Review of Anomaly Detection Based on Log Analysis

Wu Xudong
Laboratory of Wireless Network and Intelligent System
Xi'an Technological University
Xi'an, 710021, China
E-mail: wuxudong_wxd@163.com

***Abstract*—The development of the Internet and the emergence of large-scale systems promote the rapid development of society, and bring a lot of convenience to people. Then comes the problem of network security, privacy theft, malicious attacks and other illegal acts still exist, a qualified software system will log the key operation behavior of the software. Therefore, log analysis has become an important means of anomaly detection. Based on log analysis, this paper consulted the related literature on anomaly detection, elaborated the research status of anomaly detection based on log analysis from the aspects of template matching, rule self-generation and outlier analysis, and analyzed the challenges faced by anomaly detection based on log analysis.**

Keywords-Log Analysis; Distributed; Big Data; Anomaly Detection

I. INTRODUCTION

With the development of the Internet, big data and artificial intelligence have penetrated into people's lives, unknowingly changing the way people live, food, and transportation, making people's lives faster, more efficient, and easier. Research in various fields of computer is moving towards bionics, including human-like big data processing, human-like computer vision and image processing, human-like voice input, etc. These studies make the computer in a domain not only Clairvoyance, Shunfeng ear, can also save and process a large number of various types of data obtained from various aspects, forming an invisible "superman" individual.

In the past 20 years, with the rapid development of the Internet in China, people's lifestyles have undergone tremendous changes. Chinese Internet users continue to grow. According to CNNIC's 44th "Statistical Report on Internet Development in China", as of June 2019, the number of Internet users in China reached 854 million, an increase of 25.98 million from the end of 2018, and the Internet penetration rate reached 61.2%, compared with the end of 2018. An increase of 1.6 percentage points. The proportions of using desktop computers, laptop computers and tablet computers to surf the Internet were 46.2%, 36.1% and 28.3% respectively. These not only reflect the continuous increase in the number of netizens, but also the rapid and continuous growth of log data from the side.

The log records the time point selected by the developer that is worthy of attention and the changes of state or event that is worthy of attention at this point in time. It is the most important source of information for understanding the operating status of the system and diagnosing system problems. Traditionally, system maintainers use tools such as grep and awk to filter keywords such as "error" or

DOI: 10.21307/ijanmc-2020-036

"exception" in the log to find problems in system operation. When the filtering keywords cannot meet the demand, more experienced personnel will write scripts to impose more complex filtering rules. The cost of this method is very high, writing effective scripts requires a deep understanding of the target system, and these scripts written for specific target systems cannot be applied to other systems, and their versatility is poor. But even without considering the cost, this approach has become no longer feasible for today's software systems.

The ever-increasing log data scale and network security issues make network managers face severe challenges: not only need to ensure the stable and efficient operation of the network, but also need to provide secure network services as much as possible. Fortunately, in recent years, distributed computing technology has become more mature. Distributed computing platforms such as Hadoop, Spark, Flume, Storm are being accepted and applied by more and more companies, and are gradually being used in various industries for data storage. And online or offline analysis, which brings opportunities for log data anomaly detection.

At the same time, issues such as security and privacy in the network have also emerged. Distributed denial of service attacks, zombie codes, Trojan horses, ransomware, worms and other malicious software have a great negative impact on people’s lives. Once the malware operates, it may cause irreversible losses to the company’s economy. , Poses a great threat to people's privacy. A study showed that [1][2]: Random sample surveys of large-scale systems, more than half of the system failure problems were not logged. At this time, maintenance personnel are required to manually find the cause of the problem. Due to the large amount of code, The time invested is much more. High-quality software code can greatly help the detection efficiency after a program error occurs. Log records at key locations are an important means to ensure that the abnormality can be quickly located and repaired. Therefore, it is necessary to add log records to key positions of the program, and log analysis has become an important method of anomaly detection.

The Internet brings convenience to our life, but also brings a series of network security problems. The main characteristics of Internet security problems are as follows: a variety of types, all the time, causing huge losses. All kinds of human attacks, mis-operation, network equipment failure will bring network security problems. Distributed denial of service attack, zombie code, Trojan horse, blackmail program, worm virus and other malicious software appear frequently. Once the malicious software operates, it may cause irreparable loss to the company's economy and cause great impact on people's life.

The log records the time points that developers choose to pay attention to and the changes of states or events at this time point. It is the most important information source to understand the system operation status and diagnose system problems. Traditionally, system maintenance personnel use grep, awk and other tools to filter keywords in logs, such as "error" or "exception", to find problems in system operation. When the filtering keywords can't meet the requirements, senior personnel will write scripts to impose more complex filtering rules. The cost of this method is very high, writing effective scripts needs to have a deep understanding of the target system, but these scripts written for specific target system can not be applied to other systems, and the generality is very poor. But even without considering the cost, this approach is no longer feasible for today's software systems.

With the increasing scale of log data and network security issues, network managers are facing severe challenges: not only need to ensure the stable and efficient operation of the network, but also need to provide as much as possible secure network services. Fortunately, in recent years, the distributed computing

technology is becoming more and more mature. Hadoop, spark, flume, storm and other distributed computing platforms are being accepted and applied by more and more enterprises, and are gradually applied to various industries for data storage and online or offline analysis, which brings opportunities for log data anomaly detection.

This article first talks about the related knowledge of log analysis and anomaly detection, and then summarizes the current research status of log anomaly detection from the aspects of template matching, rule generation and outlier analysis, analyzes and classifies the articles that have been read, and summarizes the current The types and rules of log anomaly detection are found to be difficult to solve during the detection process. Finally, the future work of anomaly detection based on log analysis is summarized.

II. RELATED TECHNOLOGIES AND CONCEPTS OF LOG ANOMALY DETECTION

A. Log analysis

The log in the computer is a record of events generated with the operation of network equipment, applications, and systems. Each line records the date, time, type, operator, and description of related operations. Figure 1 shows a partial log record of the application. In reality, the log data generated by a system is very large, conforming to the 4V characteristics defined in big data, namely, volume, variety, velocity, and value. These log data will only occupy storage space if they are shelved, and will bring unlimited value if they are properly used. Because these log data have 4V characteristics, it also determines that manual analysis of these data is unrealistic, and log analysis tools must be used to make full use of the value of log data.

```
[2019-02-27 16:18:04] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:04] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:18:05] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:05] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:18:08] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:18:08] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:19:15] [Err] [2804] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:19:15] [Warn] [2804] utl_OpenFilterHandle Failed.
[2019-02-27 16:20:40] [Err] [2dc8] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:20:40] [Warn] [2dc8] utl_OpenFilterHandle Failed.
[2019-02-27 16:21:11] [Err] [2630] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:11] [Warn] [2630] utl_OpenFilterHandle Failed.
[2019-02-27 16:21:11] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:11] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:21:12] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:12] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
[2019-02-27 16:21:15] [Err] [684] EnumAdapters: DeviceIoControl failed with err 6.
[2019-02-27 16:21:15] [Warn] [684] Notify_AddrChange threadproc: EnableArp Fail.
2019-02-23 04:49:43.002 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-23 04:49:43.014 - Info  -      [shell] status_changed:1
2019-02-23 04:49:43.033 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-23 04:49:43.095 - Info  -      attempt to connect server cdn.oray img.com:80(125.76.247.211:80)
2019-02-23 04:49:43.097 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
2019-02-27 16:17:00.459 - Info  -      [shell] status_changed:2
2019-02-27 16:17:00.459 - Info  -      [shell] status_changed:0
2019-02-27 16:17:15.225 - Info  -      [shell] status_changed:1
2019-02-27 16:17:15.243 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:17:15.259 - Info  -      [shell] status_changed:1
2019-02-27 16:17:15.276 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:17:15.333 - Info  -      attempt to connect server cdn.oray img.com:80(125.76.247.211:80)
2019-02-27 16:17:15.347 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
2019-02-27 16:18:05.004 - Info  -      [shell] status_changed:2
2019-02-27 16:18:05.004 - Info  -      [shell] status_changed:0
2019-02-27 16:18:35.053 - Info  -      [shell] status_changed:0
2019-02-27 16:20:37.740 - Info  -      [shell] status_changed:2
2019-02-27 16:20:58.852 - Info  -      [shell] status_changed:0
2019-02-27 16:23:01.548 - Info  -      [shell] status_changed:2
2019-02-27 16:23:03.696 - Info  -      [shell] status_changed:1
2019-02-27 16:23:03.714 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:23:03.726 - Info  -      [shell] status_changed:1
2019-02-27 16:23:03.743 * Error *      [CInstallRemoteDialog] Open SunloginRemote Uninstall Reg info failed
2019-02-27 16:23:03.812 - Info  -      attempt to connect server cdn.oray img.com:80(125.76.247.211:80)
2019-02-27 16:23:03.815 +Warning+      [HTTP DOWNLOAD ON CONNECTED]
```

Figure 1. Part of the application log

Here are several current mainstream log analysis tools. Slunk is a full-text search engine for machine data and a hosted log management tool. Its main functions include: log aggregation, search, meaning extraction, grouping, formatting, and visualization of results. ELK is composed of three parts: elasticsearch, logstash, and kibana. Elasticsearch is a near real-time search platform. Compared with MongoDB, elasticsearch has more comprehensive functions and is very capable of performing full-text search. It can index, search, and sort documents, filter. Logstash is a log collection tool, which can collect various messages from local, network and other places and send them to elasticsearch. Kibana provides a visual interface on the web and has a cool dashboard.

B. Store log data

Due to the huge amount of log data and semi-structured data, the traditional structured database can not meet the storage requirements of log data. HDFS (Hadoop distributed file system) can provide high-throughput data access, which is very suitable for large-scale data sets, and it is suitable for deployment on low-cost machines, which can meet the

storage requirements of log data. In the experiment, the log data generated by the system needs to be stored in HDFS. The configured HDFS will automatically back up the data. The input data file is divided into fixed size blocks. The general size of the data block is 128MB. Each data block is stored in different nodes. Generally, each data block has three copies. The first copy is stored in the same node as the client, the second replica exists on a node in a different rack, and the third replica exists on another node in the same rack as the second replica.

C. Log data preprocessing

Log data preprocessing has three goals:

- filtering "non-conforming" data and cleaning meaningless data;
- format conversion and regularization;
- filtering and separating various basic data with different needs according to the subsequent statistical requirements.

In terms of filtering "non-conforming" data and cleaning meaningless data, the log data generated by the system may be "non-conforming" or meaningless. Before the data format conversion and normalization, a judgment needs to be added to check whether the data is standard and intentional. If not, the data is considered useless and jumps to the next data directly. In terms of format conversion and regularization, we first analyze the characteristics of the data. The fields in each record are separated in the form of spaces. According to this feature, each record is segmented according to the space as the standard. For the fields with spaces inside, we need to use regular matching for special processing. After segmentation, each field is normalized, including time format conversion, number type conversion, path completion, etc. In the aspect of filtering and separating data with different needs, the required fields are extracted according to the needs of subsequent detection algorithms.

D. Anomaly detection

Anomalies usually include outliers, fluctuation points and abnormal event sequences. Generally, given the input time series X, the outliers are timestamp value pairs (t, Xt), where the observed value xt is different from the expected value of the time series, then the observed value Xt is an outlier. Fluctuation point refers to a given input time series X, at a certain time t its state or behavior in this time series is different from the values before and after T. An abnormal time series is a part of a given set of time series X={Xi} that belongs to X but is inconsistent with most time series values on X. The abnormal point is given in the box in Figure 2.

Peng Dong et al. [3] divided anomaly detection methods into three categories: techniques based on statistical models, techniques based on proximity, and techniques based on density.

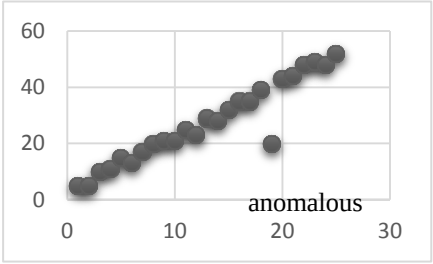

Figure 2. Outlier feature

In data mining, anomaly detection identifies items, events, or observations that do not match the expected pattern or other items in the dataset. Usually, abnormal items will turn into bank fraud, structural defects, medical problems, text errors and other types of problems. Anomalies are also known as outliers, novelty, noise, bias, and exceptions.

Especially in the detection of abuse and network intrusion, interesting objects are often not rare objects, but they are unexpected activities. This pattern does not follow the usual statistical definition of outliers as rare objects, so many anomaly detection methods will fail to deal with such data unless appropriate aggregation is carried out. On the contrary, clustering analysis algorithm may be able to detect the micro

clustering formed by these patterns.

There are three types of anomaly detection methods. Under the assumption that most instances in the dataset are normal, the unsupervised anomaly detection method can detect the unlabeled test data by finding the most unmatched instance with other data. Supervised anomaly detection requires a dataset that has been labeled "normal" and "abnormal" and involves training classifiers (the key difference from many other statistical classification problems is the inherent imbalance of anomaly detection). The semi supervised anomaly detection method creates a model representing normal behavior based on a given normal training data set, and then detects the possibility of test cases generated by the learning model.

III. RESEARCH STATUS OF LOG ANOMALY DETECTION TECHNOLOGY

Anomaly detection refers to the process of finding data patterns that do not meet expectations from the data [4]. Anomaly detection behavior based on log data can be regarded as a classification problem in essence, that is to say, distinguish normal behavior and abnormal behavior from a large amount of abnormal log behavior data, and determine the specific attack method from the abnormal behavior [5]. When the server is running, the log will record and generate the behavior of the user throughout the access process. You can find the information of abnormal users by processing the information in the log. Therefore, analyzing logs has become one of the most effective methods to detect abnormal user behavior [6,7,8]. With the rapid development of big data, log-based anomaly detection methods are divided into three categories: model-based technology, proximity-based technology and density-based technology.

A. Model-based technology

Model-based technology first builds a data model. Anomalies are objects that the model cannot fit perfectly. Since abnormal and normal objects can be regarded as defining two different classes, we can use classification techniques to build models of the two classes. However, the training set is very important in the classification technology. Because anomalies are relatively rare, it is impossible to detect new anomalies that may appear [9]. Wang Zhiyuan et al. [10] used the log template to detect anomalies in 2018. The log was cleaned first, and then the edit distance was used to cluster the text to form the log template. On the basis of the log template, TF-IDF (Word Frequency-Inverse File) was used. Frequency) to form a feature vector, and then use logistic regression, Bayesian, support vector machine and other weak classifiers to train to obtain the score feature vector, build a strong classifier based on the score feature vector and random forest, and finally use mutual information to detect the truth The correlation between the template and the clustering template, the accuracy and recall rate are used to detect the classification effect and various classifiers are compared. Siwoon et al. [11] proposed a new data storage and analysis architecture based on Apache Hive to process a large amount of Hadoop log data, using average movement and 3-sigma technology to design and implement three anomaly detection methods, these three methods They are the basic method, linear weight method and exponential weight method. The first method calculates the average line and standard deviation of anomaly detection, but there are repeated detections. In order to solve this problem, there are two other weighted detection methods, namely linear weighting and exponential weighting. In the linear weighting method, the weight is given in proportion to the position of the log item, and the exponential weight method is to give the weight exponentially on top of the basic method. Finally, the effectiveness of the proposed method is evaluated in a hadoop environment with a name node and four data nodes. Fu et al. [12] proposed a technique that does not require any specific application knowledge for anomaly detection in unstructured system logs,

including a method of extracting log keys from free text messages. The false positive rate of their experiments under the Hadoop platform is about 13%. Xu et al. [13] used source code to match the log format to find the relevant variables, extracted the features of the corresponding log variables through the bag-of-words model, and then used these features to reduce the dimensionality through the principal component analysis method, according to the maximum separability of principal component analysis Detect abnormal log files, and finally use a decision tree to visualize the results. Fronza et al. [14] used the operation sequence represented by the random index, characterized the operation in each log according to its context, and then used the support vector machine to correlate the sequence to the fault or non-fault category to predict system failure . Peng et al. [15] applied text mining technology to classify messages in log files as common cases, improved classification accuracy by considering the time characteristics of log messages, and used visualization tools to evaluate and verify the effective time for system management mode.

B. Technology based on proximity

Proximity-based technology considers proximity measures between objects, such as "distance". Zhang Luqing et al. [16] proposed a web attack data mining algorithm based on the anomaly degree of outliers, which first clustered HTTP requests, and then proposed a detection model that approximates normal distribution. The algorithm first finds the arithmetic mean of each numerical attribute value and the most frequently occurring value in each categorical attribute as the centroid of the numerical attribute and the centroid of the categorical attribute, and after synthesis, the centroid T of the data set is obtained, and the distance between the object p and the centroid T is obtained. As the abnormality of p. Finally, experiments have confirmed that the algorithm has a higher detection rate. Jakub Breier et al. [17] proposed a log file anomaly detection method, which dynamically generates rules from certain patterns in sample files and can learn new types of attacks while minimizing the need for human behavior. The implementation uses the Apache Hadoop framework to provide distributed storage and distributed data processing to support parallel processing to speed up execution. Since the incremental mining algorithm based on the local outlier factor requires multiple scans of the data set, Zhang Zhongping et al. [18] proposed a flow data outlier mining algorithm (SOMRNN) based on inverse k nearest neighbors. Using the sliding window model to update the current window requires only one scan, which improves the efficiency of the algorithm. Grace et al. [19] used data mining methods to analyze Web log files to obtain more information about users. In their work, they describe the log file format, type and content, and provide an overview of the Web usage mining process. Liang Bao et al. [20] proposed a general method for mining console logs to detect system problems. First give some formal problem definitions, and then extract a set of log statements in the source code and generate a reachability graph to show the reachability of log statements. After that, log files are analyzed to create log messages by combining information about log statements with information retrieval techniques. The grouping of these messages is tracked according to the execution unit. A detection algorithm based on probabilistic suffix tree is proposed to organize and distinguish the significant statistical characteristics of sequences. Experiments were conducted on the CloudStack test platform and Hadoop production system, and the results showed that compared with the existing four algorithms for detecting abnormalities, this algorithm can effectively detect abnormal operation. Since there are fewer abnormal points in reality, Liu et al. [21] proposed an anomaly detection algorithm based on isolation. The isolation tree created can quickly converge but requires sub-sampling to achieve high accuracy.

C. Density-based technology

Density-based technology considers objects in low-density areas as abnormal points. The density-based local outlier detection algorithm (LOF) has high time complexity and is not suitable for outlier detection of large-scale data sets and high-dimensional data sets. Wang Jinghua et al. [22] proposed a local outlier Point detection algorithm NLOF. Li Shaobo et al. [23] proposed a density-based abnormal data detection algorithm GSWCLOF. The algorithm introduces the concept of sliding time window and grid. In the sliding time window, the grid is used to subdivide the data, and the information entropy is used for all The data in the grid is pruned and filtered to eliminate most of the normal data, and finally the outlier factor is used to make a final judgment on the remaining data. Wang Qian et al. [24] proposed a density-based detection algorithm, which introduced the Local Outlier Factor (LOF), and judged whether the data is abnormal based on the LOF value of the data. The algorithm is only suitable for static data detection. Once the amount of data fluctuates, it is necessary to recalculate the LOF value of all data. The algorithm has poor adaptability and is not suitable for the detection process of dynamic data. Pukelsheim et al. [25] assumed that the data sample obeys a univariate Gaussian distribution, and judged the test sample that is outside of the distance twice or three times the variance as abnormal.

IV. CHALLENGES FACED BY LOG ANOMALY DETECTION TECHNOLOGY

There are several obstacles from the time the system abnormality occurs to the successful detection of the abnormality:

- The exception log is not recorded
- The format of exception log records is not standardized
- The exception log cannot be sent to the processing end in time
- Abnormal log sending is lost
- The detection algorithm is not accurate enough

Any occurrence of one or more of the above conditions will result in failure of the anomaly detection result.

A. Real-time

The purpose of anomaly detection is to find anomalies and find a corresponding method to float the anomaly, and the time delay from logging, to anomaly detection, to manual analysis, and to anomaly elimination is too long, which leads to anomalies that exist for too long. The losses were more serious. If real-time performance can be guaranteed, the efficiency of exception elimination will be greatly improved.

B. Detection accuracy

Anomaly detection has various factors that affect its accuracy, such as irregular log format, inappropriate algorithm, etc. These problems directly lead to a decrease in the accuracy of anomaly detection, which also determines that log anomaly detection cannot be completely separated from the intervention of technicians.

Even if the same benchmark data set is used in the literature for anomaly detection, most of them do not indicate the size or proportion of labeled data. Even the size of training and test sets and evaluation indicators are different. Different measurement combinations make the research results unable to compare with each other

C. The versatility of detection algorithms

At present, there are many anomaly detection algorithms at home and abroad, such as: Isolation Forest, One-Class SVM, Robust covariance, K-means, Principal Component Analysis, 3-ε, etc. These algorithms have their own advantages and disadvantages and are not suitable for all anomaly detection. However, due to its unstructured and non-identical characteristics of logs, a specific algorithm is needed for a specific log, or a specific

algorithm is improved to achieve a higher detection rate. The "localization" of the algorithm also requires specialized technical personnel to operate, which increases the cost of detection.

D. Tag data

In the log data, there is a large amount of data, and there are very few abnormal data. It is very difficult to mark a small amount of abnormal data in a large amount of data. There is no such publicly marked data as the experimental basis, so anomaly detection encountered great difficulties.

V. RESEARCH DIRECTION OF ANOMALY DETECTION

Based on the current research status of anomaly detection technology and the above problems, the challenges and future research directions of anomaly detection are summarized as follows:

Traffic data often have high characteristic dimensions, and the Euclidean distance in the sampling method can not measure the spatial distribution of the samples very well. The data distribution environment of supervised learning and semi supervised learning are different. Under unbalanced data, most of the existing semi supervised methods apply the traditional methods to semi supervised learning. Therefore, the traditional methods to solve the imbalance problem are not necessarily suitable for semi supervised learning and need further research. Although the research on data imbalance has achieved good results in the field of network security, there are very few researches on the imbalance problem in semi supervised learning. Most of the semi supervised methods applied in the field of anomaly detection use ensemble learning to solve the class imbalance. In the future, we can solve the problem of anomaly detection by combining the latest achievements in the field of data imbalance under semi supervision.

At present, many network traffic feature selection and extraction are limited to one dimensional features or simple combination of multi-dimensional features, while traffic anomalies usually show in multi-dimensional features. How to effectively fuse multi-dimensional features, learn data flow features from multiple perspectives, and use a small amount of labeled data for semi supervised integration algorithm synthesis results to reduce information loss is a challenging research topic.

Semi supervised dimensionality reduction is a feasible method in the field of anomaly detection. How to find a more effective way to deal with high-dimensional sparse samples and continuous variables and further improve the real-time performance of detection model is of great significance.

The learning effect of the combination of active learning and semi supervised learning strategy is better than that of single method. The combination of semi supervised learning and active learning can actively find effective supervision information. Through effective supervision information, unlabeled sample data can be used better, thus improving the accuracy of the model and solving speed. However, the research on the combination of semi supervised learning and active learning is rare, and there is a large space for improvement.

Incremental semi supervised anomaly detection is more in line with the actual anomaly detection. It makes full use of the data results processed before in the training process. It should have more in-depth research in the field of network security. In the future, we can consider introducing the incremental algorithm of natural language technology into specific anomaly detection.

Semi supervised clustering algorithm uses the traditional clustering algorithm to introduce the supervised information to complete the semi supervised learning, so it can also expand the semi supervised clustering algorithm such as density

clustering and spectral clustering. In addition, some traffic data are high-dimensional and sparse. However, most of the existing clustering algorithms are not suitable for processing high-dimensional sparse data. In future research, it is necessary to make further discussion.

In general, semi supervised learning can help improve performance by using unlabeled data, especially when the number of labeled data is limited. However, in some cases, the selection of unreliable unlabeled data may mislead the formation of classification boundaries and eventually lead to the degradation of semi supervised learning performance. Therefore, how to use unlabeled data safely is a research focus in the future.

It can combine multiple semi supervised anomaly detection methods and technologies to achieve more efficient network data detection and obtain more accurate prediction results. In addition, in semi supervised anomaly detection, it is a challenging research topic to minimize the additional impact on the network.

VI. CONCLUSION

Machine learning faces many challenges in the field of abnormal traffic detection. The biggest difficulty is the lack of label data. In practice, only a limited number of tagged data is available, while most of the data is unmarked. In addition, although there are a large number of normal access data, there are few abnormal traffic samples and various attack forms, which make it difficult to learn and train the model. Semi supervised learning is an effective solution, which can make use of both unlabeled data and labeled data, which can alleviate this problem.

For anomaly detection based on log analysis, domestic and foreign countries have made certain progress and achieved various results. Various algorithms such as template matching, automatic rule generation, outlier analysis, and statistical data have certain effects, which are of great significance to network security and intelligent operation and maintenance.

Future research will continue to focus on real-time performance to ensure that abnormalities can be detected as quickly as possible. Improve detection accuracy, minimize manual intervention or cancel manual intervention. Study the versatility of the algorithm, so that an algorithm can adapt to log analysis in different environments as much as possible.

REFERENCES

[1] Yuan D, Park S, Huang P, Liu Y, Lee MM, Tang X, Zhou Y, Savage S. Be conservative: enhancing failure diagnosis with proactive logging. In: Proc. of the 10th Symp. on Operating Systems Design and Implementation (OSDI). 2012. 293~306.

[2] Yuan D, Park S, Zhou Y. Characterizing logging practices in open-source software. In: Proc. of the 2012 Int'l Conf. on Software Engineering. 2012. 102~112. [doi: 10.1109/ICSE. 2012.6227202].

[3] Peng Dong. Intelligent operation and maintenance: building a large-scale distributed AIOps system from zero. Electronic Industry Press, 2018.7 ISBN 978-7-121-34663-7 p198-p199.

[4] Varun Chandola, Arindam Banerjee, Vipin Kumar. Anomaly Detection: A Survey[J]. Acm Computing Surveys, 2009, 41(3).

[5] Davis J J, Clark A J. Data preprocessing for anomaly based network intrusion detection: A review[J]. Computers & Security, 2011, 30(6-7):353-375.

[6] Q. Lin, H. Zhang, J. Lou, Y. Zhang and X. Chen, "Log Clustering Based Problem Identification for Online Service Systems," 2016 IEEE/ACM 38th International Conference on Software Engineering Companion (ICSE-C), Austin, TX, 2016, pp. 102-111.

[7] Pecchia A, Cotroneo D, Kalbarczyk Z, et al. Improving Log-based Field Failure Data Analysis of multi-node computing systems[C]. IEEE, 2011.

[8] Tambe R, Karabatis G, Janeja V P. Context aware discovery in web data through anomaly detection[J]. International Journal of Web Engineering and Technology, 2015, 10(1):3.

[9] Wang Xiaodong, Zhao Yining, Xiao Haili, Chi Xuebin, Wang Xiaoning. Detection method of abnormal log flow pattern in multi-node system [J/OL]. Journal of Software: 1-15 [2019-12-24].

[10] Wang Zhiyuan, Ren Chongguang, Chen Rong, Qin Li. Anomaly detection technology based on log template[J]. Intelligent Computers and Applications, 2018, 8(05): 17-20+24.

[11] Son S, Gil MS, Moon YS. [IEEE 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)-Jeju Island, South Korea (2017.2.13-2017.2.16)] 2017 IEEE International Conference on Big Data and Smart Computing (BigComp)-Anomaly detection for big log data using a Hadoop ecosystem[J]. 2017:377-380.

[12] Fu, Q., Lou, JG, Wang, Y., & Li, J. (2009). Execution anomaly detection in distributed systems through unstructured log analysis. In Proceedings of the 2009 ninth IEEE international conference on data mining, ICDM '09, (pp. 149–158). Washington, DC: IEEE Computer Society. doi:10.1109/ICDM. 2009.60.

[13] Xu W, et al. Large-scale system problems detection by mining console logs[J]. Proceedings of the Acm Sigops Symposium on Operating Systems Principles Big Sky Mt, 2013:2009.

[14] Ilenia Fronza, Alberto Sillitti, Giancarlo Succi, Mikko Terho, Jelena Vlasenko. Failure prediction based on log files using Random Indexing and Support Vector Machines[J]. Journal of Systems and Software, 2013, 86(1):2- 11.

[15] Peng W, Li T, Ma S. Mining logs files for data-driven system management. ACM SIGKDD Explorations Newsletter, 2005, 7(1):44-51.

[16] Zhang Luqing. Web attack data mining algorithm based on outlier anomaly[J]. Ship Electronic Engineering, 2018, 38(09): 105-110.

[17] Breier J, Jana Branišová. A Dynamic Rule Creation Based Anomaly Detection Method for Identifying Security Breaches in Log Records[J]. Wireless Personal Communications, 2015, 94(3):1-15.

[18] Zhang Zhongping, Liang Yongxin. Algorithm for mining outliers in flow data based on anti-k nearest neighbors[J]. Computer Engineering, 2009, 35(12): 11-13.

[19] Grace, L., Maheswari, V., & Nagamalai, D. (2011). Web log data analysis and mining. In N. Meghanathan, B. Kaushik, & D. Nagamalai (Eds.), Advanced computing, communications in computer and information science (Vol. 133, pp. 459–469). Berlin: Springer.

[20] Liang Bao, Qian Li, Peiyao Lu, Jie Lu, Tongxiao Ruan, Ke Zhang. (2018). Execution anomaly detection in large-scale systems through console log analysis. The Journal of Systems & Software 143 (2018) 172– 186.

[21] Liu F T, Ting K M, Zhou Z H. Isolation-Based Anomaly Detection[J]. ACM Transactions on Knowledge Discovery from Data, 2012, 6(1):1-39.

[22] Wang Jinghua, Zhao Xinxiang, Zhang Guoyan, Liu Jianyin. NLOF: A new density-based local outlier detection algorithm [J]. Computer Science, 2013, 40(08): 181-185.

[23] Li Shaobo, Meng Wei, Wei Jinglei. Density-based abnormal data detection algorithm GSWCLOF[J]. Computer Engineering and Applications, 2016, 52(19): 7-11.

[24] Wang Qian, Liu Shuzhi. Improvement of local outlier data mining method based on density [J]. Application Research of Computers, 2014, 31(06): 1693-1696+1701.

[25] Pukelsheim F. The Three Sigma Rule[J]. The American Statistician, 1994, 48(2):88-91.