Traffic analysis for 5G network slice based on machine learning

Feng Xie[1*], Dongxue Wei[1] and Zhencheng Wang[2]

*Correspondence:
xiefeng6640@163.com
[1] School of Artificial
Intelligence, Nanning College
for Vocational Technology,
Nanning 530008, Guangxi,
China
Full list of author information
is available at the end of the
article

Abstract

With the rise of 5G and Internet of things, especially the key technology of 5G, network slice cuts a physical network into multiple virtual end-to-end networks, each of them can obtain logically independent network resources to support richer services. 5G mobile data and sensor data converge to form a growing network traffic. Traffic explosion evolved into a mixed network type, and network viruses, worms, network theft and malicious attacks are also involved. How to distinguish traffic types, block malicious traffic and make effective use of sensor data under the background of 5G network slice, and also the significance of this study.

Keywords: 5G, Network slice, Machine learning, Traffic analysis

With the advent of the era of 5G and artificial intelligence, machine learning plays an important role in image recognition, language processing, speech recognition and other fields. In this paper, the Internet of things of home is selected as the research scene, through the construction of a sensor network, ZigBee device is connected to the Internet, to obtain temperature, humidity, power of home appliances, smoke concentration and other sensor data in the home, and then, it used the machine learning to study and analyze the flow data, identify the validity of flow, ensure the security of home network, summarize and analyze sensor data, and ensure the security of home physical environment. This study not only has important theoretical significance in machine learning, but also has a broad application prospects in smart home industry.

1 Introduce

At present, there are few studies on the combination of machine learning, traffic analysis, network slice and the Internet of things, and the function of adding sensor data to the Internet traffic participation classification is not implemented [1–4]. To solve this problem, under the background of 5G network slice, this paper proposes a home traffic analysis system combined with Internet of things [5, 6].

Firstly, this paper introduces the network slice of key role in 5G technology. Around the key concepts of machine learning and traffic analysis, it includes origin, development, features and forms [7–10]. With the wide application of artificial intelligence in various industries, people can use the high-speed processing ability of computers

Xie *et al. J Wireless Com Network*   (2021) 2021:108

Page 2 of 15

to train a large amount of data, get a model to achieve traffic classification and finally apply it in practical situations [11]. This paper introduces the design and implementation of the home traffic analysis system combined with the Internet of things in detail and demonstrates its characteristics in the process of explaining it. The first is to create the Internet of things inside the smart home, using the sensor network architecture of Zig-Bee, collect data through the ZigBee node for each sensor device and finally collect data through the coordinator to the gateway, which forward the data to the server, thus realizing the traffic classification [12–15]. In order to design and implement the system, this paper establishes a related experimental study. In the experiment, the flow dataset used for training is obtained, its data characteristics are analyzed by statistical analysis, it is filtered and cleaned, and the related algorithms of machine learning, including decision tree, random forest and regression, are used to train the flow data samples, and then, the test is carried out. Samples assess model performance. In order to verify the actual effect of the model, this paper uses the package software to capture the actual traffic data and send it to the model for evaluation and judgment, to measure the effect of the model according to the accuracy of the judgment.

In order to compare the performance differences of different machine learning algorithms in the process of feature selection, this paper describes the feature selection of two different algorithms and measures them through experiments. The experimental data show that the Chi-square filter method can get better accuracy and better comprehensive performance.

2 Related concepts and work

2.1 Network slice

The network slice can be adapted according to the requirements of each user. Different traffic and resources are treated differently by slicing, and then, network operators make different service requirements for customers of different tenant types [16–18]. Network slice is one of the key technologies of 5G, NFV is an essential technology of network slice, NFV isolated from traditional network hardware and software part and hardware by unified server deployment, the software shall be borne by the different network function (NF), so as to realize the demand of flexible assembly operations, Network slice allow to share the same physical network communication link, and complete the data exchange by virtual independent sub network, so as to better meet the needs of 5G everything connected.

2.2 Machine learning

Machine learning is an important realization method in the field of artificial intelligence. Its development can be summarized in the following stages:

In the early 1950s, machine learning began to sprout. The concept of perceptron was proposed, but this perceptor can only handle linear classification problems, not XOR logic. In the 1960s and 1970s, symbolism learning technology based on logical representation flourished, mainly researching inductive learning system based on logic [19, 20]. From the 1980s to the mid-1990s, decision, the emergence of tree makes it easy for classification algorithms to express complex data relationships in terms of knowledge

learning, but it will result in the learning process facing assumptions that are too large and complex [21, 22]. The last stage is from the mid-1990s to today, when facing a large amount of data content, statistical learning and deep learning came into being.

At present, the research work in the field of machine learning mainly focuses on three aspects, one is task-oriented research, which focuses on performance analysis and improved learning system for a set of predefined tasks; the other is to build a cognitive model, which focuses mainly on human learning process and simulates it with high-performance computer; the last is theoretical analysis, which promotes learning through theoretical exploration. Algorithmic study of the effect. In the past decade, machine learning has achieved very effective results in the application of industry, mainly focusing on weather forecasting, image recognition, voice and handwriting recognition, stock market analysis, pattern recognition and other fields. At present, major companies have launched their own machine learning platforms, such as Tensor Flow machine learning framework of Google, Microsoft Azure machine learning studio, MLflow of open source machine learning platform, Baidu machine learning (BML), Ali PAI, JD NeuCube, etc. Machine learning has become the main research direction of major companies.

The machine learning algorithm can effectively predict the gathered sensor data collected by the wireless sensor network, and then monitor the abnormalities of the home appliances in the sensor location, the future trend of the home environment and so on, to realize the intelligent service of smart home energy saving reminder [23, 24]. The combination of smart home and machine learning is of certain research value. In the foreseeable future, there will be a large number of smart home systems that apply machine learning in the market, providing people with a more convenient and intelligent home living space [25].

2.3 Traffic mining analysis

Traffic mining analysis technology is a technology that captures network traffic and continuously improves the identification algorithm and extracts traffic characteristics according to the changes of network environment [26, 27]. Up to now, the main methods of traffic identification analysis are based on port number mapping, based on network behavior characteristics, and based on machine learning.

2.3.1 Traffic identification based on port number mapping

The specific implementation of traffic identification method based on port number mapping is to identify different network applications by checking the source and destination port numbers of network packets, and mapping the port number rules used in communication according to the corresponding network protocol or network application. Kim et al. pointed out that port classification technology is effective in identifying some applications, such as WWW, DNS and MAIL, with both accuracy and recall rates higher than 90% [28].However, current P2P applications use port hopping technology and port masquerading technology to avoid traffic detection. Bleul et al. analyzed the direct-connect network and found that 70% of the observed ports were used only once [29]. Schneider et al. found that when port classification technology identifies UDP traffic, the byte accuracy is only 24%.It can be seen that the port-based traffic identification technology can no longer meet the current needs, and the limitations of this method

Xie *et al. J Wireless Com Network*   (2021) 2021:108

Page 4 of 15

are becoming more and more obvious. First, the system does not define communication port numbers for all applications, especially for some later new applications, so it is not always possible to have a one-to-one correspondence between network port numbers and applications. Second, some common protocols do not use fixed port numbers for data transmission. In addition, services of multiple network protocols can be packaged as common applications and use the same port number, while the traffic classification and identification method based on network port mapping can no longer solve these problems. The accuracy and reliability of this method are also declining, and it can no longer meet the requirements of network traffic classification and identification.

2.3.2 Flow identification based on traffic load

The traffic classification and identification method based on payload determines the network traffic category by analyzing whether the payload of a network packet matches the feature identification library. Due to the low accuracy of port mapping in traffic identification, SenS proposed an in-depth message detection method based on application-layer protocol feature fields [30]. This method needs to pre-establish the application layer feature recognition rule base of network traffic and verify whether it matches a feature recognition rule in the rule base by analyzing the key control information in the payload, so the process of network traffic classification and identification can also be considered as the process of pattern verification when using this method in practice. At present, with the increase in network bandwidth and the influx of a large amount of data into the Internet, the emergence of new network applications and the continuous updating of existing network applications, the flow identification feature rules that need to be stored in the rule base are also expanding rapidly, and the processing and storage costs of the system are increasing. Moreover, more importantly, a complete network payload analysis is not only expensive to compute, but also may involve user privacy disputes and data security leaks [31]. Therefore, it has encountered some resistance in its development process.

2.3.3 Traffic identification based on machine learning

Machine learning has the ability of data mining and can extract implicit, regular and effective information from large data. Network traffic contains huge and complex data, and now academia is focusing on the method of traffic identification based on machine learning. Machine learning has many algorithms to learn from, and network traffic has many characteristics to choose from. Andrew Moore et al. gave 248 traffic characteristics to choose from. The traffic classification and recognition methods based on machine learning can be roughly divided into the traffic classification and recognition methods based on supervised learning, the traffic classification and recognition methods based on unsupervised learning and the traffic classification and recognition methods based on reinforcement learning.

2.4 ZigBee

ZigBee is a technical standard for building personal area networks using 2.4G band communication. Its implementation of MAC and Physical Layer follows IEEE 802.15.4Standard has the advantages of simple implementation, low power consumption, automatic

Xie *et al. J Wireless Com Network*   (2021) 2021:108

Page 5 of 15

network formation and the ability to meet different functional requirements with a variety of topological structures. The disadvantage is that the product development is difficult, the development cycle is long, and the product cost is high [5, 16, 19]. ZigBee-based wireless sensor network can be set up, which can be used in many aspects such as energy-saving research as well as data collection, processing and transmission.

Like the ZigBee protocol, Z-Stack uses a hierarchical software architecture, in which the HAL manages tasks using a time slice polling mechanism and provides a multitask management mechanism. It provides driver interfaces for a variety of peripheral modules, including timers, GPIO, universal asynchronous transceiver, analog-to-digital converters, etc. It also provides other extended service items. The OSAL is used to provide application developers with the lower-level interfaces and services required by the upper-level applications.
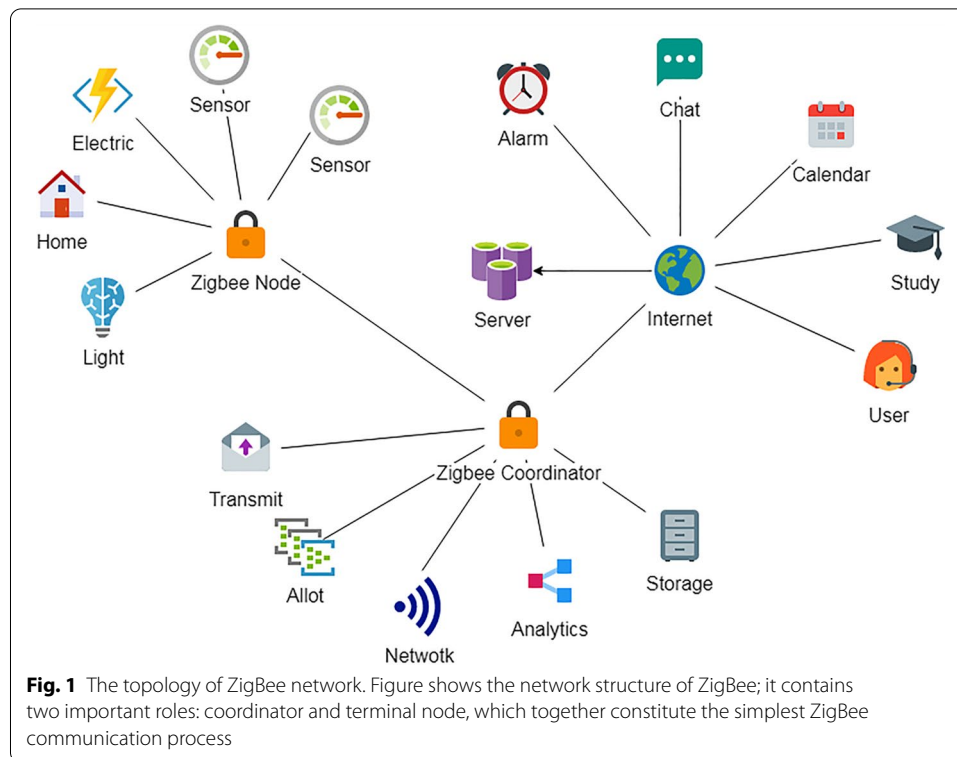
3 Methods

3.1 ZigBee networking

This paper implements ZigBee communication function based on the ZigBee protocol stack Z-Stack introduced by TI. It supports a variety of microcontrollers, including CC2530 on-chip system, CC2520 in MSP430 series and LM3S9B96 in Stellaris series. This protocol stack includes a variety of network topologies and is widely used in ZigBee industry.

The protocol stack defines how communication hardware and software communicate across layers in different hierarchies. For the data sender, the information packets sent by the user pass through each protocol layer in order from high to low, and each layer's entity adds its own unique identity information in a defined format, reaching the physical layer, transferring in the physical link as a binary stream and reaching the data receiver. When the data receiver receives the data stream, the data packet passes through each protocol layer in order from lowest to highest, and the entities of each layer extract the data information from the data packet in a predefined format that needs to be processed at this layer, and finally reach the application layer.

Figure 1 shows the network structure of ZigBee; it contains two important roles: coordinator and terminal node, which together constitute the simplest ZigBee communication process. The internal network is 2.4G wireless communication interaction. The external network is peripheral devices such as sensors and the Internet. The control of household appliances and environmental monitoring functions can be achieved through the interaction between the internal network and the external network. The coordinator role is the relay of the entire ZigBee network, which scans the current network condition, chooses the appropriate channel and network ID, and then starts ZigBee network; in addition, it will participate in assisting in configuring security parameters and application bindings within the network. In short, the coordinator role is primarily responsible for starting and configuring the network. Once it has completed its work, it can choose to switch to the router role or exit the current network. Such a change will not have any impact on the network as a whole; the terminal node itself is not responsible for the overall work of the network, it only needs to have the ability to sleep and wake up, go to

Xie *et al. J Wireless Com Network*   (2021) 2021:108

Page 6 of 15

Fig. 1 The topology of ZigBee network. Figure shows the network structure of ZigBee; it contains two important roles: coordinator and terminal node, which together constitute the simplest ZigBee communication process

sleep when not working, extend standby time and wake up quickly once it receives the wake-up command from the coordinator.

3.2 Establishing a traffic model based on machine learning

In machine learning for traffic analysis, the effectiveness of the assessment requires data support and training. This paper uses Cambridge University's Moore dataset as a training test set for traffic classification. The dataset uses a high-performance network monitor to collect data, and provides a time stamp with resolution over 35 ns, and consists of many objects, each of which is described by a set of features. This dataset was selected because it is close to the actual network situation. With a large amount of manually classified data, each object in each dataset represents a single TCP packet flow between the client and the server [16]. The characteristics of each object include classifications derived elsewhere and many derived features as inputs to probability classification techniques. The information in the features is exported using the header information alone, while the classification classes are exported using content-based analysis.

The dataset contains 10 sub-datasets, totaling 377,536 data and 249 features. The 11 traffic types involved are WWW, FTP, DATABASE, P2P, SERVICE, MAIL ATTACK, etc. Each subset of the set is characterized as shown in Table 1.

In this table, duration refers to the time from the beginning of monitoring to the end of data flow. Each sub dataset has a different number of streams and a different

Table 1 Dataset characteristics

Data subset	Duration (s)	Number of streams
Entry01	1821.8	24,863
Entry02	1696.7	23,801
Entry03	1724.1	22,932
Entry04	1784.1	22,285
Entry05	1794.9	21,648
Entry06	1658.5	19,384
Entry07	1739.2	55,835
Entry08	1665.9	55,494
Entry09	1664.5	66,248
Entry010	1613.4	65,036

duration. The streams contained in it are TCP streams, so they have clear start and end identifiers. Each stream corresponds to different traffic types. Based on machine learning, classification models can be trained to classify the actual traffic.
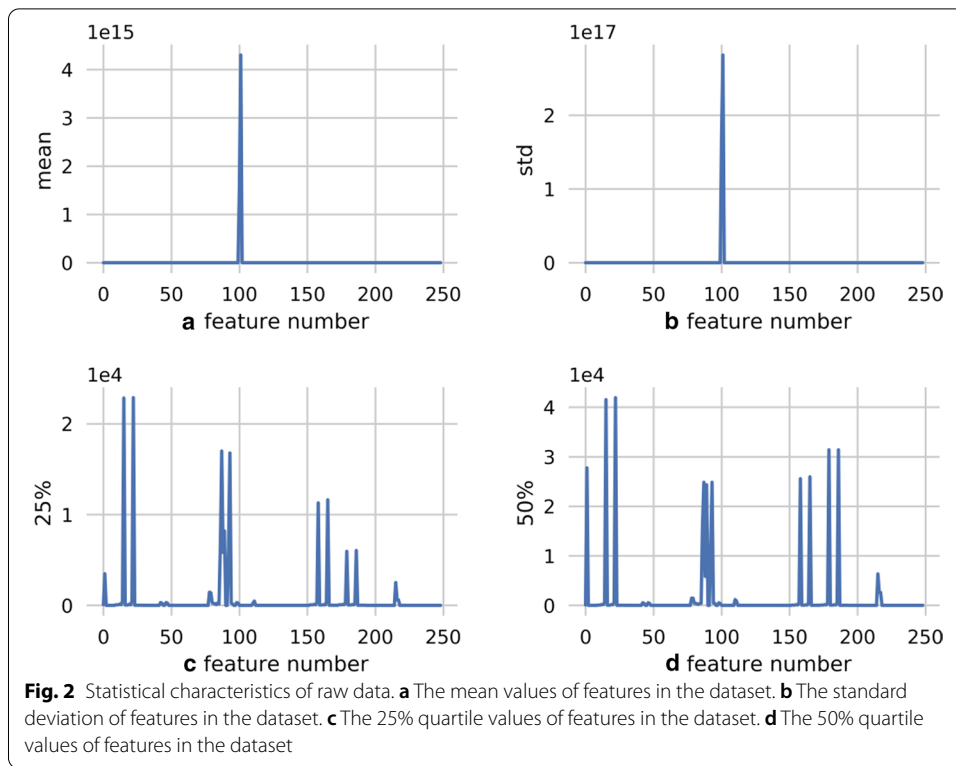
3.3 Data pre-processing

The dataset itself is not always perfect. Some datasets have different data types, such as text, numbers, time series, continuity and discontinuity. It is also possible that the quality of the data is not good, there is noise, there are anomalies, there are missing, the data is wrong, the dimensions are different, there are duplicates, the data is skewed, and the amount of data is too large or too small. In order for the data to fit the model and match the needs of the model, the Moore dataset needs to be pre-processed, detected from the data, corrected or deleted, inaccurate or inappropriate records for the model [8, 13]. Data pre-processing methods include removing unique attributes, processing missing values, attribute coding, data standardization regularization, feature selection, principal component analysis and so on.

In machine learning, most algorithms, such as logistic regression, support vector machine SVM and *k*-nearest neighbor algorithm, can only process numeric data and cannot process text. In sklearn, in addition to the algorithm used to process text, other algorithms require all input arrays or matrices during training, and cannot import text-based data. Some of the data in this dataset contains the characters Y and N, which cannot be processed directly using machine learning algorithms. You can encode Y as 1 and N as 0 through attribute encoding. During the network connection process, the maximum segment size cannot be known, so the dataset is represented by a'?', so there are consecutive features in the dataset that appear as'?'. For this reason, we use mean filling with Gauss white noise.

From Fig. 2, it can be seen that the standard deviation and mean values of some features in the dataset are unusually large, reaching 10e17 and 10e15. For such feature data, data regularization is used. For a single sample, the sample is scaled to the unit norm for each sample. The specific process is as follows:

Xie *et al. J Wireless Com Network* (2021) 2021:108

Page 8 of 15

Fig. 2 Statistical characteristics of raw data. **a** The mean values of features in the dataset. **b** The standard deviation of features in the dataset. **c** The 25% quartile values of features in the dataset. **d** The 50% quartile values of features in the dataset

Dataset D is defined as: $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_n, y_n)\}, \vec{x}_1 = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)})^{\mathrm{T}}$.

Calculate L_p norm: $L_p(\vec{x}_i) = \left(|x_i^{(1)}|^p + |x_i^{(1)}|^p + \cdots + |x_i^{(d)}|^p \right)^{\frac{1}{p}}$.

Calculate sample regularization: $\vec{x}_i = \left(\frac{x_i^{(1)}}{L_p(\vec{x}_i)}, \frac{x_i^{(2)}}{L_p(\vec{x}_i)}, \ldots, \frac{x_i^{(d)}}{L_p(\vec{x}_i)} \right)^{\mathrm{T}}$.
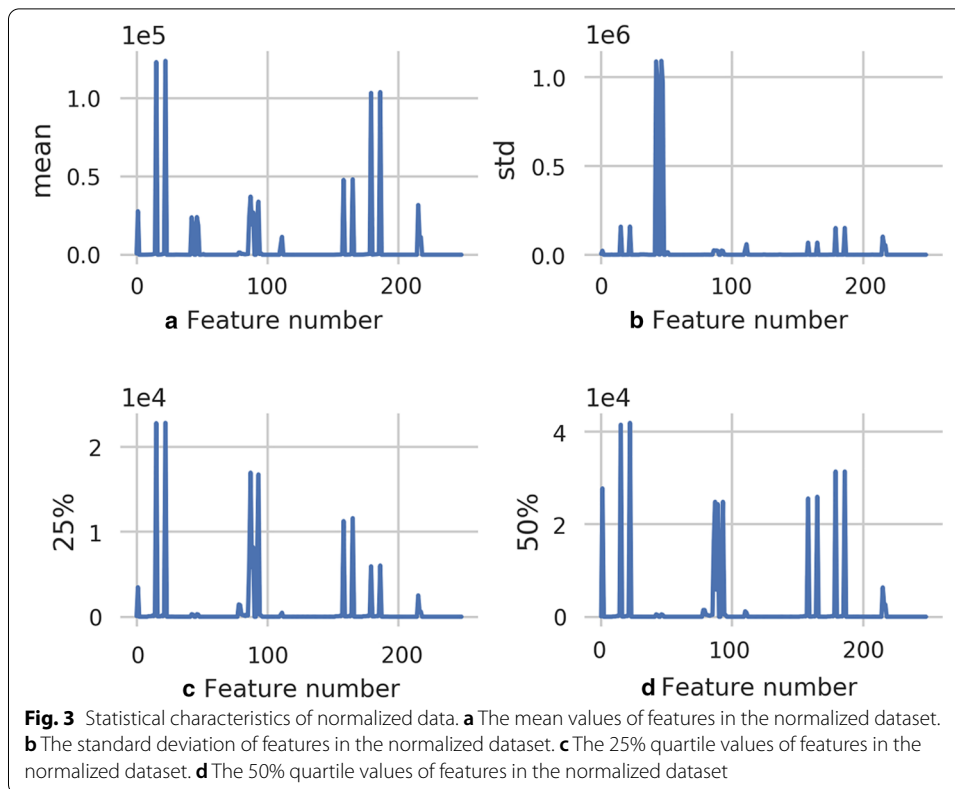
This paper recalculates the statistical characteristics by simply filling and replacing the data with abnormal features and normalizing the data. Figure 3 describes the dataset with some statistical features, including standard deviation, mean and 25% bits as median.

3.4 Data feature processing

When data pre-processing is complete, we need to select meaningful features to input into the machine learning algorithm and model for training. When exploratory analysis of data reveals that there are too many features introduced. To model and analyze directly with these features, further screening of the original features is required and only important features are retained. Generally, features are selected from two perspectives:

Whether a feature is divergent or not: If a feature does not diverge, for example, if the variance of a feature itself is small, then there is little difference in the sample on this feature. Maybe most of the values in the feature are the same, or even the values of the whole feature are the same, then this feature has no effect on sample differentiation.

Fig. 3 Statistical characteristics of normalized data. **a** The mean values of features in the normalized dataset. **b** The standard deviation of features in the normalized dataset. **c** The 25% quartile values of features in the normalized dataset. **d** The 50% quartile values of features in the normalized dataset

Relevance of features to objectives: Features that are highly relevant to objectives should be selected. In addition to the variance method, the other methods described in this paper are concerned with correlation.

According to the form of feature selection, there are three feature selection methods:

Filter: A filter method that scores each feature according to divergence or correlation, sets thresholds or the number of thresholds to be selected and selects features.

Wrapper: Packaging method that selects several features at a time based on the objective function or excludes several features, such as recursive elimination of features using a base model for multiple rounds of training. After each round of training, the features of several weight coefficients are eliminated, and then, the next round of training is based on a new set of features.

Embedded: Embedded method, first uses some machine learning algorithms and models to train, get the weight coefficients of each feature and select features from large to small according to the coefficients. Similar to the filter method, but trained to determine the quality of the features.

To explore the performance of different algorithms in the model, this paper chooses different feature selection algorithms to obtain the best traffic classification model through comparison.

3.4.1 Variance filtering

To select the optimal hyperparameter, you can draw a learning curve to find the best point of the model. However, it takes a lot of time, and the improvement of the model is limited. In this paper, variance filtering with a threshold of 0.001 is used to first eliminate some features that are obviously not needed, and then select a better feature selection method to continue to reduce the number of features. By variance filtering, features with variances less than thresholds are removed, leaving 240 features.

After selecting the variance, the next step is to select meaningful features related to the target tag, which can provide a lot of information. If the feature is not tagged, it will simply waste computing memory and possibly noise the model. Here, three common methods can be used to assess the correlation between features and labels: Chi-square, F test and mutual information.

3.4.2 Chi-square filtration

Chi-square filtering is a correlation filtering specifically for discrete tags. The Chi-square test calculates the Chi-square statistics between each nonnegative feature and label and ranks them according to the characteristics of the Chi-square statistics from high to low. Combined with the scoring criteria, the classes with the highest K-score were selected to remove features that are most likely independent of labels and unrelated to the purpose of classification. In addition, if the Chi-square test detects that all values in a feature are the same, it will prompt variance filtering using the difference first. However, the selection of K value is closely related to the performance of the model. In order to obtain the best K value, we need to find ways to explore the best K value.

The F test, also known as ANOVA, variance homogeneity test, is a filtering method used to capture the linear relationship between each feature and a label. It can be used for regression or classification, where F test classification is used for data with labels as discrete variables and F test regression is used for data with labels as continuous variables. The output statistics can be used directly to determine what kind of K we want to set. It is important to note that the F test is very stable when the data follows a normal distribution, so using F test filtering first converts the data into a normal distribution. The essence of F test is to find a linear relationship between two sets of data, assuming that there is no significant linear relationship between the data. It returns two statistics, F and P. As with Chi-square filtering, we want to select features with P values less than 0.05 or 0.01 that are significantly linear with the label, while features with P values greater than 0.05 or 0.01 are considered features that have no significant linear relationship with the label and should be deleted.

Mutual information is a filtering method used to capture any relationship (both linear and nonlinear) between each feature and the label. Similar to the F test, it can be used for both regression and classification, and it includes both mutual information classification and mutual information regression. Both classes have the same usage and parameters as the F test, but the mutual information method is more powerful than the F test, which can only find linear relationships, while the mutual information method can find any relationships. Mutual information does not return statistics with similar P or F values. It returns an estimate of the amount of mutual information between each feature and the

target, which takes a value between [0, 1]. A value of 0 indicates that the two variables are independent and a value of 1 indicates that the two variables are fully correlated.

3.4.3 Lasso

Lasso algorithm seeks the smallest sum of squares of residuals when the sum of absolute values of model coefficients is less than a constant. It is better than stepwise regression, principal component regression, ridge regression, partial least squares and so on in variable selection. It can better overcome the shortcomings of traditional methods in model selection. Lasso regression is one of the regularization methods and is a compressed estimation. It obtains a more refined model by constructing a penalty function. Using it to compress some coefficients while setting some coefficients to zero preserves the advantage of subset shrinkage and is a biased estimate for processing data with multicollinearity. Lasso is a shrinkage estimation method based on the idea of reducing the feature set. Lasso method can compress the coefficients of features and make some regression coefficients 0, which can be used for feature selection. Lasso method can be widely used in model improvement and selection. By choosing a penalty function, Lasso's ideas and methods are used to achieve the purpose of feature selection. Model selection is essentially a process of seeking sparse representation of a model, which can be accomplished by optimizing a loss + penalty function problem. The advantage of Lasso regression method is that it can make up for the deficiencies of least squares estimation and stepwise regression local optimal estimation. It can select features well and effectively solve the problem of multicollinearity among features. Its objective function can be expressed as:

$$J(\beta) = \sum (y - X\beta)^2 + \lambda ||\beta||_1 = \sum (y - X\beta)^2 + \sum \lambda |\beta|$$
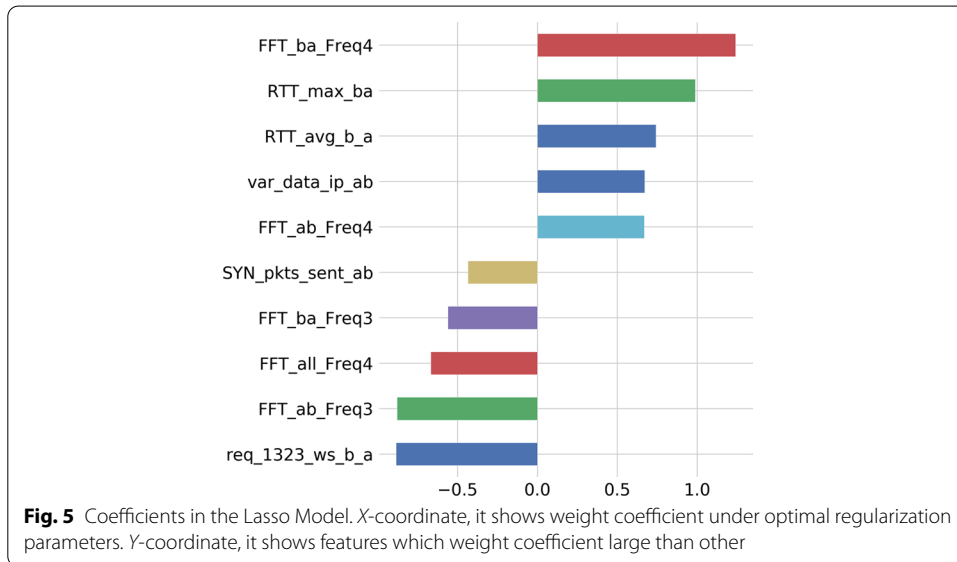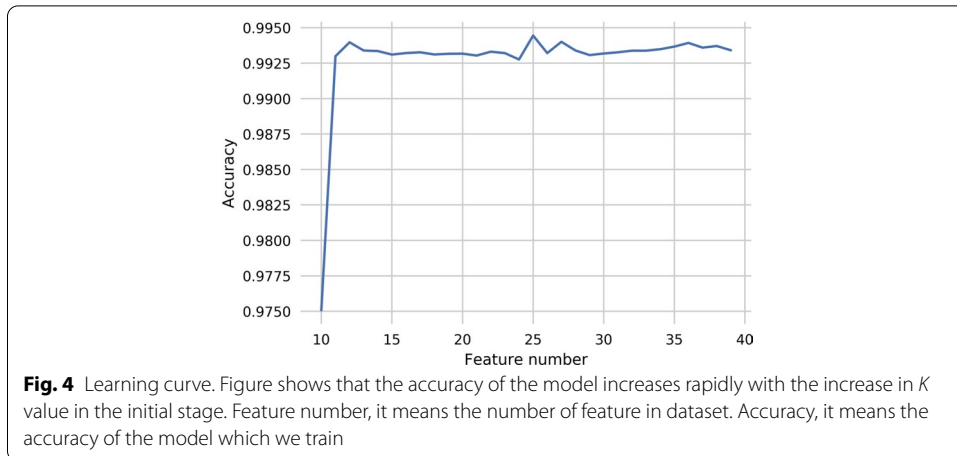
3.5 Model training

The training process of machine learning is to first define a loss function, add input samples and get prediction tests based on forward propagation. Compared with the real sample, the loss value is obtained, and then, the reverse propagation is used to update the weight value, iterating back and forth continuously until the loss function is small and the accuracy reaches the ideal value. The parameters at this point are those required by the model. That is, the ideal model is built. This paper divides the dataset into training group and testing group; the ratio is 8 to 2. First, the training data is used to preliminarily train the model, and then, a preliminary model is obtained. Then, the test data is used to test the model to see if there is any phenomenon of fitting.

4 Results and discussion

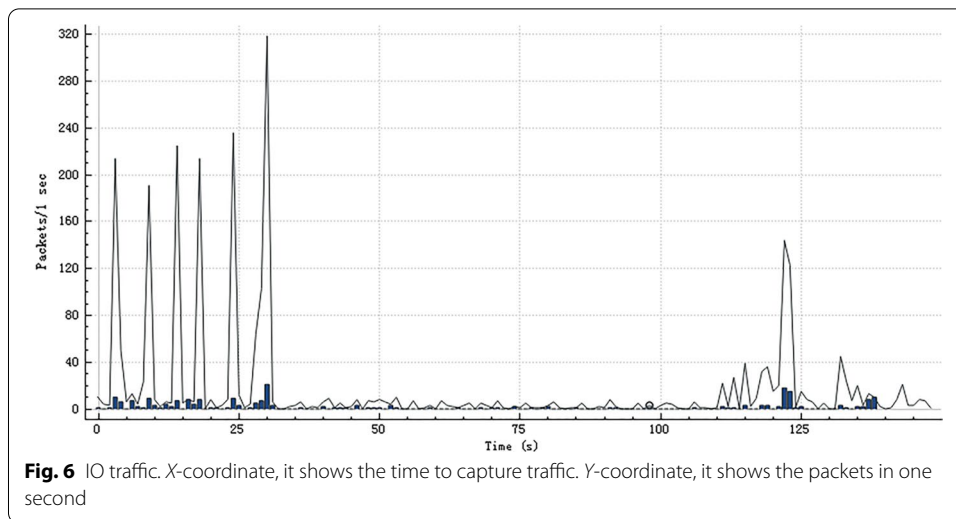### 4.1 Chi-square filtration test

This paper combines the data characteristics of the Moore dataset, chooses Chi-square filtering method, sets appropriate threshold parameters, sets K value as learning parameter through learning curve and uses the accuracy of model training as evaluation index to get learning curve.

Xie *et al. J Wireless Com Network* (2021) 2021:108

Page 12 of 15

Fig. 4 Learning curve. Figure shows that the accuracy of the model increases rapidly with the increase in K value in the initial stage. Feature number, it means the number of feature in dataset. Accuracy, it means the accuracy of the model which we train

Fig. 5 Coefficients in the Lasso Model. *X*-coordinate, it shows weight coefficient under optimal regularization parameters. *Y*-coordinate, it shows features which weight coefficient large than other

According to the results of Fig. 4, we can see that the accuracy of the model increases rapidly with the increase in K value in the initial stage. When $K=12$, the accuracy decreases, and then, there is a small fluctuation with the increase in K. From the result of the curve feedback, the best K value is 25, so it is only necessary to model according to these 25 features to get a good classification model. Finally, a random forest cross-validation training model with K 25 can be obtained, and the accuracy is 0.9944.

4.2 LassoCV test

This article uses the LassoCV class in liner model of Python to obtain the weight coefficients of each features under the optimal regularization parameters to achieve the

Fig. 6 IO traffic. *X*-coordinate, it shows the time to capture traffic. *Y*-coordinate, it shows the packets in one second

feature selection of the dataset. It selects the first five features of the absolute value of the coefficient in Fig. 5.

It can be seen from Fig. 5 that the feature weight of them is high which include FFT_ba_Freq4, RTT_max_ba, RTT_avg_b_a, var_data_ip_ab, FFT_ab_Freq4, SYN_pkts_sent_ab, FFT_ba_Freq3, FFT_all_Freq4, FFT_ab_Freq. In particular, the characteristics of FFT of packet IAT are particularly prominent, and it is very important to identify the type of traffic. In this paper, 20 features with absolute value of feature weight ranking in the front are selected for model training, and the accuracy is only 0.755, which is much lower than the previous Chi-square filtering effect. The reason is that Lasso is an algorithm for multicollinearity problems, which limits the impact of multicollinearity. Lasso performs better in applications where datasets are linearly correlated. However, in this paper, the performance of this algorithm is poor. It can also be seen that Moore dataset is nonlinear.

4.3 Grasp the flow for classification

As shown in Fig. 6, we use the software to capture the traffic data of a certain period of time, and make statistics of the data packets in the period. In the later stage, we will extract some important features of the traffic according to the Moore dataset, input them into the model to identify the traffic types that pass in this period of time, and get the statistical results of traffic categories in the periodic time window.

5 Conclusions

The Internet of things and 5G have been gradually popularized in the daily life of the general public. More and more intelligent home appliances have entered the family. Coupled with the frequent use of the Internet in life, network slice is becoming more and more mature, and the data traffic has increased dramatically. This paper selects the family as the research scene, combined with the Internet of things, and designs the family traffic analysis system, which can help family members understand the family's Internet traffic statistics, identify the invasion of malware or attacks and can also judge whether

there is abnormal according to the sensor traffic data uploaded by home appliances, and solve the problem of traffic island. It has good expansibility, high recognition accuracy and easy integration. In this paper, we use experiment to compare different machine learning algorithms in feature selection. Different algorithms perform differently in different datasets. There is no absolutely good algorithm, but in this paper, because the dataset is not nonlinear, the Chi-square filtering algorithm has obvious advantages. In this paper, the accuracy rate is almost 100%, which provides a good model reference for the later actual traffic classification.

Abbreviations
DBSCAN: Density-based spatial clustering with noise; MAC: Media access control; SFS: Sequential forward selection; EM: Expectation maximization; HAL: Hardware abstraction layer; IAT: Inter arrival time; OSAL: Operating system abstraction layer.

Authors' contributions
All authors take part in the discussion of the work described in this paper. All authors read and approved the final manuscript.

Availability of data and materials
The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Artificial Intelligence, Nanning College for Vocational Technology, Nanning 530008, Guangxi, China. [2]School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China.

References
1. Y. Sun, S. Qin, G. Feng,L. Zhang, M.A. Imran, Service provisioning framework for RAN slicing: user admissibility, slice association and bandwidth allocation. IEEE Trans. Mob. Comput. pp. 99 (2020)
2. G. Zhao, S. Qin, G. Feng, Y. Sun, Network slice selection in softwarization-based mobile networks. Trans. Emerg. Telecommun. Technol. **31**(1), e3617 (2020)
3. Y.J. Liu, G. Feng, Y. Sun, S. Qin, Y.C. Liang, Device association for RAN slicing based on hybrid federated deep reinforcement learning. IEEE. Trans. Veh. Technol. pp. 99 (2020)
4. Y. Sun, W. Jiang, G. Feng, P.V. Klaine, L. Zhang, M.A. Imran, Y.C. Liang, Efficient handover mechanism for radio access network slicing by exploiting distributed learning. IEEE Trans. Netw. Serv. Manag. **17**, 2620–2633 (2020)
5. Y. Mu, G. Feng, J.H. Zhou, Y. Sun, Y.C. Liang, Intelligent resource scheduling for 5G radio access network slicing. IEEE Trans. Veh. Technol. **68**, 7691–7703 (2019)
6. L.F. Yang, J.Y. Luo, Y. Xu, Z.R. Zhang, Z.Y. Dong, A distributed dual consensus ADMM based on partition for DC-DOPF with carbon emission trading. IEEE Trans. Ind. Inform. **16**, 1858–1872 (2020)
7. Z.R. Zhang, Y. Yu, S.N. Fu, Broadband on-chip mode-division multiplexer based on adiabatic couplers and symmetric Y-junction. IEEE Photonics J. **9**, 1–6 (2017)
8. Z. Zhang, J. Li, Y. Wang, Y. Qin, Direct detection of pilot carrier-assisted DMT signals with pre-phase compensation and imaginary noise suppression. J. Lightwave Technol. **39**, 1611–1618 (2020)
9. H.M. Huang, Y. Yu, L.J. Zhou, Y.Y. Tao, J.B. Yang, Z.R. Zhang, Whispering gallery modes in a microsphere attached to a side-polished fiber and their application for magnetic field sensing. Opt. Commun. **478**, 126366 (2020)
10. H. Li, K. Ota, M. Dong, Learning IoT in edge: deep learning for the Internet of Things with edge computing. IEEE Netw. **32**(1), 96–101 (2018)
11. S. Eugene, T. Thanassis, H. Wendy, Analytics for the Internet of Things. ACM Comput. Surv. **51**(4), 1–36 (2018)
12. L. Yu, W. Wang, C. Runze, Zigbee-based IoT smart home system% the design of internet of things smart home system based on zigbee. Electron. Test. **000**(005), 71–75 (2016)

13. C. Mingwu, T. Guilin, Non-contact monitoring system for high-temperature industrial furnace based on internet of things technology. J. Hebei North Univ. (Natural Science Edition) **035**(005), 42–45 (2019)

14. Z. Kaisheng, T. Kaiyuan, L. Ming, L. Chao, Design of agricultural greenhouse environment monitoring system based on internet of things technology. J. Xi'an Univ. Sci. Technol. **035**(006), 805–811 (2015)

15. Z.H. Wu, Research on the application of internet of things technology to digital museum construction. Acta Geosci. Sin. **38**(2), 293–298 (2017)

16. Y. Wang, The innovation of computer internet of things technology in logistics field. Log. Technol. **040**(003), 41–42 (2017)

17. T. Jiya, J. Feng, Laser detection and control system based on internet of things technology. Laser Ma. **040**(003), 153–157 (2019)

18. X.Q. Wang, Research on multi-view semi-supervised learning algorithm based on co-learning. Int. Conf. Mach. Learn. Cybern. IEEE **20**(6), 1276–1280 (2016)

19. J. Wang, X.J. Cheng, J.Q. Liu, Y.J. Wen, A enhanced algorithm based on RSSI and quasi Newton method for the node localization in wireless sensor networks. Comput. Knowl. Technol. **12**(8), 222–225 (2016)

20. Z.H. Huang, X. Xu, H.H. Zhu, M.C. Zhou, An efficient group recommendation model with multiattention-based neural networks. IEEE Trans. Neural Netw. Learn. Syst. **31**, 4461–4474 (2020)

21. Y. Sun, C. Xu, G.F. Li, W.F. Xu, J.Y. Kong, D. Jiang, B. Tao, D.S. Chen, Intelligent human computer interaction based on non redundant EMG signal. Alex. Eng. J. **59**, 1149–1157 (2020)

22. W. Wei, H. Song, W. Li, P. Shen, A. Vasilakos, Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. Inf. Sci. **408**(2), 100–114 (2017)

23. S.M.M. Gilani, T. Hong, W. Jin, G. Zhao, H.M. Heang, C. Xu, Mobility management in IEEE 802.11 WLAN using SDN/NFV technologies. EURASIP J. Wirel. Commun. Netw **67**(12), 56–62 (2017)

24. K. Dongyeon, B. Hong, W. Choi, Probabilistic caching based on maximum distance separable code in a user-centric clustered cache-aided wireless network. IEEE Trans. Wirel. Commun. **18**, 1792–1804 (2019)

25. L. Ruoyu, Indexing of CNN features for large scale image search. Pattern Recognit. **48**(10), 2983–2992 (2018)

26. R. Deboleena, P. Panda, K. Roy, Tree-CNN: a hierarchical deep convolutional neural network for incremental learning. Neural Netw. **121**, 148–160 (2018)

27. J. Hongsheng, A deep 3D residual CNN for false-positive reduction in pulmonary nodule detection. Med. Phys. **45**, 2097–2107 (2018)

28. M. Ma, Region-sequence based six-stream CNN features for general and fine-grained human action recognition in videos. Pattern Recognit. **76**, 506–521 (2018)

29. Z. Chen, D. Chen, Y. Zhang, X. Cheng, M. Zhang, C. Wu, Deep learning for autonomous ship-oriented small ship detection. Saf. Sci. **130**, 104812 (2020)

30. V.P. Rao, D. Marandin, Adaptive channel access mechanism for zigbee (IEEE 802.15.4). J. Commun. Softw. Syst. **2**(4), 283–293 (2017)

31. W.Q. Huang, M. Zhang, D. Wei, D.G. Sun, J. Shi, Efficient and anti-interference method of synchronising information extraction for cideo leaking signal. IET Signal Proc. **10**(1), 63–68 (2016)

Publisher's Note