WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 5, Number 1, Year 2025
DOI: https://doi.org/10.62051/ijcsit.v5n1.13
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Network Asset Detection Based on Weakly Supervised Learning

Guowei Zhang, Jie Zhao, Pengyuan Ma, Junjie Wu *

School of Computer Science and Technology, Taiyuan Normal University, Jinzhong 030619, China

ABSTRACT

Network asset detection technology is the basis for sorting out and counting Internet assets, timely managing vulnerable network assets and anomaly detection. Due to various shortcomings in existing methods for obtaining device fingerprint information, the accuracy of evaluation results for network asset devices is low and the functionality is single. For this reason, proposes a network asset detection method based on weakly supervised learning, which aims to simulate the real network data in the network environment as much as possible to detect network assets, achieve more accurate results, so as to connect with the task of sorting out Internet assets and timely warning network asset vulnerabilities. The results indicate that there is a significant improvement in the speed and accuracy of asset detection when using only 30% of labeled data.

KEYWORDS

Weakly supervised learning; Network assets; Device fingerprint

1. INTRODUCTION

With the continuous development of globalization and informatization, cyberspace has become of significant importance in human life, rapidly emerging as the "fifth domain" of resources contested by humanity. The internet has now become an inseparable part of our daily lives, work, and learning, and its relationship with us is becoming increasingly inseparable, even affecting economic, cultural, military, and social activities [1]. Improving network security management is an urgent issue that needs to be addressed. The exploration and organization of websites, the internet, and other critical information, even under the complex conditions of constantly changing network assets, can still quickly form a network asset database. This database serves as the foundation for all network security work, including security detection, early warning, and protection, and is essential for quickly and accurately acquiring target network asset information and detecting abnormal assets. In today's digital age, network asset detection has become a key task to ensure network security [2].

Network asset detection is an important link in information security and network management, aimed at identifying and recording various assets in the network environment, including devices, applications, and services. Traditional network asset detection methods primarily rely on network scanning and rule-based detection techniques [3]. Although these methods are effective under certain conditions, they have several limitations, including poor adaptability to complex network environments, high false-positive rates, and insufficient detection capabilities for new types of threats.

Weakly Supervised Learning (WSL) is an emerging machine learning technique designed to train models using incomplete or imprecise label information. This approach can effectively learn and make predictions in the absence of large labeled datasets by leveraging partial label data, unlabeled data, or noisy labels. The main advantage of WSL is its ability to handle the issue of scarce labeled

data, enhance the generalization ability of models, and reduce the manual labeling workload, which is particularly important in the large-scale data environment of network asset detection.

This paper aims to explore how to apply weakly supervised learning to network asset detection to address the limitations of traditional methods. The research objectives include analyzing the potential application of weakly supervised learning in network asset detection and designing and evaluating related algorithms to improve the accuracy and efficiency of device identification. Through these discussions, this paper hopes to provide new ideas and methods to enhance the intelligence level of network asset detection.

2. RELATED WORK

Relatively broader in scope, some network assets refer to the hardware equipment involved Web, Domain name, operating system, application components, traffic, and so on. Only through asset detection can we comprehensively and deeply understand the target network assets and vulnerability status, prescribe targeted solutions, and continuously improve the quality of network security management [4].

The most primitive and inefficient asset detection method in traditional detection techniques is manual statistics, which refers to the recording of network assets through regular asset surveys organized with the assistance of some software; The new network asset detection methods rely on the rapid development of computers, resulting in three main detection methods: passive detection, active detection, and search query based detection.

Active detection methods refer to constructing data packets based on various protocols and purposes, and actively sending the constructed data packets to the target network. Based on the response results of the target network, the host survival is first detected, and then fingerprint extraction and library fingerprint comparison are performed on the data packet information returned by the surviving hosts to achieve network asset detection. There are currently two types of active detection methods. The first type is based on response protocol stack fingerprint detection, where Nmap is a typical representative tool. By sending the constructed data packet to the target host, it obtains information such as host, port opening, service type, operating system, device, etc., and obtains network asset detection results; However, this detection method can easily affect the normal operation of the target host, as it requires sending a large amount of network traffic to the target network. To address this issue, Shamsi et al. proposed a second type of active detection method based on the theory of single packet response delay statistics, which only sends a single SYN packet for detection and introduces a random model to analyze and identify the timeout (RTO) fingerprint of SYN/ACK retransmission packets [5]. However, active detection also has its weaknesses, as it can be easily blocked by network defense systems. At the same time, large-scale active detection can increase network load and trigger alerts from various security devices.

Passive detection method refers to the method of collecting the traffic of the target network through various network packet capture tools without actively sending data packets to the target network. Then, the fingerprint characteristics of special fields or protocol data packets in the traffic data of the target network are analyzed to obtain asset information such as IP address, MAC address, protocol, open port, device, etc., thereby achieving passive detection of network asset information. However, passive detection also has its weaknesses, such as low detection accuracy, which cannot detect assets that actually exist in network data but have not generated traffic during a certain period of packet capture.

Non invasive detection based on search queries can be divided into detection based on general search engines and detection based on network security specific search engines [7]. Google hacking technology has a certain ability to detect network assets, because this technology can use the Google search engine to detect vulnerability targets and mine sensitive information, obtain website structure

or search for network devices, and ultimately achieve network asset detection function. In the exploration based on the special search engine for network security, Shodan search engine is very powerful, which focuses on searching all equipment information connected to the Internet. Knowing that the ZoomEye search engine developed by Chuangyu Company can directly detect network assets, users can directly search for device fingerprints, web services, etc. in ZoomEye [8].

Weakly supervised learning has achieved significant results in multiple fields. In image classification, weakly supervised methods such as multi instance learning and pseudo label generation are used to improve the performance of models in labeled scarce environments. For example, by extracting features from image collections and performing label inference, the dependence on large-scale annotated data can be significantly reduced. In the field of text analysis, weakly supervised learning improves the effectiveness of tasks such as sentiment analysis and topic modeling by utilizing semi supervised learning and transfer learning techniques. Especially when dealing with large-scale text data, weakly supervised methods can effectively utilize a small number of labeled samples and rich unlabeled data to improve classification accuracy.

In summary, although traditional network asset detection methods have achieved certain results in recognition and classification, their limitations have driven the exploration of weakly supervised learning techniques. Weakly supervised learning has shown promising prospects in enhancing the intelligence and automation level of network asset detection. On this basis, this article proposes a network asset detection method based on weakly supervised learning. By using weakly supervised learning technology, it can better adapt to complex network environments and situations where some data in the network is unlabeled, thereby improving the accuracy and robustness of network asset recognition. By identifying accurate network assets, network assets can be more accurately and conveniently sorted out, providing a good foundation for subsequent network situational awareness and anomaly detection work.

3. NETWORK ASSET DETECTION METHOD BASED ON WEAKLY SUPERVISED LEARNING

3.1. Weakly Supervised Learning Methods

Weakly supervised learning is a research method in the field of machine learning, characterized by using relatively less and incompletely labeled training data for model training. Unlike traditional supervised learning methods, weakly supervised learning can effectively utilize unlabeled data to extract hidden information and train and optimize models [9].

In traditional supervised learning, a large amount of labeled data is usually required to train the model, which requires a significant amount of time and human resources. However, in real-world scenarios, many data do not have complete labeling information, and only some data is labeled. At this point, traditional supervised learning methods are ineffective, while weakly supervised learning can solve this problem. The core idea of weakly supervised learning is to train models by using partially or incompletely labeled data [10]. In weakly supervised learning, there are three common situations: the first type is that only a small subset of the training set is labeled, while other training data is unlabeled. In this case, model training is performed, which is incomplete supervision. The second type is imprecise supervision, where the training data labels are imprecise and only have coarse-grained labels. For example, given an instance segmentation task, but given training data with only image level annotations and no pixel level annotations. The third type is inaccurate supervision, where the labels provided by the model are not all correct, and some of them contain noise. In this case, it is necessary to consider how to better learn with noise [11].

In real-world network scenarios, only a portion of the collected data has labels, therefore it belongs to the first category of weakly supervised learning: incomplete supervision. Active learning and semi

supervised learning are two types of learning methods that are not fully supervised. Active learning is a continuous process of interaction and feedback that requires domain experts to label valuable and difficult to determine data in the model, and adjust the model based on these labeled data. However, semi supervised learning does not require domain experts and directly utilizes a small amount of labeled data and a large amount of unlabeled data for model training, fully utilizing limited labeled resources while exploring the potential of unlabeled data to improve model performance and accuracy.

The method used in this article is semi supervised learning. Currently, establishing the relationship between predicted examples and learning objectives in semi supervised learning must satisfy the following three basic assumptions [13]:

(1) Smooth assumption: If two samples meet the conditions of being in a data dense area and being very close to each other, then the labels of these two samples are similar. For example, when two samples meet the above two conditions, one with a label and the other without a label, it is highly likely that unlabeled data can use the label of labeled data as its own label.

(2) Clustering assumption: There is a high probability that two samples located in the same cluster have the same class label. The clustering hypothesis focuses more on the integrity of the sample. The clustering assumption aims to ensure that the data in dense areas belong to the same class as much as possible, and the clustering assumption prioritizes the same labels under the same cluster.

(3) Manifold assumption: The manifold assumption mainly considers the local characteristics of the model. Sometimes, two data points are far apart in high-dimensional space. However, when embedding high-dimensional data into a low dimensional manifold, the data points that are far apart in high-dimensional space may be closer through dimensionality reduction. Two samples are located in a small local neighborhood in the low dimensional manifold, and they are likely to belong to the same class.

3.2. Network Asset Detection Based on Weakly Supervised Learning

In network asset detection, weakly supervised learning significantly improves the efficiency of asset discovery and classification, anomaly detection, automated asset management, and network security situational awareness by reducing reliance on large amounts of labeled data [13]. Traditional methods have lower efficiency in dealing with massive dynamic assets. Weakly supervised learning utilizes a small number of labeled samples and a large amount of unlabeled data, using semi supervised learning techniques to enhance asset discovery capabilities, enabling automatic identification and classification of devices and services in the network, improving classification accuracy and recognition capabilities for new types of assets. Weakly supervised learning reduces the need for labeling abnormal data and enhances the ability to detect abnormal behavior in network traffic. Identify abnormal patterns by learning unlabeled normal traffic data. In complex network environments, weakly supervised learning can automatically update and maintain asset databases, identifying new or changed assets by analyzing network traffic and device behavior. Existing research shows that the application of weakly supervised learning in network asset detection is still in its early stages. Although studies have explored its potential in asset recognition and anomaly detection, practical application challenges in complex environments still need to be addressed [14].

3.3. Pseudo-Labeling

The algorithm used for semi supervised learning in the experiment is Pseudo Labeling. The pseudo labeling method enhances the training dataset by using the model's predicted results on unlabeled data as pseudo labels, thereby improving the learning performance of the model.

The specific pseudo labeling algorithm is shown in Algorithm 1.

Algorithm 1 The proposed method takes a set of labeled data and a set of unlabeled data and returns a trained model
Input: labeled dataset D_L, Unlabeled dataset D_U, Model M_0, Output: The trained model M_0
1: Initialize model, M_0 2: train a model, M_0, using the samples from D_L. 3: for i=1. Maxlterations do ▷ Repeats until convergence 4: Pseudo-label D_U. using M_0 5: $D_{selecte} \Leftarrow$ Select pseudo-labels using CFT (Confidence threshold) 6: $\widetilde{D} \Leftarrow D_{selecte}+D_L$ 7: Train M_0 using the samples from $\widetilde{D}$. 8: return M_0

Algorithm 1 shows how to generate pseudo tags. The process mainly consists of the following steps:

First, you need to prepare labeled data (D_L): display labeled samples and unlabeled data (D_U): display unlabeled samples.

The first step is to initialize the model M_0. We compared DT, GB, KNN, NB, RF, SVM machine learning models under supervised learning, and finally found that the DT model achieved good performance in terms of accuracy and F1 scores. So we decided to experiment on the semi supervised learning algorithm based on the DT model;

The second step is to train the model M_0 with labeled sample data D_L;

The third step is to generate pseudo label, use the model M_0 trained in the second step to predict the unlabeled data, then select the prediction data $D_{selecte}$ with high confidence as the data to add pseudo label, and combine the data $D_{selecte}$ with the original data D_L to form a new training data $\widetilde{D}$;

The fourth step is to use the new training data $\widetilde{D}$ to train the model M_0;

The fifth step is to iterate over the third and fourth steps until all unlabeled data generate false labels, end the iteration, and return to the final model. As this experiment is a multi classification problem, the generation of pseudo tags is easily affected by the uneven distribution of samples, resulting in inaccurate results. In order to solve this problem, the number of samples of various pseudo tags selected in the first 10 iteration cycles is required to be equal [15].

3.4. Decision Tree

In this study, we discussed the construction and application of the decision tree model, and demonstrated its effectiveness in the data task of this experiment. Decision tree has become a popular choice in many practical applications because of its intuitive structure and easy interpretation. In our experiment, through multiple segmentation and evaluation of the dataset, we verified the adaptability of the decision tree when dealing with different features and sample sizes.

We analyze the main parameters of the model, including the maximum depth, the minimum number of sample splits, the minimum number of leaf node samples, and the partition criteria. The reasonable setting of these parameters is crucial to the performance of the model, which can effectively prevent over fitting and improve the generalization ability. In the experiment, we used cross validation and grid search methods to optimize these parameters to ensure that the resulting model has good

performance on the test set. In the process of optimizing partition criteria parameters, we use Gini Impurity and Entropy;

Gini Impurity

$$Gini(D) = 1 - \sum_{i=1}^{C} {p_i}^2 \quad (1)$$

As shown in Formula (1), C is the total number of categories, and $\boldsymbol{p_i}$ is the proportion of category i in dataset D. Gini impurity measures the impurity in the data set. The smaller the value, the purer the dataset. When splitting, the decision tree will select the feature that can reduce the Gini impurity most for splitting.

Entropy

$$Entropy(D) = -\sum_{i=1}^{C} p_i log_2 (p_i) \quad (2)$$

As shown in Formula (2), $\boldsymbol{p_i}$ is the proportion of category i in dataset D.

Entropy is an important concept in information theory, which represents the uncertainty of the system. The higher the value, the greater the uncertainty of the system. The decision tree selects features by calculating the information gain after each feature is split. The information gain is defined as:

$$Gain(D, A) = Entropy(D) - \sum_{v \in Values(A)} \frac{|D_v|}{D} Entropy(D_v) \quad (3)$$

As shown in Formula (3), $Values(A)$ is the value of feature A, and D_v is the subset split by feature A.

We also analyzed the performance indicators of the decision tree and other machine learning algorithms, and found that the decision tree has good performance in terms of accuracy and response time, which is very suitable for rapid response in the application scenario of network asset detection. Table 1 below shows the performance indicators of different models.

Table 1. Performance indicators of different supervised learning models

Model	Accuracy	Precision	Recall	F1 score
DT	0.70	0.77	0.70	0.72
GB	0.69	0.78	0.69	0.72
kNN	0.70	0.75	0.70	0.71
NB	0.61	0.58	0.62	0.55
RF	0.70	0.76	0.70	0.72
SVM	0.60	0.69	0.63	0.64

From the above table, it is not difficult to find that the random forest (RF) and the decision tree (DT) have the same performance, but the decision tree has a faster response speed, so the next research in this paper is also based on the decision tree. Although decision tree is a basic machine learning algorithm, it is very suitable for the application scenarios proposed in this paper because of its excellent characteristics.

4. EXPERIMENTS AND RESULTS

4.1. Dataset Description

The main data used in this experiment is the network traffic data set. The network traffic data set is derived from the public network traffic data set library and contains the marked data of normal and

abnormal network traffic. Its main characteristics include traffic, source IP, destination IP, protocol type, etc. The characteristics of the data set include diversity and dynamics. The network traffic data covers normal network data and abnormal network data with attacks, while asset detection involves different types of network devices and services. Therefore, this experiment selects normal network data from the data set. The research purpose of the experiment is to identify network devices by analyzing the normal traffic, including the characteristics of single traffic and the research of combined traffic based on traffic behavior, combined with the weakly supervised learning algorithm mentioned in this paper, and then simulate and analyze the detection of network assets in the real network world.

4.2. Experimental Setup

The experimental setup includes the following aspects:

(1) Data preprocessing: feature extraction and normalization of network traffic data, processing and label generation of network devices. Because the analysis of combined traffic based on traffic behavior is greatly affected by the MAC address of network devices, there are two most typical cases, one is that different devices of the same type have different MAC addresses, the other is that some devices do not have MAC addresses, and their communication in the network environment depends on the MAC address of other devices. Finally, the network traffic data is feature extracted. In order to ensure the generalization of the experimental results, the MAC feature is removed, and finally 30 traffic features are left, which together constitute the device fingerprint.

(2) Model selection: semi supervised learning (pseudo label) and decision tree in weakly supervised learning are selected as the main models.

(3) Training and verification: Some data in the training set have labels, and all data in the verification set have labels, which are used to verify the performance of the model.

(4) Parameter setting: The decision tree model sets the feature selection standard as Gini coefficient, the maximum depth of the decision tree is 26, the maximum number of features considered when splitting each node is 26, and the minimum number of samples required for node subdivision is 6.

(5) Evaluation indicators: because the data set used in this experiment involves multiple classifications, and faces unbalanced data, because the evaluation indicator information cannot only observe the accuracy rate, but also pay attention to F1 scores, and consider comprehensively [16].

4.3. Result Analysis

The experimental results show that the weakly supervised learning model outperforms the traditional methods in the branch task equipment identification of network asset detection. Table 2 below shows the performance indicators of weakly supervised learning and supervised learning:

Table 2. Comparison of performance indicators between weakly supervised learning and supervised learning

Date	10%		20%		30%		all	
Model	acc	F1	acc	F1	acc	F1	acc	F1
DT	0.52	0.40	0.63	0.57	0.69	0.70	0.70	0.72
Ours	0.62	0.56	0.70	0.71	0.72	0.74		

The table above shows the comparison of indicators between weakly supervised learning and traditional supervised learning methods in the identification task of network asset detection equipment. It can be seen that the weakly supervised learning model has the same performance as traditional machine learning. Even if only 30% of the original data is labeled, a more accurate model

than supervised learning can still be obtained. After careful analysis, the reason why the preparation rate is not high is that the equipment has some similarities: they are either similar equipment produced by the same company for similar purposes or different models of the same equipment. Therefore, it seems impossible to separate them perfectly, at least when observing at the network level. However, these devices may use very similar hardware and software, show similar behavior, and have similar vulnerabilities and prevention methods. Therefore, from the perspective of equipment identification, it is reasonable to regard these equipment as labels of the same kind. When this is done, the accuracy rate increases from 72% to 88%.

4.4. Discussion

The experimental results show that weakly supervised learning has advantages in network asset detection. Semi supervised learning can effectively learn equipment features under limited tag data, which improves the accuracy of equipment recognition. The decision tree model can identify quickly and has high accuracy. On the whole, the weakly supervised learning technology not only improves the efficiency of asset detection in practical applications, but also reduces the dependence on a large number of tag data, adapting to the dynamic changes of the network environment. Future research can further explore the combination of weakly supervised learning and other advanced technologies to achieve more efficient network asset detection and management.

5. SUMMARY AND PROSPECT

5.1. Summary

This study discusses the application of weakly supervised learning in network asset detection, mainly using semi supervised learning pseudo label method. Through experiments, we find that this method has more advantages than traditional methods when dealing with complex network traffic and asset scanning data. It can identify devices with high accuracy when using few labeled training data. These results show that weakly supervised learning can not only significantly improve the efficiency of network asset detection, but also provide a more flexible solution for adapting to the dynamic network environment and network security management in the future.

5.2. Future Research Direction

Future research can be carried out in the following directions. First of all, more accurate and simplified device fingerprints can be extracted, which can quickly and accurately identify devices in today's increasingly complex equipment types, so as to facilitate real-time reflection of changes in the network environment, thus improving the response ability and decision-making efficiency of the security team, and timely warning of abnormal vulnerabilities in complex and changing network activities. Secondly, more complex weakly supervised learning algorithms and model structures can be explored, such as combining deep learning and migration learning technologies, to further improve the accuracy and robustness of network asset detection. Finally, we study how to optimize the model parameter setting and training process to achieve more efficient real-time monitoring and detection capabilities. In addition, the combination of real-time data flow and dynamic network analysis will help to better adapt to changes in the network environment and further enhance the network security protection capability.

REFERENCES

[1] Long Zhu. Research on Cyberspace Asset Detection and Analysis Technology [J]. Wireless Internet Technology, 2023, 20 (03): 149-151.

[2] Xianzhe Yang. Research on security early warning based on fingerprint detection method of network assets [D]. Zhengzhou University of Light Industry, 2022. DOI: 10.27469/d.cnki.gzzqc.2022.000015.

[3] Mingxing Li. Research and Implementation of Network Asset Detection Technology Based on Deep Learning [D]. Beijing University of Posts and Telecommunications, 2023. DOI: 10.26969/d.cnki.gbydu.2023.001486.

[4] Yongxia Wang, Yuxuan Jiang, Xiangzhe Yuan. Research and Analysis of Intranet Asset Detection Technology [J]. Network Security Technology and Application, 2024, (05):18-20.

[5] Chendong Wang, Yuanbo Guo, Shuaihui Zhen, etc. Research on Network Asset Detection Technology [J]. Computer Science, 2018, 45 (12): 24-31.

[6] Lei Shao, Xiao Yu, Jianzhang Wu. Research on Key Technologies of Network Asset Detection [J]. Network Security and Data Governance, 2022, 41 (11): 3-9+35. DOI: 10.19358/j.issn.2097-1788.2022.05.001.

[7] Yu Shen. Research and application of automatic identification method of network assets [D]. Shanxi University, 2021. DOI: 10.27284/d.cnki.gsxiu.2021.001759.

[8] Lijun Cheng, Zhiyong Zhang, Yuguang Zhang, etc. Research on Cyberspace Asset Detection and Analysis Technology [J]. Security Science and Technology, 2021, (03):13-19.

[9] Ziqiang Li, Wei Yang, Xianfeng Yang, etc. Semi automatic classification data annotation method based on weak label dispute [J]. Journal of Electronics, 2024, 52 (08): 2891-2899.

[10] Haowei Cheng, Wenjie Zi, Shuang Peng, etc. A 3D Mesh building facade extraction and semantic segmentation method based on semi supervised learning [J]. Journal of Zhengzhou University (Science Edition), 2023, 55 (04): 8-15. DOI: 10.13705/j.issn.1671-6841.2022161.

[11] Zhi-Hua Zhou, A brief introduction to weakly supervised learning, National Science Review, Volume 5, Issue 1, January 2018, Pages 44–53, https://doi.org/10.1093/nsr/nwx106

[12] Yu Zhong, Zhennan Huang, Huichao Xie, etc. A network abnormal traffic detection method based on semi supervised learning [J]. Journal of Guangxi University (Natural Science Edition), 2024, 49 (03): 563-574. DOI: 10.13624/j.cnki.issn.1001-7445.2024.0563.

[13] Tianhao Wang. Research on general ellipse target detection method based on deep learning [D]. Southeast University, 2023. DOI: 10.27014/d.cnki.gdnau.2023.000352.

[14] Xianzhe Yang, Yifeng Yin, Hongtao Zhang, etc. Research on Network Asset Detection and Early Warning of Education System [J]. Computer Applications and Software, 2023, 40 (10): 322-328.

[15] Lee D H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks [J]. 2013.

[16] Kostas K, Just M, Lones M A. IoTDevID: A Behavior-Based Device Identification Method for the IoT [J]. IEEE internet of things journal, 2022, 9(23):23741-23749.

[17] Wang X, Gao J, Wang J, et al. Self-Tuning for Data-Efficient Deep Learning [J]. 2021. DOI:10.48550/arXiv.2102.12903.

[18] Mingliang Yao, Ning Lu, Zhuanyan Bai, etc. Construction method of device fingerprint search engine for network asset vulnerability assessment [J]. Journal of Electronics, 2019, 47 (11): 2354-2358.