Clustering Analysis based Power Security Big Data Aggregation in Ubiquitous Power Internet of Things

1st Guofei Guan
Jiangsu Fangtian Power Technology Co., Ltd
Nanjing, China
ggf_mail@126.com

2nd Xinyuan Hu
College of Automation & College of Artificial Intelligence
Nanjing University of Posts and Telecommunications
Nanjing, China
1219053930@njupt.edu.cn

3rd Yan Xu
Jiangsu Fangtian Power Technology Co., Ltd
Nanjing, China
13952040432@163.com

4th Qiqi Luan
Jiangsu Fangtian Power Technology Co., Ltd
Nanjing, China
704998432@qq.com

5th Chunpeng Li
Jiangsu Fangtian Power Technology Co., Ltd
Nanjing, China
chunpeng_lee@126.com

6th Feng Jiang
Jiangsu Fangtian Power Technology Co., Ltd
Nanjing, China
15850686088@139.com

***Abstract*—Data aggregation is a key technology in the research of massive security data intrusion detection, which can effectively improve the deficiencies of a large number of repeated alarms and high false alarm rates in practical applications. This paper summarizes the existing clustering analysis based power security big data aggregation, shows the importance and effectiveness of data aggregation technology for massive power security data processing, and provides a theoretical basis for the important technical problems and development trends of the current research.**

Keywords—Network security; Intrusion detection; Data aggregation; Cluster analysis

I. INTRODUCTION

With the growing of sensor nodes in Ubiquitous Power Internet of Things, a large amount of data is generated in the network. It will consume a lot of resources such as communication, energy, computing and storage to process and transmit these data [1]. Due to the limited resources of sensor nodes (energy, computing power, storage capacity), how to effectively deal with the large amount of data generated in Ubiquitous Power Internet of Things has become an urgent problem to be solved. In addition, in some sensor applications, users are not concerned about the original data of each node, but a macro feature of the overall network data [2-3]. In this context, the concept of data aggregation came into being.

Data aggregation means that the intermediate nodes in the network do not directly forward data after receiving the data of the predecessor nodes, but performs arithmetic processing on the received data and the data perceived by its own node (for example: averaging operation, sum operation , find the maximum operation, find the minimum operation, etc.), and finally generate a single data and forward it to the successor node. Data aggregation aggregates multiple data into a single data according to the operation processing rules, which can effectively reduce the energy consumption of the node and reduce the occupation of bandwidth resources. Due to the limit of the energy, calculation, bandwidth and other resources of sensor nodes, data aggregation technology can increase the life cycle of the network, and it has become a key technology in Ubiquitous Power Internet of Things [4].

Meanwhile, in order to cope with the increasingly prominent network security issues, the research in recent years on intrusion detection systems and various network security products has gradually shifted from improving the efficiency and accuracy of security event alarms to the correlation analysis of security events , providing a more powerful technical support for the overall control of the network security situation [5]. How to associate low-level alarms and network security events with more useful information for network security managers has become the key point of the research in the network security area. In the process of data generating information and then knowledge, there are mainly two important transformation processes: on the one hand, it is from data to information conversion, which mainly uses related technologies of data aggregation to transform the basic data generated by the collected information system into an organized and meaningful information database to facilitate further analysis; on the other hand, it is from information to knowledge conversion, mainly through multi-source information association and aggregation technology. In order to help managers better make corresponding decisions, we transform the machine-generated original information into the effective knowledge which make decision makers easier to understand and grasp the overall situation [6].

The theoretical basis of power security big data aggregation includes data integration technology and multi-source information aggregation technology. The key problem solved by data integration is how to integrate interrelated distributed heterogeneous data sources, so that users can access these data sources in a transparent manner. The key problem to be solved by multi-source information aggregation is how to combine information from various aspects in a regular way, thus pushing out more valuable information.

This article describes several common clustering analysis based data aggregation algorithms, and makes detailed and specific comparison and analysis of related data aggregation schemes, which can provide reference and guidance for current and future related research.

II. CLUSTER ANALYSIS AND DATA MONITORING

In a complex network environment, security data will be dispersed in different system environments in different time series, describing the same or different security information, and lacking a unified and standardized description of each

Scientific Research Project of Jiangsu Fangtian Power Technology Co., Ltd （KJ201917）

978-1-7281-5748-1/20/$31.00 ©2020 IEEE

other. The knowledge between safety information is difficult to effectively organize and interact in a unified manner, which leads to great difficulties in the judgment and handling of safety events [7]. Security data aggregation is essentially a process of information aggregation, which means that high-volume, low-value alarms and logs generated by one or more security devices are converted, deduplicated, merged, and related operations to form low-volume, high-value security incidents or security knowledge for higher-level security analysis.

Whether it is to use outlier analysis to find outliers, or to construct a classifier to classify intrusion events, it is to detect intrusion by training on unknown data, and it is an anomaly detection algorithm based on unsupervised cluster learning. Unsupervised clustering is a learning method based on statistical theory. Its biggest feature is that according to the principle of Vapnik structural risk minimization, the generalization ability of the learning algorithm is improved as much as possible, and the unsupervised clustering algorithm which can be guaranteed that there is still a high detection rate in the case of insufficient prior knowledge is applied to intrusion detection, so that the intrusion detection system has better detection performance [8]. The clustering algorithm performs feature analysis on the training data set and divides similar data into the same class.

The core idea of cluster analysis is to divide the target objects distributed in space into several subsets according to the detailed division rules. Each subset can be classified into specific categories according to the characteristics of the target. The cluster analysis algorithm is relatively subjective, mainly relying on prior knowledge or predefined functions to judge the quality of the clustering results. Therefore, in order to obtain more accurate clustering results, the effectiveness and repeatability of clustering analysis algorithms must be analyzed before use [9].

The current mainstream clustering algorithms mainly include fuzzy K-means clustering, artificial neural network clustering, Bayesian clustering and cross-modal multi-source heterogeneous aggregation methods. Among them, the characteristics of fuzzy logic are very suitable for uncertain expression and reasoning. Neural networks have good knowledge generalization and structural fault tolerance. Bayesian estimation overcomes the shortcomings of traditional probability and statistical methods that do not rely on sample data but determine accuracy and reliability subjectively or empirically. The following will introduce and analyze these methods in detail.

III. CLUSTERING ALGORITHM ANALYSIS

A. Fuzzy K-means Clustering

Currently, the traditional K-means clustering algorithm proposed by JB. MacQueen is one of the most common clustering algorithms used in intrusion detection systems. Given the value of k, the traditional K-means clustering algorithm iterates repeatedly until the square error criterion function converges, and outputs k clustering result sets. The algorithm is relatively scalable and efficient, and has significant effectiveness in processing large data sets. It is one of the classic algorithms for solving clustering problems.

The traditional K-means iterative algorithm is a clustering algorithm based on the target function, which can minimize the clustering between cluster samples and clustering centers to form K clusters. If the samples within clusters are close to each other and the samples between clusters are far away, K-means can be used to obtain the desired clustering results. However, such data clusters are encountered in real applications [10]. For example, the points farther from the cluster center in the data set contributing to the update of the cluster center are the same as the points closer to each other, because their "weights" are all 1. However, distant points may lead to undesirable clustering results. In order to overcome the shortcomings of K-means, a clustering algorithm based on FKM is proposed, which considers the weight (membership) to control the degree of contribution of data samples in the update process of the cluster center. The samples assigned to larger membership values make a greater contribution to updating the cluster centers, and those samples far from the cluster centers are given lower membership values to reduce their impact on the clustering results [10].

Fuzzy K-means algorithm (FKM) is to classify clustering into a nonlinear programming problem with constraints, and obtain fuzzy division and clustering of pattern sets through optimal solution. The FKM algorithm iteratively optimizes the objective function, which assigns fuzzy membership to sample data and updates the clusters according to the assigned membership. The assigned degrees of membership play the role of weight values, that is, they represent the degree to which the sample contributes to update the cluster center. The amount of contribution depends on the choice of fuzzifier m. The FKM algorithm steps are as follows:

$$v_j = \frac{\sum_{i=1}^{N} u_j(x_i)^m x_i}{\sum_{i=1}^{N} u_j(x_i)^m}, j = 1, ..., C \qquad (1)$$

And

$$u_j(x_i) = \frac{1}{\sum_{k=1}^{C} (\frac{d_{ji}}{d_{ki}})^{2/(m-1)}} \qquad (2)$$

Among them, the distance d_{ji} （d_{ki}） represents the distance between the cluster center v_j （v_k） and the sample x_i . Equation (2) represents the fuzzy membership of the sample x_i , which depends on the relative distance between the sample x_i and all cluster centers and the choice of fuzzifier m .

B. Artificial neural network clustering

Artificial neural network is a competitive learning algorithm. A network formed by a large number of neurons (or processing units) connected to each other is named after simulating the structure of the human brain nervous system. Deep learning is an artificial neural network with many levels and more complex structures. The rapid development of deep learning has taken artificial neural network related research to a higher level [11]. The clustering method needs to have three basic conditions: (1) In order to make each input sample have a different output in the network as possible, except for some randomly distributed parameters, each unit must be the same; (2) Each unit has limited strength; (3)There are certain competitive mechanisms among the units. The competition function can be expressed as:

$$\begin{cases} S_i = \omega_j^{\,T} X = \sum_{i=1}^{4} \omega_{ji} x_i \\ Y_i = f(S_i - \theta) \end{cases} \tag{3}$$

Among them: ω_j is a weight vector, X is the input sample, S_i is the input of each neuron, and the output of the neuron Y_i is the mapping of the input and the threshold of this neuron on the transfer function. After building a network model, the artificial neural network learns knowledge from the input data to adjust the weight vectors and thresholds of the neurons until the output error of the network (mostly the norm of the output and expected results) reaches the expectation and the training ends. The parameter adjustment methods of artificial neural network can be generally divided into two types: grid search and random search. In addition, using artificial intelligence optimization algorithms to optimize parameters is also a very promising way [12].

Competitive learning methods can be divided into two types: hard competitive learning and soft competitive learning. The hard competition learning mechanism uses the optimal weight vector to match the input pattern, while the soft competition learning uses the similarity to match the input pattern. Self-Organizing Map (SOM) [13] is an algorithm that uses artificial neural networks for clustering. This method processes all the sample points one by one, and maps the cluster center to the two-dimensional space, so as to realize visualization. Yin [14] proposed an improved Visualization Self-Organizing Map (VISOM) to greatly improve the visualization characteristics of the traditional SOM algorithm. The artificial neural network clustering based on projection adaptive resonance theory proposed by Cao et al. [15] further improves the performance of the algorithm. Deep learning is another leap-forward development of artificial neural networks. At present, the clustering algorithm based on deep learning has also become a hot issue of research [16-18].

C. *Bayesian clustering*

Anomaly detection method based on Bayesian clustering refers to the discovery of different sets of data in the data. These classes reflect the basic generic relationship. Members of the same class are more similar than other members, so that abnormal user classes can be distinguished, and intrusion events can be inferred. The Autoclass Program proposed by Cheeseman and Stutz in 1995 is an unsupervised data classification technology. Autoclass applies Bayesian statistical technology to search and classify the given data. Its advantages are: (1) according to the given data, automatically determine and determine the number of types; (2) does not require special similarity measurement, pause rules and clustering criteria; (3) can be continuous attributes and discrete attributes [19- 20].

Statistics-based anomaly detection methods classify observed behaviors. So far, the technology used has mainly been supervised classification, which is to establish user behavior profiles based on observed user behavior. While the Bayesian classification method allows the ideal number of classifications, user groups with similar profiles, and compliance with natural classifications that conform to the user's feature set.

D. *Cross-model multi-source heterogeneous polymerization*

The multi-source heterogeneous security monitoring aggregation algorithm builds an adaptive method for various types of security monitoring data domains to make a more comprehensive description of network threats, and further digs out potential unknown network threats. The original data source domain S contains various types of security monitoring data with tags, security alarm data S_1, network log data S_2 and system log data S_3, etc., $S=\{S_1, S_2, ..., S_n\}$, where the collection of tags can be Expressed as Y_s. At the same time, the unknown network of label set Y_t is the target domain t. Suppose that the edge distribution $P_s \neq P_t$ and the conditional distribution $P(Y_s|S_1) \neq P(Y_s|S_2) \neq P(Y_s|S_3)$ $\neq \cdots \neq P(Y_s|S_n) \neq P(Y_s|t)$. The target of the first stage is to obtain the converted source domain S_{new1} and the converted target domain S_{new2} by looking for a subspace that does not change the view. The goal of the second stage is to use the converted source domain S_{new1} to predict the label Y_t of the unknown network target domain t, that is, a potential unknown network threat. Through the adaptation of the domain, the offset between the edge distribution (P_s, P_t) and the conditional distribution $\left(P(Y_s|S_1), P(Y_s|S_2), P(Y_s|S_3)\right.$ $\left., ..., P(Y_s|S_n), P(Y_s|t)\right)$ between the domains is reduced. Next, by adaptively adjusting these two offsets to construct adaptive methods for various types of security monitoring data domains, a more comprehensive description of cyber threats is made. With the help of Nonnegative Matrix Factorization, a robust view-invariant subspace is obtained. In the fusion stage of multi-source heterogeneous data, it explores subspaces that do not change with modalities for multi-source heterogeneous data. In particular, we hope that in the low-dimensional space, the mixed categories and the structures that change with the mode can be disassembled, and these structures can be better guided by the regularization of the added graph. In the stage of adaptive distribution embedding, the algorithm will increase the importance of conditional distribution and edge distribution.

IV. ALGORITHM COMPARISON ANALYSIS

K-means clustering is widely used for large-scale data with its fast convergence speed, and the ability of being easily expanded; the disadvantage is that it tends to aggregate the number of samples relatively close to each other, with density as the main means of differentiation to perform cluster analysis, and the selection of initial cluster centers and outlier (or noise) data will have a more obvious impact on cluster analysis. Because clustering generally judges the similarity between samples as a whole, it is not suitable for clustering high-dimensional data sets [10].

Artificial neural network clustering is a clustering algorithm based on machine learning. The main advantages include: (1) The flexibility of clustering, which can theoretically reduce the error infinitely by adjusting the parameters; (2) The parallelization of clustering operations. According to the different dimensions of the data set, the number of input nodes can be adjusted for parallel calculation; (3) The algorithm is easy to implement. At the same time, the main disadvantage of the algorithm is that it is prone to learning problems and unstable clustering results [15].

Bayesian clustering is currently limited to theoretical discussions and has not yet been applied in practice. It has not been well solved that how the automatic classification program handles inherent sequential data, and how to consider the statistical distribution characteristics in the classification. Due to the inherent characteristics of statistical methods, automatic classification programs also have problems such as the selection of anomaly thresholds and preventing attackers from interfering with the type distribution [20].

Cross-modal multi-source heterogeneous polymerization algorithm is a newer adaptive method for constructing various types of security monitoring data domains, which can make a more comprehensive description of network threats. The algorithm explores the modal invariant subspace of the security alarm data and the network log data in the original data through decomposition and subspace learning. In the adaptive distribution embedding phase, the decision boundary of the aligned target domain and the converted source domain can be obtained thereby preventing the samples in the target domain from being erroneously estimated.

V. APPLICATION AND CHALLENGES OF DATA AGGREGATION TECHNOLOGY

A. Application of clustering technology in power system

The power distribution Internet of Things management platform can not only implement platform functions such as link management, data processing, device management, application management, and identity management. Platform information security is the primary prerequisite for ensuring business and data security and supporting the company's application development [21]. In response to the need of the current massive safety data monitoring, the project carries out research work on massive safety data management and analysis, and massive safety monitoring data aggregation technology. At present, in order to avoid network data security issues, a series of security devices such as firewalls, intrusion detection systems and antivirus software detection systems have been deployed in the State Grid environment to provide a full range of detection capabilities for detecting attacks and threats in the network environment. However, the simple accumulation of security equipment has not completely solved the various problems in the complex network environment, and even a large amount of redundant alarm information has disturbed the administrator's timely response to important attacks in the system [22]. In order to cope with these problems and help administrators analyze and identify attacks and threats from massive security data, it is necessary to effectively manage and analyze these security data. Among them, the aggregation of multi-source data to reduce redundant information and tap the connection of information between various devices is the main solution.

Clustering is an important data division method. The objects in the same cluster are similar to each other, but different from the objects in other clusters. In the era of big data, clustering is one of the effective ways to deal with big data because of the characteristics of less manual intervention, data driving and strong adaptability of big data. Clustering technology is also widely used in data processing of multiple links in the power system. Reference [13] proposed to divide the high-speed collected massive recording oscillograph data of fault by semi-supervised clustering technology to realize the effective screening of the fault data set in the multi-channel massive recording oscillograph data of fault and reduce the manual workload. Literature [17] proposed an electric vehicle emergency power supply aggregation scheme based on fuzzy K-means algorithm for the mobile energy storage characteristics of electric vehicles. In [4], it is difficult to deal with the problem of using electricity information to collect big data in the traditional way. The MapReduce K-means clustering technology is studied to implement data-driven user classification. However, these technologies mainly use static clustering schemes, which are suitable for offline data and cannot be directly used for data stream processing.

B. Challenges faced by clustering technology

Under the situation of massive security monitoring data aggregation technology in the post-processing process of alarm information filtering, aggregation and correlation, it is a difficult problem to solve that how to maintain or improve its detection rate, while reducing the false alarm rate. Due to the different detection technologies used in various types of security equipment and the different methods for classifying attacks, there is no obvious attack feature that can cover all attacks. Therefore, it is the key technology of project research to realize the real cross-heterogeneous security equipment alarm correlation. How to process the alarm information more deeply on the basis of aggregation and association, including intrusion intention recognition based on further association between different associated systems, and how to integrate the correlation results into the system security status assessment, and combine the alarm correlation system with the automatic intrusion response system. In view of the opportunities and challenges facing the existing smart grid, we analyze from the following three aspects [23-24]:

(1) How to resolve the contradiction between the quality of alarm information and the number of alarms in the association. According to the specific situation, how to maintain or increase the detection rate while reducing the false alarm rate during the post-processing of alarm information filtering, aggregation and correlation is a difficult problem to solve.

(2) In practical applications, there will be characteristics of complexity and diversity in the data set. Choosing any kind of clustering algorithm may not be applicable. Therefore, we can research on multiple algorithm fusion problem on the basis of understanding the advantages and disadvantages of the basic clustering algorithm.

(3) As the application of cloud computing, Internet of Things and big data technology matures, in this case, the complexity of the data to be clustered is unprecedented. Therefore, in the context of complex data, how to explore the effectiveness of clustering is also an important and difficult research hotspot.

VI. CONCLUSIONS

With the rapid development of network technology, the current network attacks are showing diversified and complicated characteristics, which put forward higher requirements for alarm aggregation and correlation technology, showing the following development trends: (1) Intelligent processing methods, flexible response to complex networks environment and attack methods. (2) Association rule mining which means that how association rules are generated and used in alarm association, network situation analysis, etc. (3) Visualization which means to assist the security administrator to analyze the attack patterns and intentions hidden behind a large number of warning messages. (4) Situation assessment, that is, how to quantify the warning information to analyze the network situation and get a more accurate and reliable security state of the system. (5) Intrusion response, judging the actual intrusion and attack behavior, reducing the response times and economic costs [24].

In short, data aggregation technology requires features such as accuracy, real-time, strong adaptability, and good scalability, which can more effectively improve the applicability of massive safety data monitoring systems and other safety equipment to enhance the overall network security protection and emergency response capability.

REFERENCES

[1] Yijie Zhu, Yulong Yang, Shuai Li, et al. Research on Network Security Situation Awareness Platform for Big Data Environment [J]. Network Security Technology & Application, 2018, 215(11):55-57.(in Chinese)

[2] Goryczka S, Xiong L. A Comprehensive Comparison of Multiparty Secure Additions with Differential Privacy[J]. IEEE Transactions on Dependable and Secure Computing, 2015: 1-1.

[3] Sijia Zhang, Chunhua Gu, Mi Wen. Research on Classification of Data Aggregation Scheme in Smart Grid [J]. Computer Engineering and Applications, 2019, (12): 83-89. (in Chinese)

[4] Xueyan Liu, Qiang Zhang, Zhanming Li, et al. Data aggregation and access control method for smart grid communication system [J]. Automation of Electric Power Systems, 2016, 40(14): 135-144. (in Chinese)

[5] Bao H, Lu R. A lightweight data aggregation scheme achieving privacy preservation and data integrity with differential privacy and fault tolerance[J]. Peer-to-Peer Networking and Applications, 2017, 10(1): 106-121.

[6] Abdallah A, Shen X S. A Lightweight Lattice-Based Homomorphic Privacy-Preserving Data Aggregation Scheme for Smart Grid[J]. IEEE Transactions on Smart Grid, 2017, 9(1): 396-405.

[7] Gopikrishnan S, Priakanth P. HSDA: hybrid communication for secure data aggregation in wireless sensor network[J]. Wireless Networks, 2017, 22(3): 1-18.

[8] Talburt J R, Pullen D, Penning M. Evaluating and Improving Data Fusion Accuracy[M]. 2019.

[9] Wang T, Qin X, Ding Y, et al. Privacy-Preserving and Energy-Efficient Continuous Data Aggregation Algorithm in Wireless Sensor Networks[J]. Wireless Personal Communications, 2018, 98(1): 665-684.

[10] Liu Y, Liu C, Zeng Q A. Improved trust management based on the strength of ties for secure data aggregation in wireless sensor networks[M]. Kluwer Academic Publishers. 2016.

[11] Chen J, Chen Y, Du X, et al. Big data challenge: a data management perspective[J]. Frontiers of Computer Science, 2013, 7(2): 157-164.

[12] Shangce Gao, Mengchu Zhou, et al. Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(2): 601-614.

[13] Manyika J, Chui M, Brown B, et al. Big data: The next frontier for innovation, competition, and productivity[J]. 2011.

[14] YIN H. VISOM: a novel method for multivariate data projection and structure[J]. IEEE Transactions on Neural Networks, 2002, 13(1): 237 -243.

[15] CAO Y, WU J. Dynamics of projective adaptive resonance theory model: the foundation of PART algorithm[J]. IEEE Transactions on Neural Network, 2004, 15(2): 245 -260.

[16] ZHANG Y, LU J, LIU F. Does deep learning help topic extraction? a kernel k-means clustering method with word embedding[J]. Journal of Informetrics, 2018, 12(4): 1099 -1117.

[17] HAN J, TAO J, WANG C. FlowNet: a deep learning framework for clustering and selection of streamlines and stream surfaces [J]. IEEE Transactions on Visualization and Computer Graphics, 2018, 11 ： 678 -689.

[18] ZHAO Z, BARIJOUGH K, GERSTLAUSER A. DeepThings: distributed adaptive deep learning inference on resource-constrained IoT edge clusters[J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2018, 37(11): 2348 – 2359.

[19] ZHANG Y, ZHOU X, SHI H, et al. Corrosion pitting damage detection of rolling bearings using data mining techniques[J]. International Journal of Modeling, Identification and Control, 2015, 24 (3): 235 - 243.

[20] ZHOU Y, YU J, WANG X. Time series prediction methods for depth-averaged current velocities of underwater gliders [J]. IEEE Access, 2017, 5: 5773 -5784.

[21] GAO J, WANG Y, LI J. Bounds on covering radius of linear codes with Chinese Euclidean distance over the finite non chain ring F 一 2 + vF(2) [J]. Information Processing Letters, 2018, 138: 22-26.

[22] ANTER A, HASSENIAN A E, OLIVA D. An improved fast fuzzy c-means using crow search optimization algorithm for crop identification in agricultural[J]. Expert Systems with Applications, 2019, 118: 340 - 354.

[23] ZHAN J, WANG R, YI L. Health assessment methods for wind turbines based on power prediction and Mahalanobis distance[J]. International Journal of Pattern Recognition and Artificial Intelligence, 2019, 33(2):1951001.

[24] ANTOINE G B, CATHY M R, ANDREA R. Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data[J]. Journal of Applied Statistics, 2019, 46(1): 47 -65.

[25] Zongke Geng, Changbin Wang, Zhenguo Zhang. Fuzzy clustering algorithm based on fuzzy c-means and adaptive particle swarm optimization [J]. Computer Science, 2016, 43(8): 267-272. (in Chinese)