Research of Network Behavior Pattern Recognition Based on Power Monitoring System

YangYingfu Wang 1, Junxia Shi1, Huixun Li2*, Yongwei Chen1, Jianqiang Shen1, Min Kan1, Ye Liang2

1State Grid Huzhou Power Supply Company, Huzhou Zhejiang China;
2*NARI Group Corporation/State Grid Electric Power Research Institute, Beijing China

Lihuixun123@163.com*

Abstract: Based on the network security protection of the power monitoring system, this paper proposes a pattern recognition model for the operation behavior of the operation and maintenance personnel of power monitoring system. Through the technical research of the alarm pattern recognition of the network security management platform, and carries out the operation type data of the network security management platform. Clustering analysis uses historical data to intelligently train the clustering model. Through training, the risk level classification of the operation behavior is obtained, and then real-time data is introduced into the model for detection, which can judge the degree of risk of the operation event in real time. With the improvement of the K-means clustering algorithm and the application of big data analysis, it further improves the intelligence level of the platform and provides technical support for the intelligent protection of network.

Keywords: Power Monitoring System, Pattern Recognition, Clustering Analysis.

1. Introduction
The development of the power grid carries the mission of optimizing energy resources allocation, reducing energy consumption, effectively using new energy and promoting technological progress in emerging industries. It has become an important part of nation's energy strategy. As an important part of the smart grid dispatching and control system, the safety protection of substations is essential to the coordinated safety protection of the entire smart grid dispatching and control system. With the deployment and construction of the network security management platform of power monitoring system in various regulatory agencies, new demands have been put forward for the practicability, professionalism and scalability of network security management. At present, the network security monitoring device collects too much alarm data from the substation, which cannot be processed manually and identified efficiently. Therefore, only by effective correlation analysis of these large amounts of security alarm information and further merging and integration of alarms can the power monitoring system network security status be effectively monitored, network security alarm information can be reduced, and maintenance personnel's operation and maintenance efficiency can be improved.
In order to strengthen the safety awareness and skill level of operation and maintenance personnel, reduce unnecessary and improper maintenance operations, strengthen quality of operation and maintenance personnel, and improve the stability and safety of the network environment of the power monitoring system, this paper builds an operational behavior evaluation model suitable for the operation and maintenance personnel of the power monitoring system by the technical research of the alarm pattern recognition of the network security management platform.

2. Related research

The problem of behavioral anomaly detection usually includes two sub-problems, namely, defining the criteria for dividing normal behavior patterns and abnormal behavior patterns, and selecting methods or techniques for mining abnormal behavior patterns. A specific model is usually used to describe normal behavior, and then based on this model, the current behavior pattern is pattern-matched to judge whether the current behavior is abnormal.

2.1. Method based on statistics

The premise of this type of detection method is to assume that a given sample data set conforms to a mathematical distribution model, such as Gaussian distribution[1]. Then establish the mathematical distribution model based on the sample data set, and the criterion for judging whether it is abnormal is to see whether a certain data object can better conform to the mathematical distribution model. If it does not, it is judged as abnormal.

2.2. Method based on neural network

Anomaly Detection methods generally require adaptive learning processing on a given training data set, and then memorize and form normal behavior patterns or intrusive behavior patterns. It can effectively prevent many types of abnormal or intrusive behaviors. The main reason is that this technology can realize self-regulation based on the current situation of the system. This technique can better model normal behavior or intrusion behavior on high-dimensional data.

2.3. Method based on machine learning

Detection methods using machine learning algorithms are mainly based on data mining or case learning related technologies to establish normal behavior patterns to discover abnormal behaviors, such as distance-based clustering. The main advantage of this technology is to ensure high detection efficiency, while the false detection rate is relatively low, and the data dimensionality reduction technology not only reduces the computational complexity of data processing, but also removes the noise data in the data set, which is more effective to discover user behavior patterns.

2.4. Behavior sequence analysis based on continuous behavior

The user's operation behavior is isolated in time, such as clicking the mouse at T1 and tapping the keyboard at T2. To describe the user's behavior and habits requires a continuous characterization, mining the correlation between the actions. The idea of sliding windows can be used to make certain combinations of behavior sequences. The sequence width is selected based on the time interval. At regular intervals, the operation behavior of the most recent period is counted, and the movement event of the entire sequence is triggered when the statistical time is reached. Each movement moves the behavior record of the first time period in the sequence out of the queue, and adds a new behavior record at the end of the sequence, and the time width is denoted as W. For example, select the time width W=10s and perform statistics every 1s. In this way, a multi-dimensional array with a length of 10 in the time dimension can be used to implement the behavior sequence, and every 1s the entire sequence will be triggered to move backward by 1. The associated information in the behavior sequence can be effectively mined by using such a design.

3. Cluster-based pattern recognition
3.1. Brief introduction of clustering algorithm

Clustering is a special kind of classification, which is different from classification analysis method. Cluster analysis is a method of gathering information according to the principle of information similarity without knowing the class to be defined in advance. The process of clustering is carried out according to the principle of maximum similarity within a class and minimum similarity between classes. Clustering comes from many fields, including mathematics, computer science, statistics, biology, and economics. In different application fields, many clustering techniques have been developed. These technical methods are used to describe data, measure the similarity between different data sources, and classify data sources into different clusters.

Based on the above-mentioned characteristc of the clustering technology, it can be seen that it is suitable for the classification of similar things that are not yet known. Such characteristics are exactly what we need to study network anomaly detection systems. In the network, the characteristics of network data are very difficult to control and are affected by many factors. Therefore, clustering techniques can be considered in the research to characterize the characteristics of network data sources, and then network operation behaviors can be classified according to these characteristics. To find out the overall characteristics from a large number of network operation behaviors to distinguish whether these behaviors are intended to be attacked. A single data packet cannot be judged, and an overall judgment of multiple data packets is needed to judge whether there is an attack attempt.

3.2. Improvement of K-means algorithm

The idea of the K-means algorithm is for the analyzed data set, the first step is to input the user's expected classification number k[2]. According to the input parameter k, randomly select k points as the initial point of the cluster center. The second step is to calculate the distance between each cluster center and other data objects, and categorize these data objects into the cluster with the nearest cluster center, and then recalculate the cluster centers for the new class. Repeating the above process repeatedly until there is no change in the adjacent clustering center, then the clustering criterion of the algorithm has converged, the adjustment for the analysis data object is completely finished, and the algorithm is terminated. After the algorithm classification is completed, the similarity in the clusters of the cluster centers is higher, and each cluster has a lower similarity between each other. The Euclidean distance is usually used to measure similarity, and the objective function J to evaluate the quality of division is defined as:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{c_i} d_{ij}(x_j, z_i)$$

(3.1)

Where $z_i$ is the cluster center of class $c_i$, $x_j$ is the data point in cluster $c_i$, and $d_{ij}$ is the distance between $(x_j, z_i) x_j$ and $z_i$. The objective function $J$ is the sum of the distances between each data object point and the cluster center of the cluster. The smaller the value of $J$ is, the more compact and independent the cluster is, and the clustering effect is better. The K-means algorithm seeks a better clustering method by continuously optimizing the value of $J$. When the minimum value is selected, the clustering center will no longer change and the clustering algorithm ends.

The K-means algorithm is mainly composed of four steps[3]. The first step is to select the initial center point of the cluster; the second step is to classify the sample object points; the third step is to adjust the cluster center; the fourth step is to output the clustering results. The second step and the third step are performed alternately through iteration.

The description of the classic K-means algorithm is as follows:

Input: the data set to be processed and analyzed $S\{x_1, x_2, ..., x_n\} n \in N$, the number of clusters is $K$.

Output: $K$ clustering results

Proceed as follows:

Step 1: Randomly select $K$ initial cluster centers $z_1, z_2,...,z_k$ from the sample point set $S\{x_1, x_2, ..., x_n\} n \in N$;
Step 2: Cluster the sample set according to the initial clustering centers \( z_1, z_2, \ldots, z_k \) to obtain K classes \( \{c_1, c_2, \ldots, c_k\} \), and the method for determining \( c_i \) is as follows: For any sample point \( x_i \in S \), if there is \( d(x_i, z_i) < d(x_i, z_m) \), where \( m \neq j \), then \( x_i \) belongs to class \( c_i \);

Step 3: Readjust the centroid of the cluster according to the formula, \( Z'_1, Z'_2, \ldots, Z'_k \) where \( n_i \) is the number of sample points contained in the cluster \( c_i \).

\[
Z'_i = \frac{1}{n_i} \sum_{x_{i,j}} x_j
\]  

(3.2)

Step 4: Judge whether the iteration is terminated. When the objective function \( J \) converges, the algorithm can end and output the optimal clustering result \( \{c_1, c_2, \ldots, c_k\} \). Otherwise, let \( Z_i = Z'_i \), return to step 2 and execute again. In order to prevent the infinite loop that cannot meet the termination condition of the iteration, the algorithm usually sets a maximum number of iterations to terminate the iteration.

Step 5: Output the clustering results and the algorithm ends.

Aiming at the advantages and disadvantages of the classic K-means algorithm, the algorithm is improved[4]. First, according to the different effect of each attribute of the data object in the clustering process, the basis for assigning a corresponding weight to each attribute is the coefficient of variation weighting method to improve the accuracy of the clustering analysis results. The improved clustering initialization center selection algorithm can obtain the determined high-quality initial clustering center, avoiding the problem that the classic K-means algorithm is sensitive to the initial center value and easy to fall into the local optimum.

3.3. Pattern recognition system composition

The pattern recognition system model based on cluster analysis is composed of five modules: network data collection module, data object preprocessing module, cluster analysis module, separate classification module, new data detection and analysis module[5]. The functions of each module are as follows:

1) Network data collection module

The basis of pattern recognition is data collection. This module mainly extracts data from the network security management platform database.

2) Data preprocessing module. In this module, the main purpose is to preprocess the collected network data to standardize it, convert it into a suitable data format, and perform operations such as denoising and filtering the data.

3) Cluster analysis module

After preprocessing the data objects, the module performs clustering algorithm analysis and classification on these data (usually operation records), and generates cluster analysis results.

4) Separately label modules

After the analysis and processing of the previous module, several classification families will be generated, and each cluster contains some link records. Because of the differences in the characteristics of abnormal link records and normal link records, they will be put into different clusters if they are not similar. In this way, the classes (clusters) can be marked separately. Those that contain less data are abnormal, and those that contain more data are normal.

5) New data detection and analysis module

After completing the separate identification of the clustering results of the data objects, the standard class results can be used to detect. After the data is standardized, the distance to each cluster center is calculated, and then the cluster with the closest distance is summarized. If the cluster of this category is marked as normal, then the new data object is judged to be normal data, otherwise it is considered that the new data object is abnormal data.

4. Implementation of pattern recognition module
4.1. Data sources
Database and server: Elasticsearch system. Using Elasticsearch system for system monitoring and big data analysis. Save security alarm data to HDFS and Elasticsearch indexes in the big data platform in real time or near real time. Use nGram and non-nGram indexes to perform Elasticsearch performance in a cloud environment. The data set is the data of the commands and directories in the network security management platform of power monitoring system from December 27, 2017 to 20, 20, 2018. These data include SSH access data (133142 groups) and local access data (8526 groups). The overall concept map includes the data layer, AI layer, network layer, and physical layer related to algorithms and applications.

4.2. Behavior level definition
First, the user operations are divided into 4 levels, the risk level ranges from low to high. The greater the number of levels, the higher the risk value. The detailed classification is as follows:
1) Level 1 user behavior: (No risk behavior, green alert)
   Operate in the "/home" directory; The commands only include "ls, cd, ifconfig, netstat, ping" to search only for files, search for IP, search for ports, and test network connection operations.
2) Level 2 user behavior: (low risk behavior, yellow alert)
   Operations under the directory "/opt or /usr or /etc or /var or /proc or /tmp"; Do any operation commands.
3) Level 3 user behavior: (medium risk, orange alert)
   Operations under the "/home" directory; Command "rm, cp, su, passwd, chown";Operations under the "/root" directory and "/" operations; Command "ls, cd, ifconfig, netstat, ping".
4) Level 4 user behavior: (high-risk behavior and abnormal behavior, red alert)
   In the "/root" directory and the "/" operation; The commands entered include commands "rm, pwd, reboot, pkill, su, chown", delete files, change passwords, restart the system, terminate processes, modify permissions and other operations; Does not meet any historical data, historical operation catalog; Implement high authority command operations. The definition of user behavior levels is shown in the following table:

| Operation path | Retrieval instruction | Modify instruction |
|----------------|-----------------------|--------------------|
| /home          | Level                 | Level2             |
| /opt, /usr, /etc, /var, /proc, /tmp | Leve2 | Leve2|
| /root, /        | Level3                | Level3             |
| No match any historical data | Leve4   | Leve4             |

Tab.1 User Behavior Level Definition

After defining the user behavior level, preliminary data is collected from the database and filtered, and all the data will be programed to set the vectorization and digitization of the text label.

4.3. Definition of outliers and initial value selection
According to the role of each attribute of the data object in the clustering process, When assigning a corresponding weight to each attribute, the coefficient of variation weighting method is used.

Through the selection rules of the initial clustering center, the method of randomly extracting the initial center points is changed, and the initial center of the clustering of the K-means algorithm is reasonably determined, which fundamentally eliminates the blindness of the initial center point selection, and also ensures the obtained initial center is high quality[6]. This solves the problem that the K-means algorithm is sensitive to the initial center value of the cluster and easily falls into the local optimum.
In this project, the outlier is the data point far away from other data, that is, starting from the most primitive outlier definition, avoiding the complexity of calculating the neighborhood features of each data object, and based on the result of agglomerated hierarchical clustering. The first n global outliers with the greatest degree of isolation are detected from the top to bottom of the clustering tree[7].

In determining the initial point, two factors should be considered: density factor and clustering factor. Since the location of the cluster center is always in a place where the data objects are dense, there are always some data objects that are closer to the cluster center[8]. If these data objects can be found and used as the initial center of the cluster, it can avoid the problems of the K-means algorithm due to the unreasonable selection of the initial center point.

1) Measurement of data object point density
In a data set, when data objects are distributed in clusters, the more other data object points around a data object point, the greater the data object distribution density at that data object point, and the impact of the data object point on classification is bigger[9]. Therefore, each data object point has a surrounding distribution density problem. For each data object point, the point distribution density function is defined as follows:

\[ p_i = z_i \sum a \cdot z_i \]  

(4.1)

\[ Z_i \] is a parameter about the distance between data object points, which is defined as follows:

\[ Z_i = \sum_{a \geq 1} \left| \frac{1}{d_a} \right| a \]

(4.2)

\[ d_a = \| x_i - x_j \| \]

(4.3)

\[ p_i \] indicates the degree of influence of the i-th data object point on the classification. The larger the \( p_i \) is, the more sample points around the data object point \( x_i \) is, and the greater the density of the data object point \( x_i \); the smaller the \( p_i \) is, the fewer sample points around the data object point \( x_i \). The density of data object points is also smaller.

2) Select the data object point with high density as the initial center of the cluster
The \( p_i \) is calculated to find the data object points with higher density, and the distance between the selected data object points is ensured to be as large as possible, otherwise the selected cluster center will inevitably gather in the data object density in the highest area, two clusters may overlap. Therefore, the first cluster center is selected according to its density, and the distance from other cluster centers needs to be considered when selecting other cluster centers. The metric chosen in the research is the product of distance and density. The units of them are different, and required to be normalized. For a given data object, transform the distance from the data object point \( x_j (j = 1,2, ..., N) \):

\[ d'_j = \frac{d_{ij}}{\sum_{a \geq 1} d_a} (j = 1,2, ..., N) \]

(4.4)

Assuming that several cluster centers have been selected, when adding a new cluster center, it is necessary to consider making the total distance from the existing cluster center larger, and the density of the data object points is also larger. Selecting \( x_j \) among the remaining data objects so that the product of the data object points distribution density and the normalized distance between the data object point and the selected cluster center is the largest:

\[ w_i = p_i \sum_{n \geq 1} d_{ij} \]

(4.5)

\( (n_1, n_2, ..., n_i) \) is the sequence number of the determined k (\( k \geq 1 \)) cluster initial centers in the data object.

3) Initialize clustering center algorithm flow
The algorithm flow is as follows:
Input: the pending data object set \( w_i \) and the number of clusters \( K \)
Output: cluster initial point set \( M \)
Step 1: Calculate the density of each data object according to formula (3.1);
Step 2: Initialize the initial clustering center point set M and set it to the empty set $M=\{\}$, the initial accumulation parameter $w_i = 0(i=1,2,\ldots,N)$;

Step 3: Set $j=1$, select the data object point with the largest distribution density $m_1$ (the $v_1$-th point) as the initial center point of the first cluster, namely:

$$p_{i_1} = \max(p_i),(i=1,2,\ldots,N)$$

$$M = M \cup \{M_1\}$$

Step 4: Calculate the normalized distance $d_{ij}$ from the initial center point of the selected cluster $m_j$ to all data objects according to formula (3.4), and accumulate $w_i$:

$$w_i = w_i + p_i d_{ij},(i=1,2,\ldots,N)$$

Step 5: Let $j=j+1$, select the $m_j$-th point (the $v_j$-th point) of the data object corresponding to the maximum value as the $j$-th cluster initial center point, which is:

$$w_{v_j} = \max\{w_i\},(i=1,2,\ldots,N),i \neq (v_1,\ldots,v_{j-1})$$

$$M = M \cup \{M_j\}$$

Step 6: Repeat steps 4 and 5 until an initial center point of the cluster is found;

Step 7: Output the center point set found, and the algorithm ends.

Since the first point of the cluster's initial center point set is determined (that is, the maximum density point), the search for the initial center points of other clusters should be based on the farthest distance, so the initial center points of the clusters obtained are basically certain, this eliminates the randomness in the selection of the initial center point of the cluster, avoids blindness, and also ensures that a higher-quality initial center point of the cluster is obtained.

5. Design of university education intelligent agent

University education intelligent agent is composed of intelligent computing and big data analysis platform, data center, the new smart campus system, the adaptive evolution system of teaching system, the whole life cycle data analysis system, expert system in man-machine integration, public cloud services platform and so on, which uses 5G network to realize connectivity. By integrated use of big data processing, deep learning, knowledge graph, man-machine integration, intelligent decision, network security technologies, it can collect continuous accumulation of multi-source data and conduct analysis and assistant decision-making, and form an open, adaptive and iterative evolution intelligent system.

6. Conclusion

This paper improves the accuracy of cluster analysis through the definition of outliers and the selection of initial values, establishes a new cluster analysis model, and clusters the operation type data of the network security management platform, uses historical data to intelligently train the K-means clustering model, obtains the risk level classification of the operation behavior through training, and then introduces the real-time data into the model for detection, which can judge the operation event’s degree of danger. The research results of this project have further improved the intelligence level of the network security management platform through the application of big data analysis technology in the network security platform, and provided technical support for the intelligent protection of network security.

References
[1] Shaoxian T and Wenjun C 2008 F.Intrusion detection based on unsupervised clustering and hybrid genetic algorithm J.Computer Applications 02 409-411
[2] Srinivas M and Patnaik M 1994. Adaptive probabilities of crossover and mutation in genetic algorithm. Man and Cybernetics
[3] KDD Cup1999 Data 2007 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
[4] Yongguo L, Kefei C, and Xiaofeng L 2004 A Genetic Clustering Method for Intrusion
Detection. *Pattern Recognition*

[5] Shengyi J, Qinghuai L, Hui W and Zhonglou M 2005 Clustering-Based and Supervised Intrusion Detection Method *J. Mini-Micro Systems* 06

[6] Tiezhu L, Jian-cheng L and Ye W 2002 F A Novel Clustering-Based Method to Network Intrusion Detection *J. Journal of National University of Defense Technology* 02

[7] Schultz M, Eskin E and Zadok E 2001 Data mining methods for detection of new malicious executables *Proceedings of IEEE Symposium on Security and Privacy (IEEE S&P)*

[8] Eskin E 2000 Anomaly detection over noisy data using learned probability distributions *International Conference on Machine Learning*

[9] Portnoy L, Eskin E and Stolfo S J 2001 Intrusion detection with unlabeled data using clustering *In Proceedings of ACM CSS Workshop on Data Mining Applied to Security (DMSA)*