection. IEEE Transactions on Industrial Informatics 15: 1809-1819.
[10] Zhang Y, Ai Q, Li Z, Xiao F, Rao Y (2020) Feature Extraction Based Electricity Theft Detection for Edge Data Center. Automation of Electric Power Systems 44(9): 128-134.
[11] Chicco G, Ionel O, Porumb R (2013) Electrical Load Pattern Grouping Based on Centroid Model With Ant Colony Clustering. IEEE Transactions on Power System 28: 1706-1715.
[12] Xiang Y, Jiang H, Pan P, Sun C (2021) Study on K-means clustering algorithm of quadratic power coupling. Computer Engineering and Applications 57(14): 95-102.
[13] Tan F, Chen H, He J (2021) Oil temperature forecasting of UHV shunt reactor based on K-means clustering method and similar period. Electric Power Automation Equipment 41(6) 213-219.
[14] Zhai F, Yang T, Cao Y, Li S (2020) Key management method of smart meter based on blockchain and K-means algorithm. Electric Power Automation Equipment 40(8): 38-46.
[15] Yang Y, Qian Q, Liang Y (2021) Completion of Missing Data Through Curvilinear Proportional Expansion and Substitution Based on k-means Clustering. Electrical Automation 43(2): 50-52.
[16] Lv Y, Qian B, Hu R, Zhang Z (2021) Enhanced Artificial Bee Colony algorithm to solve semiconductor final test scheduling problem. Acta Electronica Sinica 49(9): 1708-1715.
[17] Pang Y, Liu S (2018) Optimization of MIMO radar sparse array based on modified artificial bee colony. Systems Engineering and Electronics 40(5): 1026-1030.
[18] Cai J, Wan H, Sun Y, Qin T (2021) Artificial bee colony algorithm based self-optimization of base sta-tion antenna azimuth and down-tilt angles. Telecommunications Science 37(1): 69-75.
[19] Yang H, Liu Q, Ruan Y (2021) Selection of typical daily cooling and heating load of CCHP system based on k-means clustering algorithm. Thermal Power Generation 50(3): 84-90.
[20] Senoussaoui M, Kenny P, Stafylakis T, Dumouchel P (2014) A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization. IEEE/ACM Transactions on Audio, Speech, and Language Processing 22: 217-227.
[21] Tang D, Liu Y, Xiong Z, Ma T, Su T (2020) Early Warning Method of Electricity Anti-theft in Distribution Station Area Based on Spatiotemporal Correlation Matrix. Automation of Electric Power Systems 44(19): 168-176.
[22] Zhong Z, Li M, Zhang Y (2021) Improved k-means clustering algorithm for adaptive k value in machine learning. Computer Engineering and Design 42(1): 136-141.