Transformer User Identification Method Based on Abnormal Point Detection and Large Data Clustering

Chen Liu¹, Junping Wang¹, Zhongxiao He²*, Meng Wu¹, Jialing Bai¹, Libo Lian²

¹State Grid Sichuan Electric Power Corporation Metering Center, Chengdu 610000, China
²Tsinghua Sichuan Energy Internet Research Institute, Chengdu 610000, China
*Corresponding author’s e-mail: zhongxiao_he@163.com

Abstract. Solving the mismatch of transformer and users is a key step to promote the intelligent management of distribution network. The rapid popularization of big data technology makes it possible to achieve low-cost and high-efficiency transformer user identification. Since load transferring may occur during the operation of distribution network, the user data to be identified may be mixed with the operation data of other transformer users. If the users’ data of other transformers is not eliminated, the identification accuracy will be greatly reduced. In this paper, a transformer user identification method based on local outlier detection and improved k-means algorithm is proposed. Firstly, the analysis data is preprocessed by the local factor algorithm, the user data which does not belong to the transformer to be analyzed is eliminated. According to the characteristics of application scenarios, improved k-means algorithm is proposed, including determining the number of clusters, initial centroid and selecting correlation coefficient as the index of evaluating sample similarity. Finally, the k-means algorithm is used to cluster the pre-processed data to accurately identify the transformer users. Example results show that the proposed method can effectively improve the accuracy of transformer user identification and maintain high stability in different data environments.

1. Introduction

Transformer is directly connected with the end user, so its management will directly affect the safety and experience of power consumption. Therefore, it is important to promote the fine management of planning, operation, monitoring of distribution network. Line loss and power flow calculation in distribution network are important foundations for advanced applications such as distribution network planning and operation. However, China's power system develops late, its initial development planning is not perfect and distribution lines are perplexing. Meanwhile, the distribution area file data is on paper records. With the accumulation of time, the missing data is more, and its update is not timely. Therefore, the results of power flow and line loss calculation are inaccurate [2-3], and it is difficult to promote the intelligent management of distribution network. Therefore, solving the problem of user identification is a key step to realize the intelligent management of distribution network, and it is also an important basis to realize the urban intelligence [4-5].

Traditional user identification methods mainly include manual recording and relying on communication information. However, the former has some problems such as high labour cost and low efficiency, while the latter often has the situation of "tandem area". Massive data will also be generated during the operation of the distribution network. The trend of data change is closely related...
to the regional topology. The popularization of large data technology has provided a new way for user identification. User identification technology based on large data has the advantages of low cost and good effect. At present, there are few related studies and they are still in the initial stage. Literature [8] makes a preliminary analysis of the application scenario of large data in Transformer-user relationship verification; Literature [9] proposes a line-to-line relationship verification method based on distribution network operation data, and takes the correlation coefficient between voltage curves as the verification standard. Document [10] based on the grey correlation analysis results of measurement data, the identification of user's station area and phase is realized. However, the above study did not take into account that in the process of distribution network operation, some users may have transferred to other transformers in the process of distribution network operation, and the distribution network archives records cannot be recorded in time. As a result, users in non-analysis area are classified into analysis area, so the final collected operation data may adulterate the operation data of other areas. As the basic data, the accuracy of user identification will be greatly reduced. Therefore, it is necessary to pre-process the analysis data and eliminate the user data that does not belong to the analysis area.

For this reason, this paper proposes a user identification method based on outlier detection and large data clustering. Firstly, the user data collected are pre-processed based on Local Outlier Factor (LOF) algorithm to distinguish the correct user set of transformer connection relationship in the station area and eliminate the user data that does not belong to the station area to be identified. Then, the improved k-means algorithm is used to cluster the pre-processed data, and the accurate identification of users in low-voltage stations is realized. The example analysis shows that the method proposed in this paper is effective and can effectively improve the accuracy of user identification in the platform area.

2. Full-text research ideas

Distribution network is characterized by "closed-loop design, open-loop operation", and its structure is radial. Because the load of the system varies at different time, the voltage will fluctuate. The closer the electrical distance is, the closer the voltage change trend of users is. It shows the following rules:

(1) The consistency of voltage variation of users in the same station area is higher than that of users in different stations;

(2) In the same station area, the adjacent users on the same line have higher correlation than the adjacent users on different lines.

Based on the above rules, k-means clustering method can be used to cluster the collected operational data, so as to realize the accurate identification of users in the station area.

In this paper, the output voltage of transformer's low-voltage side and the voltage data of user's ammeter are used as analysis objects. Assuming that there are N transformers in a certain area, the collection of the outlet voltage at the low-voltage side of the transformer is as follows:

\[ X_{\text{trans}} = [X_{1A}, X_{1B}, X_{1C}, ..., X_{fA}, ..., X_{NC}] \]  (1)

Where, \( X_{fA} = [x_{fA1}, x_{fA2}, ..., x_{fAn}] \), \( x_{fAn} \) represents the A-phase voltage of the \( f \) transformer at time \( n \).

Assuming that there are \( M \) users in the station area, the set of user-side voltages to be identified can be expressed as:

\[ X_{\text{user}} = [X_1, X_2, ..., X_f, ..., X_M] \]  (2)

Where, \( X_f = [x_1, x_2, ..., x_n] \), \( x_{fn} \) represents the voltage of the \( f \) user at time \( n \).

However, in the process of distribution network operation, if transformer overload occurs, some users will be relocated to nearby stations, thereby alleviating transformer load pressure. As shown in Figure 1, when the load of Transformer 1 is too large, it will be considered to migrate one of the C phase users of Transformer 1 to the A phase of Transformer 2. However, in the actual operation of the distribution network, the migration of some users is often not recorded in time, resulting in the
migration of users still being counted into the original station area. If such users are divided into phases as clustering analysis data, the final identification accuracy will be greatly affected. For this reason, this paper first pre-processes the voltage data collected by the local anomaly factor algorithm, eliminates the user data whose local outlier factor is greater than 2, and obtains the correct user set of the transformer connection relationship in the station area.

In addition, according to the characteristics of specific application scenarios, K-Means clustering algorithm is improved, including determining the number of clusters in advance, initial centroid, and using correlation coefficient as similarity evaluation index, so as to improve the accuracy of user identification in the platform area. The idea of this paper is shown in Figure 2.

3. Transformer user relationship identification method

3.1. Local Outlier Factor
Local factor algorithm is a density-based anomaly detection algorithm. Local outlier factor is used to evaluate the degree of anomaly of samples relative to local neighborhood. In this paper, the local anomaly factor algorithm is used to eliminate user data that do not belong to the station area. It is assumed that the voltage data is aggregated into \( D \), where \( p \) is any sample in \( D \). The main algorithm flow is as follows:

(1) K-distance of sample \( p \): \( k - \text{dist}(p) \)

\( k - \text{dist}(p) \) represents the distance from \( P \) to \( k \), which is defined as follows:

\[
k - \text{dist}(p) = d(p,o)
\]

(3)
Where, \( d(p,o) \) represents the distance between \( p \) and \( o \), formula (1) satisfies the following conditions:

a) There are at least \( k \) samples \( p \in D \) in data set \( D \) except sample \( p \), which satisfies \( d(p,p) \leq k - \text{dist}(p) \);

b) There are at most \( k-1 \) samples \( p \in D \) in data set \( D \) except sample \( p \), which satisfies \( d(p,p) > k - \text{dist}(p) \);

(2) K-neighborhood of sample \( p \): \( N_k(p) \)

\( N_k(p) \) represents all samples’ spatial distance from \( p \) is not greater than or equal to \( k - \text{dist}(p) \) except \( p \). The expressions are as follows:

\[
N_k(p) = \{ o \in D \setminus \{ p \} \mid d(p,o) \leq k - \text{dist}(p) \}
\]  

(4)

(3) Reachable distance between sample \( p \) and other samples: \( r_k(p,o) \)

\( r_k(p,o) \) is the reachable distance between \( p \) and \( o \). The reachable distance of \( k \) points nearest to \( p \) is equal. The expression of reachable distance is as follows:

\[
r_k(p,o) = \max \{ (k - \text{dist}(p), d(p,o)) \}
\]  

(5)

(4) Local reachable density of sample \( p \): \( lrd_k(p) \)

\( lrd_k(p) \) is the reciprocal of the average reachable distance between the sample in the k-distance and sample \( p \). The higher the value, the greater the possibility of belonging to the same category. The expression of \( lrd_k(p) \) is as follows:

\[
lrd_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} r_k(p)}
\]  

(6)

Where, \( |N_k(p)| \) is the number of samples for \( p \).

(5) Local outlier factor of sample \( p \): \( \text{LOF}_k(p) \)

\( \text{LOF}_k(p) \) is the average of the ratio of the local reachable density of \( p \)'s K-neighborhood sample to its own local reachable density. If the degree of outlier of sample \( p \) is higher and the degree of outlier of neighborhood \( k \) of sample \( p \) is lower, the relative density of sample \( p \) is lower. The higher the degree of alienation of sample \( p \), the lower the local reachable density of sample \( p \), and the lower the degree of alienation of K-neighborhood sample of sample \( p \). The higher the local reachable density of sample \( p \), the larger the local alienation factor of object \( p \). The local outlier factor is close to 1 for a sample embedded in a uniform cocoon. Therefore, no matter whether the cocoon is dense or sparse, the object in the cocoon will never be marked as outlier.

\[
\text{LOF}_k(p) = \frac{\sum_{p \in N_k(p)} lrd_k(p)}{|N_k(p)|}
\]  

(7)

The value of k-value parameter in this paper refers to reference [11]. When the distribution network is in normal operation and there is no outlier data, the LOF value of most samples is approximately equal to 1. The larger the distance between sample \( P \) and other samples, the larger the reachable distance of sample \( P \) and the smaller the local reachable density of sample \( P \). Therefore, when \( \text{LOF} > 1 \), the larger the \( \text{LOF} \), the greater the possibility that the sample is abnormal data. Therefore, in order to ensure the reliability of data pre-processing, the criterion for identifying abnormal data samples is that the LOF value is greater than the setting value. Considering the actual operation status of the distribution network and the actual needs of user identification in the station area, based on the principle of LOF algorithm, a large number of simulation experiments verify that when the setting value is set to 1.5, the identification of abnormal data can be better realized. Therefore, in this paper, if the number of local outlier factors is greater than 1.5, the user is determined not to belong to the station area, and the data is eliminated.
3.2. Improved k-means algorithm

Based on the clustering analysis data obtained by the local anomaly factor algorithm, this paper uses the improved K-Means algorithm to cluster the collected voltage data. The core of K-Means algorithm is to randomly select k points in the data set as the initial centroid. According to the distance between the sample and the K centroid, the samples are classified into the classes with the smallest distance. Then, the mean center of the sample is used as the new centroid, and iteration is repeated until the error function is less than the prescribed threshold.

The number of clusters and the initial centroid will greatly affect the accuracy of K-Means algorithm. However, due to the particularity of the scene identification, the number of clusters and the accuracy of the initial centroid enhancement algorithm can be set in advance according to the actual operation. In addition, the trend of data change is an important judgment basis for completing user clustering. In this paper, correlation coefficient is introduced to replace the typical Euclidean distance, which is the main basis for measuring the direct similarity of samples. The main process of improving K-Means algorithm is as follows:

(1) Initialization: Setting the number of clusters
Assuming that the number of transformers in the station area is N, the final clustering number must be 3N.

(2) Choosing the initial center of mass
The outlet voltage of low voltage side of transformer is set as the initial center of mass.

(3) Data classification
The classical k-means algorithm takes the Euclidean distance as the criterion to measure the similarity of samples. However, the Euclidean distance can only represent the absolute distance between samples, but it cannot reflect the trend of data change. Section 1 shows that the fluctuations of voltage data of users in the same station area are similar. For this reason, correlation coefficient $C_{pq}$ is introduced to describe the consistency of fluctuation of sample data. The expression of correlation coefficient is as follows:

$$C_{pq} = 1 - \frac{(X_p - \overline{X}_p)(X_q - \overline{X}_q)}{\sqrt{(X_p - \overline{X}_p)(X_p - \overline{X}_p)} \sqrt{(X_q - \overline{X}_q)(X_q - \overline{X}_q)}}$$  \hspace{1cm} (8)

Where, n represents the data dimension of the sample, $n = \text{sampling frequency} \times \text{sampling time}$. $\overline{X}_p$ and $\overline{X}_q$ represent the data set of sample p and q in time dimension n. $X_{p\text{d}}$ and $X_{q\text{d}}$ represent the data of sample p and q in d dimension. $(X_p - \overline{X}_p)$ is the transformation of $(X_q - \overline{X}_q)$.

(4) Centroid Reset
After classifying the classes according to the correlation coefficients, the mean centroid of each class is calculated and used as a new centroid.

(5) Convergence Judgment
Determine whether the distance between the new centroid and the original centroid is less than the threshold. If satisfied, the classification is completed by the end of the algorithm. If not, the steps (3) and (4) are repeated until satisfied.

4. Example analysis

To verify the effectiveness of the proposed method, 92 users under a transformer in a Sichuan campus were selected for analysis. Among them, 92 users belong to the analysis area according to the electrical distance and archives records. The data sampling frequency of low voltage side voltage of users and transformers is 1 h, and the sampling time is 7 days, then the data dimension is 168. Firstly,
the data is pre-processed to eliminate the abnormal data. The number of local outlier factors of some users is shown in Table 1, sorted by the number of local outlier factors from high to low.

According to Table 1, there are 4 users whose local outlier factor is higher than 1.5. Among them, 88 users still belong to this area. The result of manual checking for this area is consistent with that of data pre-processing for non-analytical users. According to the above analysis, the proportion of user data in non-analysis area is 4.34%. Distribution network has a large number of downstage areas. The proportion of actual abnormal data caused by the improper update of archives should not be underestimated.

| Serial number | User    | Local outlier factor |
|---------------|---------|----------------------|
| 1             | User 7  | 2.312                |
| 2             | User 32 | 2.036                |
| 3             | User 2  | 1.832                |
| 4             | User52  | 1.765                |
| 5             | User61  | 1.321                |
| 6             | User92  | 1.026                |

Table 2 shows the effect of data pre-processing on the accuracy of user identification in the station area. From Table 2, it can be seen that the accuracy rate of user identification without using local anomaly factor algorithm data pre-processing is 63.83%, and the proportion of user data belonging to the station area is 93.62%. From this, it can be deduced that the accuracy rate of user identification belonging to the station area is 68.17%, while the accuracy rate of data pre-processing can reach 100.00%. Therefore, the existence of user data in non-analysis station area not only leads to the existence of user data. The overall accuracy of user identification decreases, and the proportion of misjudgement of users belonging to the station area increases. Therefore, the pre-processing of clustering analysis data in advance can effectively improve the accuracy of user identification.

To verify the stability of the proposed method, Table 3 shows the accuracy of user identification when the clustering analysis data contains different proportions of non-analytical user data, and table 4 shows the accuracy of user identification when the data volume is different in the case of fixed abnormal data (6 users). Among them, method 1 uses improved K-Means algorithm to cluster users without data pre-processing, and method 2 is the method proposed in this paper.

Table 2 shows that the accuracy of method 1 decreases with the increase of user data ratio in non-analytical area, especially when the user data ratio in non-analytical area is 2.22% to 4.34%, the accuracy of method 1 decreases sharply. Compared with method 1, method 2 maintains 100.00% accuracy under different user data ratios of non-analytical stations. The accuracy is very high and stable, which proves the effectiveness and practicability of this method.

As shown in Table 3, with the increasing amount of user data, the accuracy of method 1 and method 2 is improved. However, even in a low-density data environment, method 2 can still maintain a high accuracy rate, while method 1 cannot maintain a large fluctuation. This further proves the stability of the proposed method in different data environments.

| Serial number | outlier data ratio | Accuracy of method 1 | Accuracy of method 2 |
|---------------|-------------------|----------------------|----------------------|
| 1             | 2.32%             | 97.78%               | 100.00%              |
| 2             | 4.56%             | 65.21%               | 100.00%              |
| 3             | 63%               | 63.83%               | 100.00%              |
| 4             | 8.33%             | 62.50%               | 100.00%              |
### Table 3. Accuracy of user identification under different data volumes

| Sampling frequency | Sampling duration/day | Data volume | Accuracy of method 1 | Accuracy of method 2 |
|--------------------|-----------------------|-------------|----------------------|----------------------|
| 30min              | 1                     | 48          | 33.86%               | 98.12%               |
| 15min              | 1                     | 96          | 45.21%               | 100.00%              |
| 1h                 | 7                     | 168         | 63.83%               | 100.00%              |
| 15min              | 2                     | 192         | 65.51%               | 100.00%              |
| 30min              | 6                     | 288         | 88.26%               | 100.00%              |

5. Conclusion
Aiming at the mismatch of user lines in distribution stations, a method of user identification based on anomaly detection and improved K-Means algorithm is proposed. Through the local factor algorithm, the correct user set connected with the transformer in the analysis area is obtained. Based on this, the user in the area is clustered to realize the accurate identification of the user in the area. The method proposed in this paper has the following advantages:

1. Pre-processing the clustering analysis data, eliminating the user data in the analysis area, effectively improving the user identification accuracy.
2. The K-Means algorithm is improved to make it more suitable for practical application scenarios and enhance the practicability of the algorithm.
3. In different data environments, it can maintain high accuracy and good stability of the algorithm.

References
[1] Yang Rui. Research on radio network user identification technology [D]. Chongqing University.
[2] Jinli W, Wanxing S, Qipeng S, et al. Research on Intelligent Power Quality Monitoring and On-Line Simulation for Distribution Network[J]. Power System Technology, 2014, 38(2):515-519.
[3] Ya L I , Liping L , Baiqing L I , et al. Calculation of Line Loss Rate in Transformer District Based on Improved K-Means Clustering Algorithm and BP Neural Network[J]. Proceedings of the CSEE, 2016, 36(17).
[4] Shu-Yong C , Shu-Fang S , Lan-Xin L I , et al. Survey on Smart Grid Technology[J]. Power System Technology, 2009, 33(8):1-7.
[5] Shiming T, Wenpeng L, Dongxia Z, et al. Technical forms and key technologies on energy internet[J]. Proceedings of the CSEE,2015,35(14): 3482-3494.
[6] Yudong Z . An Automatic Determination Method of Smart Meters Transformer Area Distinguish[J]. North China Electric Power, 2016.
[7] Luan W, Yixin Y U, Wang B. AMI Data Analytics[J]. Proceedings of the CSEE, 2015, 35(1):29-36.
[8] Peng W , Jiaying L , Shen G , et al. Distribution System Data Analytics and Applications[J]. Power System Technology, 2017.
[9] Zeyang T , Kunpeng Z , Kan C , et al. Substation Area Topology Verification Method Based on Distribution Network Operation Data[J]. High Voltage Engineering, 2018.
[10] Mingming P, Shiming T, Bo W, et al. Research on user’s transformer attribute identification and stealing electricity detection method based on smart meter data[J]. Smart Power, 2017, 45(12).
[11] Xie J. Fuzzy Least Square Support Vector Machine Applied to Detect Damage for Fiber Smart Structures[C]// International Symposium on Intelligent Information Technology Application Workshops. 2008.