Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3792205, 14 pages
https://doi.org/10.1155/2022/3792205

Research Article

An Unknown Protocol Identification Method for Industrial Internet

Xinghui Zhu, Ziheng Jiang, Qiyuan Zhang, Shuangrui Zhao, and Zhiwei Zhang

School of Computer Science and Technology, Xidian University, Xi'an 710071, China

Correspondence should be addressed to Shuangrui Zhao; zhaoshuangrui@xidian.edu.cn

Received 24 June 2022; Accepted 18 August 2022; Published 5 September 2022

Academic Editor: Changyan Yi

This paper focuses on the problem of protocol identification in the industrial internet and proposes an unknown protocol identification method. We first establish an industrial internet protocol detection model to classify known protocols, unknown protocols, and interference signals and then store the unknown protocols for manual analysis. Based on the Eps-neighborhood idea, we further develop an Eps-neighborhood hit algorithm and propose an identification method to identify unknown protocols, where the supervised learning classification of unknown protocol detection is realized. Finally, extensive experimental results are provided to illustrate our theoretical findings. It indicates that the proposed method has an average screening accuracy of 94.675% and 95.159% for unknown protocols encoded in binary and ASCII, respectively, while the average screening accuracy of known protocols in binary and ASCII encoding is 94.242% and 94.075%.

1. Introduction

Industrial internet has become an indispensable component of intelligent manufacturing and has been widely used in many applications, such as product traceability, product life management, supply chain optimization, and health management [1–4]. Since the industrial internet has the characteristics like large scale, complex structure, and difficult management [5–7], it is urgent to establish a flexible and scalable platform to detect and identify industrial internet protocols and to realize the interconnection under such scenario [8]. In particular, the identification of industrial internet protocols can be divided into two categories: known protocol identification and unknown protocol identification [9]. The research and implementation of the former have been relatively mature, while the latter remains an open problem. How to solve the problem of identification of unknown protocols has been an important difficulty in the field of network security [10–12].

Compared with known protocols, unknown protocols have the characteristics of unknown format, unknown length, unknown characteristics, and unknown traffic, which make it more difficult to be detected and classified. In order to achieve the purpose of detecting unknown protocols, Liu et al. [13] proposed a port-based network traffic classification method with the advantages of fast recognition speed, high precision, and good performance. Zhang and Chen [14] used a small amount of labeled data to classify unknown protocols based on the semisupervised learning, which effectively improved the classification accuracy. By using a feature selection technique, Singh [15] proposed an unsupervised clustering method for unknown protocols classification, where a higher performance than K-means clustering accuracy was achieved. Ma and Qin [16] used the convolutional network to identify unknown protocols and treated the network flow load as image data, while Wang et al. [17] proposed a zero-knowledge classification model for unknown protocols in a bit stream. Jung and Jeong [18] considered a system where a deep belief network was combined and then proposed an extraction algorithm to realize the classification of unknown protocols based on average histogram features. Liu and Lang [19] proposed a traffic detection and identification method to detect traffic of unknown protocol, where the density-based spatial clustering of applications with noise (DBSCAN) clustering algorithm and convolutional neural network (CNN) algorithm was jointly utilized.

Table 1: Common binary encoding protocol part message table.

Protocol	Feature1	Feature2	Feature3	Feature4	Feature5	Feature6
MC3E	0x50	0x00	0x00	0xFF	0xFF	0x03
MC4E	0x54	0x00	0x01	0x00	0x00	0x00
MC4C	0x10	0x02	0x12	0x00	0xF8	0x00
COTP	0x03	0x00	0x00	0x19	0x17	0x0E
S7comm	0x03	0x00	0x00	0x19	0x17	0x0F
USS	0x02	0x0F	0x0F	Request data		0x43
Profibus-DP	0x68	0x05	0x05	0x68	0x83	0x81
MPI	0x68	0x1F	0x1F	0x68	0x83	0x81
PPI	0x68	0x1B	0x1B	0x68	0x02	0x00
Modbud RTU	0x01	0x03	0x0F	Data	0x32	0x43
Modbus TCP	0x00	0x0F	0x00	0x00	0x00	0x06
DF-1	0x10	0x01	0x01	0x10	0x02	0x01

It is worth noting that in the existing methods, the aforementioned literature mainly focuses on how to improve the performance and accuracy of the system when the number of unknown protocols is relatively small. If there exist a large number of unknown protocols and interference signals, these methods will not meet the requirements of the industrial internet. To address this problem, this paper establishes an industrial internet protocol detection model and develops an Eps-neighborhood hit algorithm. The main contributions of this paper are as follows:

(i) We establish an industrial internet protocol identification framework to classify 18 common known protocols, unknown protocols, and interference signals

(ii) We investigate for the first time the application of the DBSCAN clustering algorithm in the area of the industrial internet protocol identification. Based on the Eps-neighborhood idea, we propose an Eps-neighborhood hit algorithm to identify the unknown protocols

(iii) The experimental results verify our theoretical findings and demonstrate that the proposed Eps-neighborhood hit algorithm can effectively distinguish known and unknown protocols and improve the performance and accuracy of the system

The rest of the paper is organized as follows. Section 2 presents the industrial internet protocol detection model. Section 3 proposes the Eps-neighborhood hit algorithm. Experimental results are provided in Section 4, followed by the conclusion in Section 5.

2. System Model

We consider a system model which consists of a preprocessing module, a known protocol and unknown protocol screening module, a known protocol classification module, and an unknown protocol and interference signal screening module. We assume that the model's message data are all binary codes, ASCII codes, or interference signals. When the model receives a processing signal, the message data collected from the industrial programmable logic controller and distributed control system is divided into binary code and ASCII code, and then the features are extracted by principal component analysis in the message preprocessing stage.

By using the Eps-neighborhood hit algorithm at the filter, the filtered known protocol packets are submitted to the corresponding known protocol packet processing module for classification, while the filtered unknown protocol packets with interference signals are submitted to unknown industrial protocols and interference signal screening module. Note that the latter module further exploits the DBSCAN algorithm to discard the filtered interference signals. Moreover, when the number of identified unknown protocol packets reaches the threshold, it is added to the protocol database as a training data set for a single protocol, so that the unknown protocol packets can be added to the protocol database.

3. The Unknown Protocol Identification Method

In this section, we propose an unknown protocol identification method for industrial internet, where the principal component analysis (PCA) feature dimensionality reduction, Eps-neighborhood hit algorithm, and DBSCAN clustering algorithm are jointly employed.

3.1. Preprocessing Module

3.1.1. Feature Dimension Selection. According to [20], the DBSCAN algorithm needs to traverse the target point and perform Euclidean distance calculation with other points, such that the performance of the algorithm is very high in large-scale multidimensional data operations. In order to reduce the performance consumption of the DBSCAN algorithm, this paper uses principal component analysis to reduce the multidimensional features of the original data to two dimensions for classification. In the following, "Main

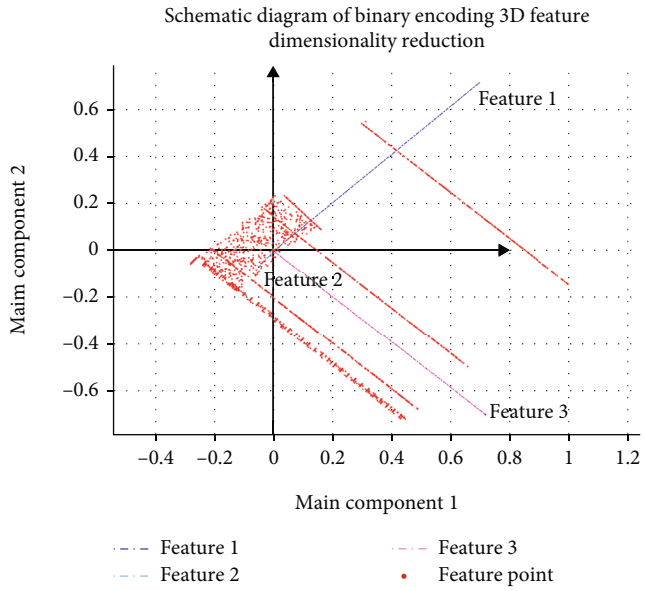

Figure 1: 3D schematic.

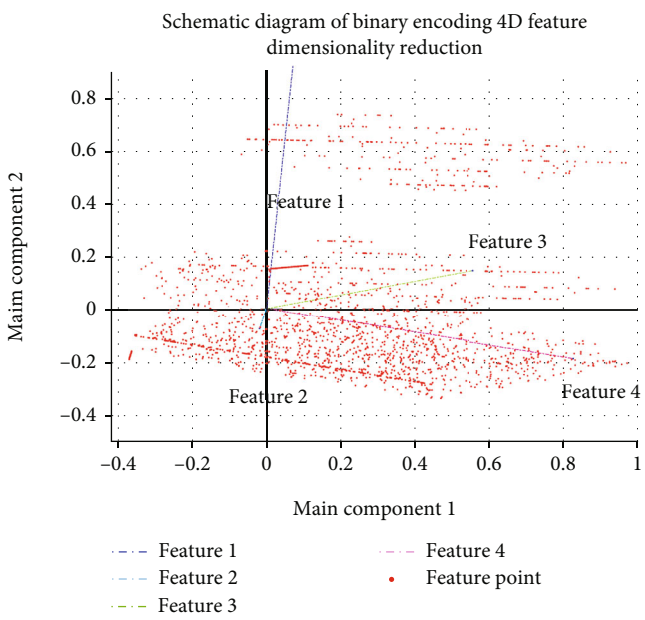

Figure 2: 4D schematic.

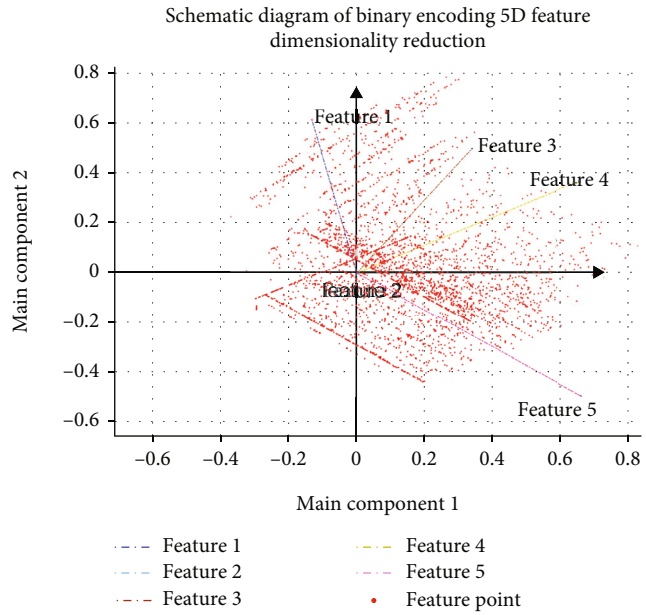

Figure 3: 5D schematic.

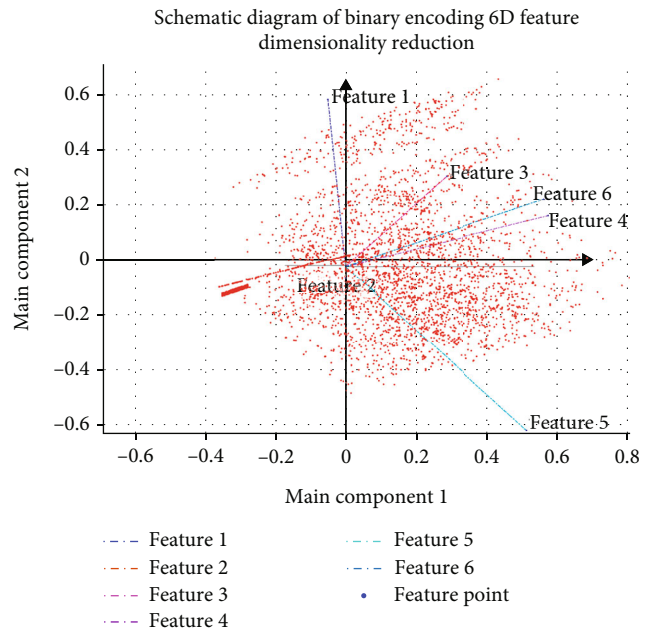

Figure 4: 6D schematic.

component 1" and "Main component 2" are the two main features of the data after dimensionality reduction.

(1) Binary Encoding Protocol Raw Dimension Selection. From [21], binary encoding protocols are transmitted in binary data streams. Table 1 includes 12 common binary encoding protocols and their corresponding protocol clusters, and the hexadecimal sample packets are divided into 6 main features. Based on Table 1, it can be seen that all binary encoding protocols use eight binary bits as a data link layer encoding unit and show obvious protocol characteristics in the converted hexadecimal value. For example, each protocol in the Profibus protocol family uses 0x68 as the first bit of the protocol message and the hexadecimal code of the message length as the protocol two or three bits, while each protocol in the S7 protocol family uses 0x03, 0x00, and 0x00 as the protocol, the first three digits of the protocol packet. Therefore, in this paper, these hexadecimal message bits with protocol features are used as the original feature dimension of PCA dimensionality reduction to realize the classification of different protocols.

Select the message data sets of several known binary encoding protocols in the table to perform principal component analysis and reduce the dimension of features 1 to 6 of

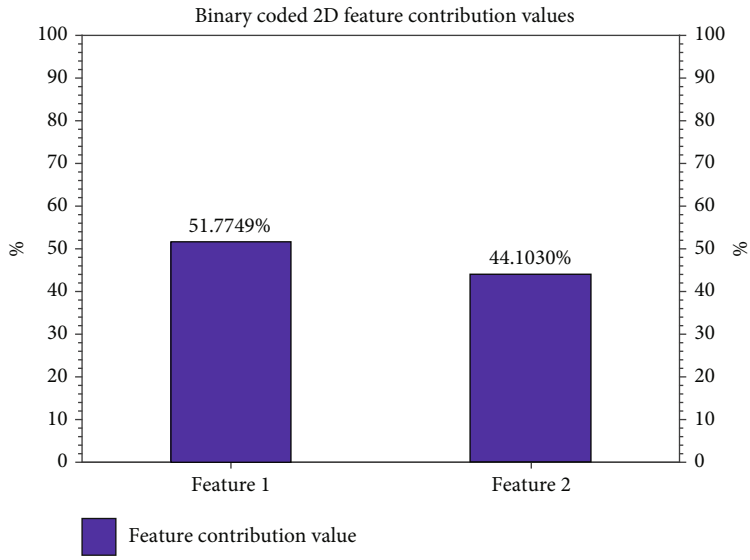

Figure 5: Contribution value of 2D feature after binary encoding dimensionality reduction.

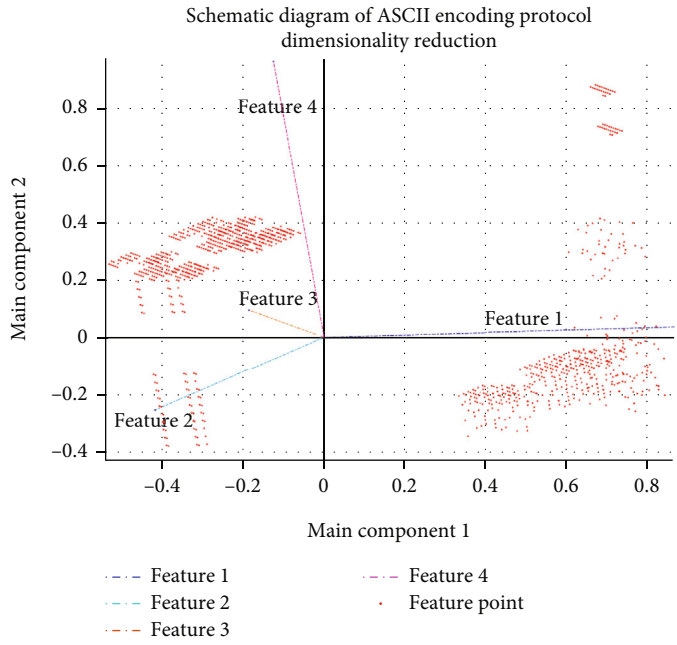

Figure 6: Schematic diagram of ASCII encoding protocol dimensionality reduction.

each protocol in the table to two dimensions. The protocol packet data test set for determining the original dimension of PCA includes 12,000 samples of 12 protocols. Figures 1–4 are the two-dimensional distribution diagrams of the principal component analysis feature dimensionality reduction of the three-dimensional, four-dimensional, five-dimensional, and six-dimensional original features, respectively.

It can be seen from the feature dimensionality reduction diagram that the six-dimensional original feature has good convergence for some protocol packet data, but no obvious principal components are extracted for the feature points of other protocols. The four-dimensional original features and the five-dimensional original features conform to the feature point density distribution of different protocols, but due to the transformation of the principal component coordinate system, the distribution of the two-dimensional principal components is not clearly distinguished. The three-dimensional original features have good convergence after dimensionality reduction by PCA, and the principal components are also relatively obvious.

Figure 5 shows the principal component contribution value of the two-dimensional feature obtained by the binary encoding of the original feature of the three-dimensional protocol after the PCA feature dimensionality reduction [22]. The original features contributed 44.1030% of the information, and when the dimensionality reduction of the three-dimensional original features was reduced to two-dimensional, the principal components 1 and 2 contributed

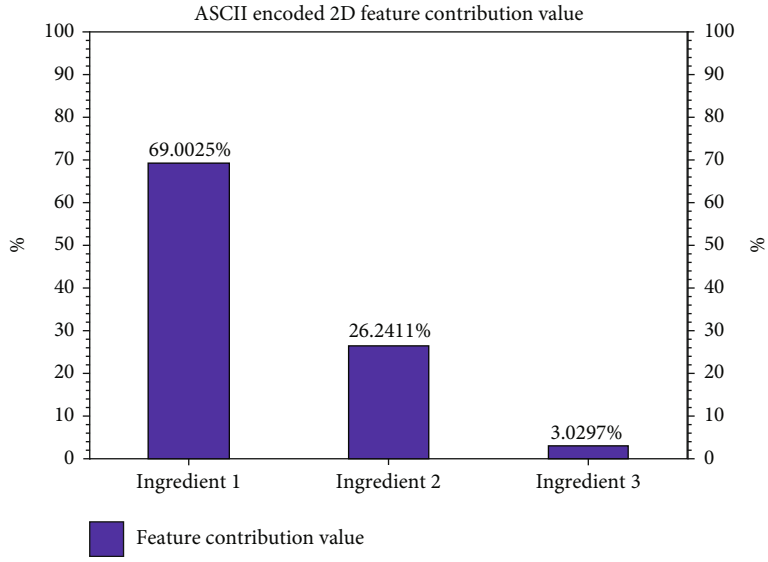

Figure 7: ASCII encoded 2D feature contribution value.

1 **input:** Known protocol feature set D, input message feature set P, neighborhood distance *eps*, minimum number of neighborhood points *minpts*;
2: **Output:** Known protocol classification dataset;
3: **If** $P_i \in V$ **then**
4: Check eps(P_i, D, P, eps);
5: **If** the number of E_i midpoints is greater than or equal to *minpts* **then**
6: Add point P_i to the set of visited points V;
7: Add point P_i to known protocol classification dataset S;
8: **Else**
9: The number of midpoints in E is equal to 0;
10: **End if**
11: **Else**
12: Check eps(T_g, D, P, eps)
13: **If**$T_g \in D$ and the number of E_g midpoints is greater than or equal to *minpts***then**
14: Add point P_i to the set of visited points V;
15: Add point P_i to known protocol classification dataset S;
16: **Break**;
17: **End if**
18: **End if**

Algorithm 1: Eps neighborhood hit algorithm.

95.8779% of the information of the original features. Therefore, this paper selects the binary-coded protocol dataset of the original features of the 3D protocol as the input of PCA.

3.1.2. ASCII Encoding Protocol Original Dimension Selection. Since the ASCII protocol encoding [23] format is transmitted through single characters 0-9, a-z, A-Z encoded as ASCII codes, corresponding to ASCII codes 48 to 57, 65 to 90, and 97 to 122, the data link layer is converted to hexadecimal. In the form of 0x30 to 0x39, 0x41 to 0x5A, and 0x61 to 0x7A, each characteristic bit of the protocol is expressed as a density distribution that follows the ASCII code range.

Figure 6 shows the relationship between the four-dimensional original features of the ASCII encoding protocol reduced to two-dimensional features. In this paper, sample data is added to the messages of the same protocol in different formats. The principal component analysis uses 1000 sample data of each ASCII encoding protocol. The data set includes a total of 10000 message sample data of 6 common ASCII encoding industrial protocol families. As can be seen from Figure 6, the four-dimensional original feature input accurately extracts the features of each ASCII-encoded industrial protocol message, and the feature points of the six different protocol family message samples are clearly divided.

Figure 7 shows the three principal components and their feature contribution values after dimensionality reduction. It can be seen from the figure that the contribution value of the principal component to the original feature is 69.0025%, the contribution value of the principal component to the

original feature is 26.2411%, and the principal component contributes 26.2411%. The contribution value of the three pairs of original features is 3.0297%, and the contribution value of the two-dimensional target dimension, namely, principal component one and principal component two, is 95.2436%. According to the contribution value of each feature in the figure, after denoising by principal component analysis, when the target dimension is two-dimensional, the information reflected by principal component one and principal component two must be greater than 95.2436% of the original sample. Here, the four-dimensional feature input can be used as the second dimension.

3.1.3. Preprocessing Process. The principal component analysis method is used to realize the message preprocessing process [24]. First, the encoding format of the input message data is judged according to the hexadecimal value range of the message data bits. Select the corresponding binary code or ASCII code known protocol database to obtain the corresponding known protocol training data set. The obtained known protocol training data set is submitted to the principal component analysis method together with the input message data, the protocol features of the data are extracted, and the feature dimension is reduced to two dimensions. The training data set follows the industrial protocol specification on the feature bits of each message and uses random values for other nonfeature bits and data bits, so as to eliminate the interference of specific data on the training results.

3.2. Screening Module. In this section, we propose an Eps-neighborhood hit algorithm to address the problems of high training cost and slow recognition speed of traditional supervised machine learning algorithms.

3.2.1. Eps-Neighborhood Hit Algorithm. The proposed Eps-neighborhood hit algorithm is based on the given neighborhood distance eps, and the minimum number of neighborhood points minpts is used to determine the neighborhood hit on the input packet feature set.

(i) If the point in the feature set of the input message is the core point under the current neighborhood distance, it is determined that the feature point is in the known protocol cluster, and the corresponding message is determined as the message of the known protocol

(ii) If the concentrated point has no neighbor point under the current neighbor distance, it is determined to be an unknown protocol or a packet feature point with an interference signal

(iii) If the point in the feature set of the input message is a boundary point under the current neighborhood distance, then traverse other feature points in the neighborhood of the point

(iv) If there are core points belonging to the known protocol feature set in these points, it indicates that the input point in the cluster of the core point, the corresponding packet is determined to be the packet of the known protocol

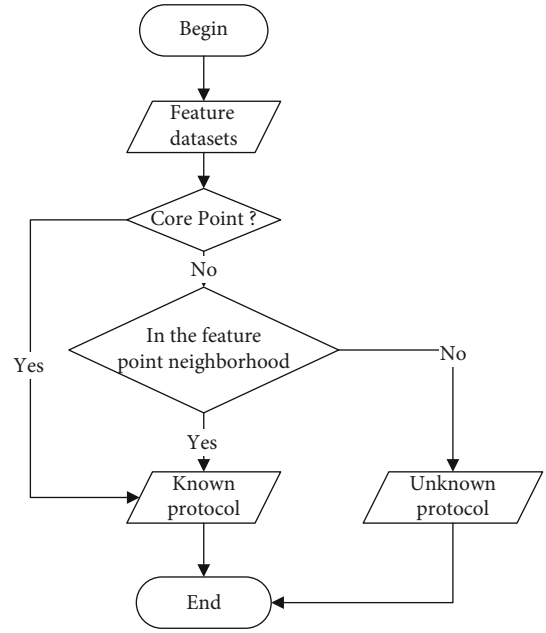

Figure 8: Flowchart of unknown packet screening method.

Formally, we summarize it in Algorithm 1. Note that this algorithm simplifies the operation process of the DBSCAN algorithm and distributes the feature points in the cluster of the known protocol training data set. The corresponding protocol packets are identified as known protocols, and the unknown protocol packets and the interference signal noise points are separated.

The key idea of the algorithm is derived from the DBSCAN algorithm [25]. We note that the conventional DBSCAN algorithm uses the two parameters of Eps neighborhood distance and the minimum number of domain points to calculate the core points, boundary points, and noise points in the data set. The advantage of the DBSCAN algorithm is to use the density and distribution of feature points for clustering instead of specifying the number of clusters of feature points, which has a better clustering effect for irregularly distributed feature points. In the clustering process, the feature points are divided into core points, boundary points, and noise points, which is convenient for boundary division of clusters and removal of noise points. This is in line with the characteristic distribution of unknown industrial communication protocols and is helpful to divide the unknown protocols by density clustering.

Here, in order to prove that the input message belongs to a known protocol, it is necessary to prove the characteristic points of the input message first. In the cluster of the known protocol feature set, that is, the core point of the known protocol feature set exists in the neighborhood of the feature point of the input message. If the input packet core point and the core point in the known protocol feature set to form a density reachable relationship, then the input packet feature point in the neighborhood of the core point also belongs

Table 2: Algorithm clustering fitting rate statistics table.

Algorithm	Ps 1	Ps 2	Ps 3	Ps 4	Ps 5	Ps 6
DBSCAN	65.45%	67.95%	99.91%	99.33%	72.22%	99.54%
K-means	66.34%	64.77%	55.86%	74.82%	78.61%	90.19%
Meanshift	64.26%	65.63%	99.43%	51.65%	93.84%	53.52%

Table 3: Binary code unknown protocol recognition accuracy table.

Algorithm	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Mean
DBSCAN	96.80%	94.96%	94.79%	96.03%	92.12%	96.55%	94.23%	95.79%	93.56%	91.92%	94.67%
K-means	90.84%	89.99%	85.49%	85.25%	84.65%	84.52%	85.77%	90.43%	85.75%	89.11%	87.18%
Meanshift	84.72%	83.92%	85.12%	79.31%	85.05%	80.20%	80.44%	85.15%	77.08%	82.62%	82.36%

Table 4: ASCII encoding unknown protocol recognition accuracy table.

Algorithm	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group 10	Mean
DBSCAN	97.02%	94.71%	97.36%	94.06%	95.32%	92.06%	95.89%	96.86%	92.46%	95.87%	95.16%
K-means	91.05%	89.13%	90.67%	84.04%	86.40%	90.80%	87.44%	84.94%	86.62%	87.47%	87.86%
Meanshift	93.57%	92.77%	91.83%	91.73%	87.45%	93.62%	90.37%	88.10%	93.08%	91.24%	91.38%

to the known protocol feature cluster and can be classified as a known protocol.

3.2.2. Screening Process. Next, we further propose a method for filtering known industrial protocol packets and unknown packets based on the Eps-neighborhood hit algorithm. Figure 8 shows the flow chart of the known industrial protocol packet and unknown packet filtering method. The screening method firstly inputs the two-dimensional feature data set and uses the Eps-neighborhood hit algorithm to detect whether the input message data is in the known protocol data set cluster in the two-dimensional feature distance. If the feature point of the input message is a core point in the current two-dimensional feature data set, it is identified as a known protocol message. While if the feature point of the input message is not a core point, it is judged whether there is a core point in the neighborhood of the feature point. When it does not exist, it is identified as an unknown protocol packet or a packet with an interference signal.

3.3. Identification of Unknown Protocols and Interference Signals

3.3.1. Clustering Algorithm Selection. Table 2 shows the cluster fitting rate and average cluster fitting rate of the DBSCAN algorithm, *K*-means algorithm, and meanshift algorithm for each protocol. The average cluster fitting rate of the DBSCAN algorithm is 84.07%, the *K*-means algorithm is 71.77%, and the meanshift algorithm is 71.39%. The DBSCAN algorithm has the best fitting effect on the known industrial communication protocols than the other two algorithms. Therefore, we use the DBSCAN algorithm to cluster the unknown protocols.

Case 1. There are different clusters in the protocol dataset. Suppose there are n types of known protocols, the number of protocol data feature points of known protocol i is c_i, the number of feature points belonging to protocol i in algorithm cluster a is m_{ia}, and the number of feature points belonging to protocol i in algorithm cluster b is m_{ib}. Then, the formula for calculating the cluster fitting rate P_i is expressed as

$$P_i = \frac{\max\left(m_{ia}, m_{ib}\right)}{c_i} \times 100\%. \tag{1}$$

Case 2. The feature points of the protocol dataset all belong to a certain cluster. Assuming that there are n types of known protocols, the number of protocol data feature points of known protocol i is c_i, and the number of feature points in algorithm cluster a is m_a, then, the calculation formula of cluster fitting rate P_i is as follows:

$$P_i = \frac{m_a - c_i}{c_i} \times 100\%. \tag{2}$$

As shown in formulas (1) and (2), the cluster fitting rate P_i of a protocol is equal to the proportion of feature points of algorithm clusters with the largest number in the data set of the protocol. The average protocol fitting rate P is expressed as

$$P = \frac{1}{n}\sum_{i=1}^{n} P_i. \tag{3}$$

3.3.2. Identification Process. According to the characteristics of the unknown packets and the characteristics of the interference signals mentioned above, this paper combines the principal component analysis method and the DBSCAN clustering algorithm to identify unknown industrial protocol packets and interference signals. The unknown protocol

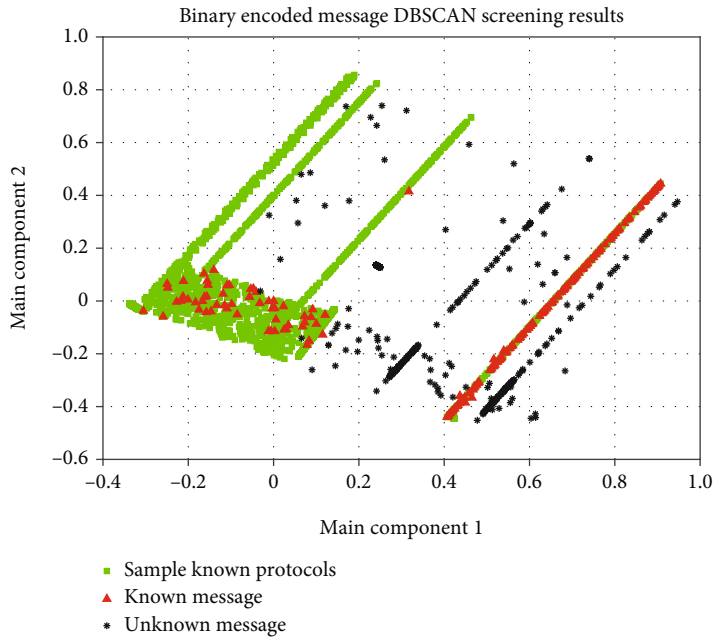

Figure 9: Binary encoded packet filtering results.

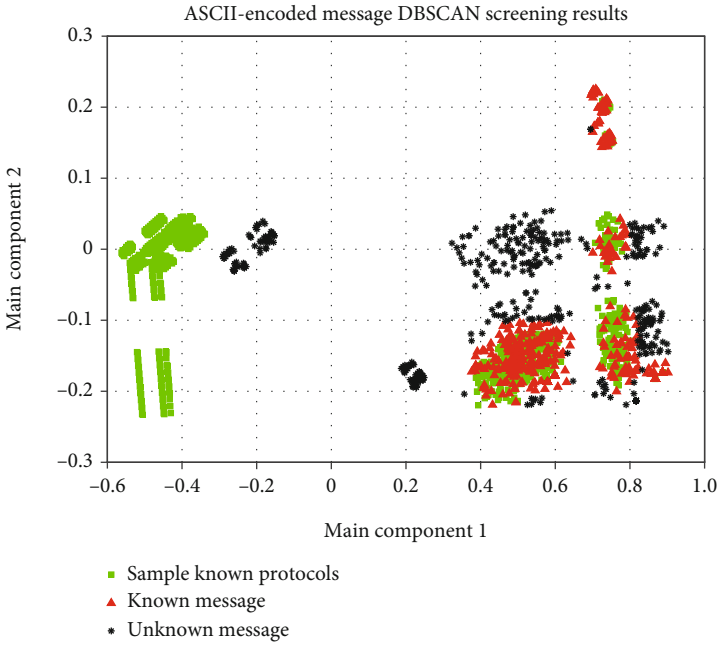

Figure 10: ASCII-encoded packet filtering results.

packets and the packets with interference signals screened out by the known industrial protocol packets and the unknown packet screening method are passed into the DBSCAN clustering algorithm. First, the two-dimensional characteristics of the unknown protocol packets and the interference signal packets are obtained. The data set is passed into the DBSCAN clustering algorithm, and then, DBSCAN clustering is performed on the two-dimensional feature data set. According to the characteristics of the unknown protocol clustering and distribution in a specific dimension and the principle of DBSCAN clustering, the feature clusters obtained by clustering correspond to these features. Therefore, the feature points of core points and boundary points are the feature points of unknown protocol packets, and the noise points can be regarded as the feature points of packets with interference signals (Tables 3 and 4).

4. Results and Discussions

4.1. Settings. The hardware verification in this paper uses a total of seven industrial programmable logic controllers (PLCs) from Siemens, Mitsubishi, Omron, and other brands

Table 5: Binary coding protocol screening result recognition rate statistics table.

Group	Number of known packets	Number of unknown packets	Misidentification	Known packet recognition rate	Unknown packet recognition rate
1	512	488	29	94.336%	94.057%
2	475	525	33	93.053%	93.714%
3	448	552	19	95.759%	96.558%
4	454	546	31	93.172%	94.322%
5	491	509	38	92.261%	92.534%
6	475	525	32	93.263%	93.905%
7	531	469	32	93.974%	93.177%
8	453	547	24	94.702%	95.612%
9	478	522	21	95.607%	95.977%
10	540	460	20	96.296%	95.652%
Mean	485.7	514.3	27.9	94.242%	94.551%

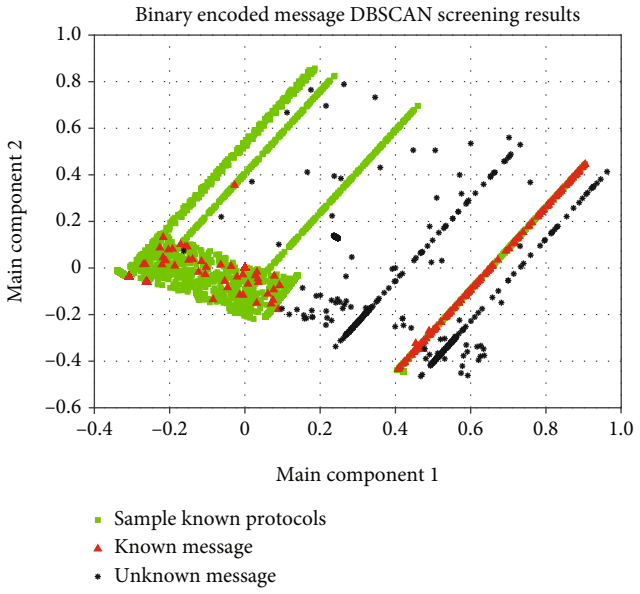

Figure 11: Binary encoded packet filtering results.

and one industrial Internet access gateway with independent intellectual property rights to simulate PLC communication in actual industrial scenarios. A laptop is used to connect the RS485 communication network through the USB to RS485 converter, and the RJ45 Ethernet interface is used to connect to the PLC's Enternet LAN, to realize the PLC's host computer communication and to simulate unknown protocol messages and interference signal messages. Combined with the mixing of unknown protocol packets, interference signal packets and known protocol packets in the actual communication link, we test the software and hardware access function and unknown protocol separation function of industrial equipments.

In this section, the metric of the average screening accuracy is used to measure the performance of the proposed method. Note that the average screening accuracy rate represents the average hit rate of the algorithm for different known protocols. As an algorithm that uses the DBSCAN clustering principle as the judgment criterion, the neighborhood hit rate can intuitively reflect the recognition effect of the known protocol packets under the current parameters.

4.2. Experimental Results. Figure 9 shows the operation results of the first test group among the ten test groups of the binary coding protocol. Here, 1000 pieces of sample data for each binary coding protocol and a total of 12,000 pieces of sample data for 12 protocols are used as the known protocol training samples. A total of 1000 mixed data sets are used as the test set, and the number of each message is random, the eps is 0.02, and the number of minpts is 3. Assume that the number of input known packets is m_0, the number of input unknown packets is n_0, the number of misrecognition of known packets is m_1, the number of misrecognition of unknown packets is n_1, the calculation formula of the known packet recognition rate P_m is as the following formula, and the formula of the unknown packet recognition rate P_n is the same.

$$P_m = \frac{n_1 + m_1}{m_0} \times 100\%. \tag{4}$$

The binary coding protocol test group includes 512 known packets collected and 488 unknown protocol/interference signal packets. The input screening method identifies 533 known packets and 467 unknown packets. There are 4 unknown packets that are misidentified known packets, and 25 unknown packets are misidentified as known packets. The known packet recognition rate is 94.336%, and the unknown packet recognition rate is 94.057%.

Figure 10 shows the running results of the first test group among the ten test groups of the ASCII protocol. Similarly, 1,000 pieces of sample data for each ASCII protocol and 10,000 pieces of sample data for 6 protocol families and subprotocols are used as the known protocol training samples. The known protocol packets, unknown protocol packets, and interference signals are collected. A total of 1000 mixed data sets of packets are used as the test set, the number of each packet is random, the eps is 0.02, and the number of minpts is 3.

Table 6: Binary coding protocol clustering results recognition rate statistics table.

Group	Number of total packets	Number of unknown packets	Misidentification	Location packet clusters	Unknown protocol recognition rate
1	467	437	14	4	96.796%
2	512	496	25	6	94.960%
3	543	518	27	7	94.788%
4	529	504	20	9	96.032%
5	485	457	36	11	92.123%
6	497	464	16	5	96.552%
7	457	433	25	7	94.226%
8	527	499	21	12	95.792%
9	507	481	31	9	93.555%
10	454	421	34	10	91.924%
Mean	497.8	471	24.9	8	94.675%

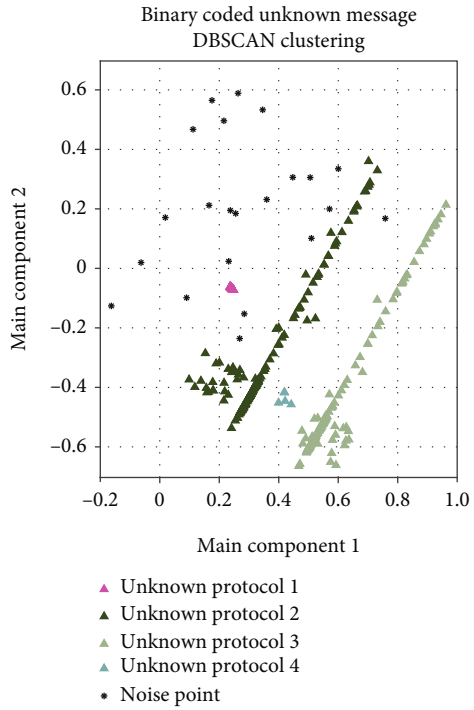

Figure 12: Clustering diagram of binary coded unknown packets.

The ASCII encoding protocol test group includes 519 known packets collected and 481 unknown protocol/interference signal packets. The input screening method identifies 547 known packets and 453 unknown packets. Therein, 1 unknown packet is misidentified known packets, and 29 unknown packets are misidentified as known packets. The known packet identification rate is 94.220%, and the unknown packet identification rate is 93.763%.

From the test results of binary-coded protocol packets and ASCII-coded protocol packets, it can be concluded that the feature recognition rate of known protocol packets of the proposed Eps-neighborhood hit algorithm is above 94%, which performs well to identify known protocol packets (Table 5).

4.2.1. Verification of Unknown Protocols and Interference Signal Identification Method. In this subsection, ten groups of unknown industrial protocol messages in binary and ASCII codes are screened out by the screening method, and the DBSCAN algorithm, K-means algorithm, and meanshift algorithm are used to perform clustering and identification fitting rate comparisons. Due to the different quantity and density of unknown protocol packets screened by the screening method, we use the protocol identification accuracy rate to measure the accuracy rate of unknown protocol and interference signal identification. The protocol identification accuracy rate of the algorithm is expressed as the ratio of the difference between the number of unknown protocol packets in a certain unknown protocol packet data set and the sum of the algorithm clustering misidentified packets and the number of unknown protocol packets in the protocol packet data set. Suppose the unknown protocol packet data set t has a total of at messages, n algorithm clusters in the unknown protocol data set, and q_i misidentified packets in each cluster, then the identification accuracy can be expressed as

$$R(t) = \frac{a_t - \sum_{i=1}^{n} q_i}{a_t} \times 100\%. \tag{5}$$

Table 3 is a statistical table of the accuracy rate of binary-coded unknown protocol recognition for the three clustering algorithms. The binary-coded unknown protocol/interference signal mixed message data set screened by the input screening method is 4978 in ten groups. The average recognition accuracy of the DBSACN algorithm is 94.67%, the average recognition accuracy of the K-means algorithm is 87.18%, and the average recognition accuracy of the meanshift algorithm is 82.36%.

Table 3 is the ASCII code unknown protocol recognition accuracy statistics table of the three clustering algorithms. The input screening method selects ASCII code unknown protocol/interference signal mixed message data sets in ten groups with a total of 4834 pieces. The average recognition accuracy of the DBSCAN algorithm is 95.16%, the average

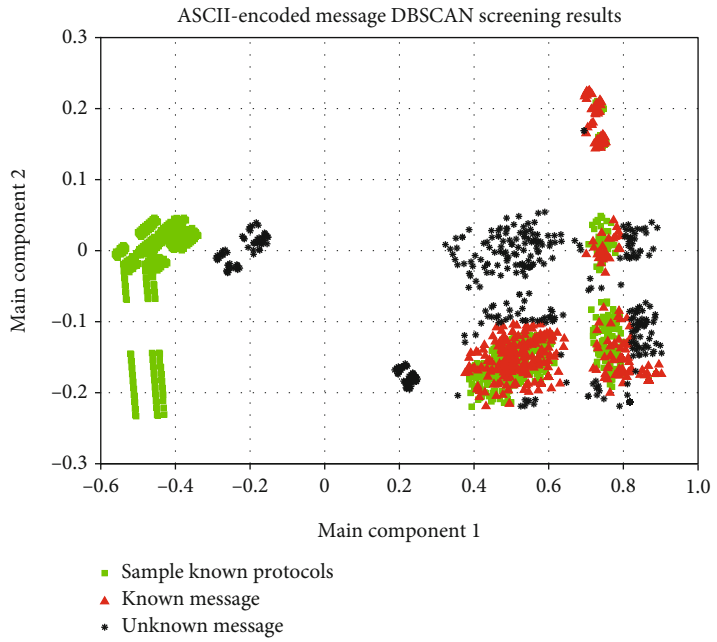

Figure 13: ASCII-encoded packet filtering results.

recognition accuracy of the K-means algorithm is 87.86%, and the average recognition accuracy of the meanshift algorithm is 91.38%.

Note that the DBSCAN algorithm is used as a clustering algorithm for identifying unknown protocols and interference signals, which meets the needs of identifying unknown protocols. The accuracy rate is higher than that of K-means algorithm and meanshift algorithm.

4.3. Performance Analysis in Real Industrial Environment

4.3.1. Binary Encoding Mixed Packet Test. Figure 11 shows the experimental verification of binary-coded unknown/known industrial protocol message screening. According to the distribution of known packets and unknown packets, some random interference signal packets or unknown protocol packets with characteristic points near known protocol samples may be identified as known packets. Therefore, known packets are tested in the test. The number of misrecognition is generally less than the recognition rate of unknown packets.

The sixth test group of the binary coding protocol randomly generated 475 known packets and 525 unknown protocol packets plus interference signal packets. The input screening method identified 503 known packets and 497 unknown packets. Therein, 2 unknown packets are misidentified as unknown packets, and 30 unknown packets are misidentified as known packets. The recognition rate of known packets is 93.263%, and the recognition rate of unknown packets is 93.905%.

Table 5 shows the screening results of ten groups of binary-coded known packets. The accuracy evaluation criteria of the screening algorithm of the paper is the recognition rate of known packets, the recognition rate of unknown packets, and the number of misidentified packets. The average number of misidentified packets in the ten groups of test data is 27.9 per thousand, the average recognition rate of known packets is 94.242%, and the average recognition rate of unknown packets is 94.551%.

Table 6 is a statistical table of the recognition rate of the clustering results of the ten groups of binary coding protocol test groups. The algorithm accuracy evaluation criteria of the unknown industrial protocol packet/interference signal identification method are the unknown protocol packet recognition rate and the number of misidentified packets. The binary-coded unknown packets are screened out by each group of input packets in the test group, the average number of misidentified packets in the ten groups of test data is 24.9 per 497.8, and the average recognition rate of unknown protocol packets is 94.675%.

Figure 12 shows the clustering of unknown packets in the first test group. The first test group inputs 467 unknown packets, including 437 unknown protocol packets of 4 types and 30 interference signal packets. After clustering operation, 4 unknown protocol clusters are obtained, and 22 interference signal packets are identified. Therein, 11 interference signal packets are mistakenly identified as unknown protocol packets, 3 unknown protocol packets are mistakenly identified as interference signal packets, and the unknown protocol recognition rate is 96.796%.

4.3.2. ASCII Encoding Mixed Packet Test. Figure 13 shows the results of the first test group of the ASCII encoding protocol, in which 519 known packets and 481 unknown protocol packets plus interference signal packets are randomly generated. It is not difficult to see from the figure that the eps algorithm has a good clustering effect on ASCII packets and can effectively distinguish known protocols from unknown protocols, giving full play to the advantages of supervised learning and clustering algorithms. The

Table 7: ASCII encoding protocol screening result recognition rate statistics table.

Group	Number of known packets	Number of unknown packets	Misidentification	Known packet recognition rate	Unknown packet recognition rate
1	519	481	30	94.220%	93.763%
2	450	550	33	92.667%	94.000%
3	509	491	18	96.464%	96.334%
4	476	524	44	90.756%	91.603%
5	527	473	19	96.395%	95.983%
6	498	502	27	94.578%	94.622%
7	462	538	25	94.589%	95.353%
8	538	462	36	93.309%	92.208%
9	544	456	27	95.037%	94.079%
10	468	532	34	92.735%	93.609%
Mean	499.1	500.9	29.3	94.075%	94.155%

Table 8: Statistical table of recognition rate of ASCII encoding protocol clustering results.

Group	Number of total packets	Number of unknown packets	Misidentification	Location packet clusters	Unknown protocol recognition rate
1	453	436	13	7	97.018%
2	523	491	26	6	94.705%
3	483	454	12	11	97.357%
4	498	471	28	9	94.055%
5	468	449	21	12	95.323%
6	493	466	37	14	92.060%
7	531	511	21	18	95.890%
8	440	414	13	13	96.860%
9	431	411	31	16	92.457%
10	514	484	20	12	95.868%
Mean	483.4	458.7	22.2	11.8	95.159%

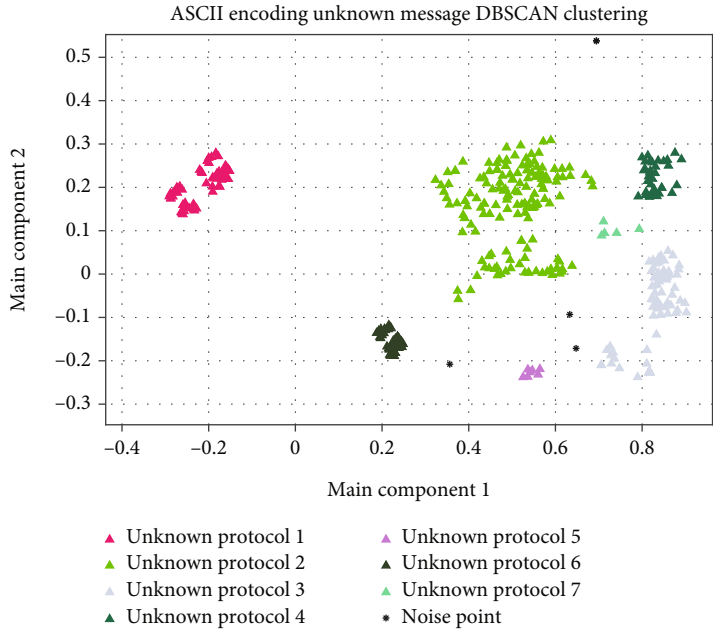

Figure 14: ASCII-encoded location message clustering diagram.

experimental results show that the model identifies 547 known packets and 453 unknown packets, while 1 known packet is misidentified as an unknown packets and 29 unknown packets are misidentified as known packets. The text recognition rate is 94.220%, and the unknown message recognition rate is 93.763%.

Table 7 shows the statistical table of the running results of ten groups of ASCII encoding protocol test groups. The average number of falsely identified packets in the test data is 29.3 per thousand, the average known packet identification rate is 94.075%, and the average unknown packet identification rate is 94.075%. Table 8 shows the statistical table of clustering results for groups 1 to 10 in the test group of ASCII-encoded unknown packets. We can see that the unknown protocol recognition rate is 95.159%.

Figure 14 shows the clustering of ASCII-encoded unknown packets in the first test group. The first test group entered 453 unknown packets, including 437 unknown protocol packets of 5 types and 30 interference signal packets. After clustering with the same algorithm, 7 unknown protocol clusters with a total of 449 packets are obtained. 4 interference signal packets are identified, while 13 interference signal packets are misidentified as unknown protocol packets. The unknown protocol recognition rate is 97.018%.

5. Conclusion

This paper proposed an Eps-neighborhood hit algorithm to separate known industrial protocol packets from unknown packets based on the classical DBSCAN algorithm. The application of the DBSCAN clustering algorithm in the area of the industrial internet protocol detection was also investigated. With the help of the proposed algorithm, we designed an industrial internet adaptive access system, where adaptive protocols for industrial hardware equipment access are identified and classified effectively. It indicates that the proposed method has an average screening accuracy of 94.675% and 95.159% for unknown protocols encoded in binary and ASCII, respectively, while the average screening accuracy of known protocols in binary and ASCII encoding is 94.242% and 94.075%, which has the potential to be implemented in actual industrial scenarios.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (Grant no. 2018YFE0207600), in part by the Natural Science Foundation of China (NSFC) under Grant 61972308 and in part by Natural Science Basic Research Program of Shaanxi (Program no. 2019JC-17).

References

[1] A. Amjad, F. Azam, M. W. Anwar, and W. H. Butt, "A systematic review on the data interoperability of application layer protocols in industrial IoT," *Access*, vol. 9, pp. 96528–96545, 2021.

[2] M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial internet of things and industry 4.0," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 10, pp. 4674–4682, 2018.

[3] L. Liu, M. Zhao, M. Yu, M. A. Jan, D. Lan, and A. Taherkordi, "Mobility-aware multi-hop task offloading for autonomous driving in vehicular edge computing and networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2022.

[4] J. Feng, L. Liu, Q. Pei, and K. Li, "Minmax cost optimization for efficient hierarchical federated learning in wireless edge networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 11, pp. 2687–2700, 2021.

[5] J. Q. Li, F. R. Yu, G. Deng, C. Luo, Z. Ming, and Q. Yan, "Industrial internet: a survey on the enabling technologies, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1504–1526, 2017.

[6] E. Sisinni, A. Saifullah, S. Han, U. Jennehag, and M. Gidlund, "Industrial internet of things: challenges, opportunities, and directions," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 11, pp. 4724–4734, 2018.

[7] I. Bedhief, L. Foschini, P. Bellavista, M. Kassar, and T. Aguili, "Toward selfadaptive software defined fog networking architecture for IIoT and industry 4.0," *Proc. 2019 IEEE 24th Int. Workshop Comput. Aided Model. Design Commun. Links Netw.(CAMAD)*, pp. 1–5, 2019.

[8] J. Yue, M. Xiao, and Z. Pang, "Distributed BATS-based schemes for uplink of industrial internet of things," *Proc. ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2019.

[9] J. Lin and L. Liu, "Research on security detection and data analysis for industrial internet," *Proc. 2019 IEEE 19th Int. Conf. Softw. Qual. Rel. Secur. Companion (QRS-C)*, pp. 466–470, 2019.

[10] T. Cerquitelli, D. J. Pagliari, A. Calimera et al., "Manufacturing as a data-driven practice: methodologies, technologies, and tools," *Proceedings of the IEEE*, vol. 109, no. 4, pp. 399–422, 2021.

[11] X. Cui, Y. Li, Y. Liu et al., "Analysis methodology for differential mode interference in energy supply system of hybrid DC breaker," *IEEE Transactions on Electromagnetic Compatibility*, vol. 61, no. 6, pp. 1967–1978, 2019.

[12] T. J. Levy, U. Ahmed, T. Tsaava et al., "An impedance matching algorithm for common-mode interference removal in vagus nerve recordings," *Journal of Neuroscience Methods*, vol. 330, article 108467, 2020.

[13] Y. Liu, W. Li, and Y. Li, "Network traffic classification using k-means clustering," in *Proc. 2nd Int. Multi-Symp. Comput. Comput. Sci*, pp. 360–365, IMSCCS, 2007.

[14] J. Zhang and C. Chen, "An effective network traffic classification method with unknown flow detection," *Manage*, vol. 10, no. 2, pp. 133–147, 2013.

[15] H. Singh, "Performance analysis of unsupervised machine learning techniques for network traffic classification," *Proc. 2015 5th Int. Conf. Adv. Comput. Commun. Technol.*, pp. 401–404, 2015.

[16] R. Ma and S. Qin, "Identification of unknown protocol traffic based on deep learning," *Proc. 2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pp. 1195–1198, 2017.

[17] W. Wang, B. Bai, Y. Wang, X. Hei, and L. Zhang, "Bitstream protocol classification mechanism based on feature extraction," *Proc. 2019 International Conference on Networking and Network Applications (NaNA)*, pp. 241–246, 2019.

[18] Y. G. Jung and C.-M. Jeong, "Deep neural network-based automatic unknown protocol classification system using histogram feature," *The Journal of Supercomputing*, vol. 76, no. 7, pp. 5425–5441, 2020.

[19] H. Liu and B. Lang, "Network traffic classification method supporting unknown protocol detection," *Proc. 2021 IEEE 46th Conference on Local Computer Networks (LCN)*, pp. 311–314, 2021.

[20] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino, "Euclidean distance geometry and applications," *SIAM Review*, vol. 56, no. 1, pp. 3–69, 2014.

[21] R. Hu, S. Huang, M. Wang, L. Zhou, X. Peng, and X. Luo, "Binary thermal encoding by energy shielding and harvesting units," *Physical Review Applied*, vol. 10, no. 5, article 054032, 2018.

[22] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *Access*, vol. 8, pp. 54776–54788, 2020.

[23] S. K. Mukhopadhyay, M. Omair Ahmad, and M. N. S. Swamy, "ASCII-character-encoding based PPG compression for telemonitoring system," *Biomedical Signal Processing and Control*, vol. 31, pp. 470–482, 2017.

[24] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[25] M. Kryszkiewicz and Ł. Skonieczny, "Faster clustering with dbscan," in *Intelligent Information Processing and Web Mining*, pp. 605–614, Springer, 2005.