Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Network Traffic Classification Based on Deep Learning

To cite this article: Jun Hua Shu *et al* 2018 *J. Phys.: Conf. Ser.* **1087** 062021

View the article online for updates and enhancements.

Network Traffic Classification Based on Deep Learning

Jun Hua SHU[1], Jiang JIANG[2], Jing Xuan SUN[3]

School of control and computer engineering, North China Electric Power University, Beijing 102206, China.

xxdyx110@126.com

Abstract. With the rapid development of the Internet, the network has been expanding. In order to allow network operators to provide better quality of service, but also the effective supervision and management of the network, identify the traffic type of network traffic classification technology has become a hot topic in recent years. In this paper, a method of network traffic classification based on deep learning is proposed. Compared with the traditional machine learning method, the accuracy of network traffic classification based on deep learning has been improved obviously.

1. Introduction

In recent years, with the rapid development of the Internet, the network traffic data also showed explosive growth, while giving people convenience, but also to the effective network management, security, network environment has brought great challenges, such as virus flooding、it is difficult to monitor the network of unhealthy content, P2P applications take a lot of network bandwidth and other issues.

In view of the problems in these networks, network researchers have proposed capacity planning, traffic scheduling and other strategies to improve the operational efficiency of the network. However, the premise of these strategies is to classify the traffic, so the network traffic classification technology is more and more by many network research scholars and network service providers. At present, the network traffic classification technology is mainly used in service quality and traffic engineering, network security monitoring, network management technology[1].

2. Research status

The traditional method of network traffic classification is based on the network port implementation. Through the network protocol and application software for data transmission when the general port requirements, the port number and the specific network protocol, application software one by one to determine the use of the port of the network traffic type to complete the identification of network traffic. However, there are many new applications (Streaming, Gaming and P2P) appearing in addition to traditional applications, the port-based protocol identification method is no longer reliable[2].

The classification method based on feature field recognition is to pre-establish the application layer identification rule base for network traffic generated by each network application. It is to establish the application layer identification rule base of the network traffic generated by each network application in advance, to rule out the data content of the data stream to be identified, and to determine the application type of network traffic according to the matching result. However, the classification method based on feature field recognition can only identify the existing traffic type, but also can not contribute to the encrypted data. Therefore, in today's widely used load encryption technology and traffic types continue to emerge, based on the feature field recognition classification method has a greater limitation.

In view of the limitations of the above methods, in recent years, more and more network researchers use the statistical characteristics of network traffic and machine learning methods to classify network traffic. Commonly used in the network traffic classification of machine learning methods are Naive Bayesian method and C4.5 decision tree method.

3. Machine learning classification method

3.1. Naive Bayesian

The naive Bayesian method is a method of modeling the probability relation between attribute set and type variable by Bayes theorem. Suppose there are n data streams, $\{x_1, x_2, \cdots, x_n\}$ in the network, each data stream has m attributes $\{A_1, A_2, \cdots, A_m\}$, $x_i = \{A_1{}^{(i)}, A_2{}^{(i)}, \cdots, A_m{}^{(i)}\}$ represented the characteristic attribute of the i-th data stream. $Y = \{y_1, y_2, \cdots, y_k\}$ denote k traffic application types, that is, the target type. For each data stream x_i, need to find a mapping $F: x \to Y$, indicates that each network traffic x is classified as a traffic application type. For an unclassified data stream x, the network type y_i to which it belongs depends on the posterior probability $p(y_j|x)$. The Bayesian method calculates the posterior probability as follows:

$$p(y_j|x) = \frac{p(y_j)p(x|y_j)}{\sum_{y_j} p(y_j)p(x|y_j)} \qquad (1)$$

$p(y_j)$ is the prior probability of y_j, which is independent of the traffic to be classified. $p(x|y_j)$ is the probability density function of x in the case of given y_j. The key to naive Bayesian is to give the probability $p(x|y_j)$. Naive Bayesian assumes that all attributes A_i are independent and obey the Gaussian distribution. Calculate the maximum posterior probability:

$$p(y_{max}|x) = max\, p(y_j|x) \qquad j\epsilon(1, |Y|) \qquad (2)$$

you can calculate the best category y_{max} of x.

Although the Naive Bayesian independence hypothesis is difficult to set up, the Naive Bayesian method performs better than most of the more complex classification methods and can handle very complex situations. However, the naive Bayesian method uses the proportion of various types of samples in the training data to estimate the prior probabilities of the various samples, which can lead to the instability of classification accuracy[3,4].

3.2. C4.5 Decision Tree

C4.5 Decision tree method is to use the information entropy in the training data set to construct the classification model, so as to realize the classification of unknown samples.

The training phase of the decision tree consists of two processes of building and pruning. In the process of building the algorithm, the training samples are selected and the samples are divided according to certain rules. Each partition is divided according to a certain attribute, and a node is generated. This attribute is the basis for dividing the sample in the classification[5].

The pruning process uses the remaining training samples as the test set, prudently verifies the generated decision tree records, trims or adds nodes to the incorrectly classified branches, and finally forms the decision tree.

C4.5 algorithm occupies an important position in the research of network traffic classification based on machine learning. It has good classification accuracy and data processing speed. However, in the case of huge amount of network traffic data, tree generation time is greatly affected[6,7].

4. Network traffic classification based on deep learning:

With the rise of Big Data and the improvement of computer computing ability, the deep learning has been getting more and more researchers attention. Machine learning algorithms such as naive Bayesian and decision trees are highly dependent on the choice of feature attributes, such as the absence of feature selection for naive Bayesian classification, the accuracy rate is only about 65%. However, the average classification accuracy after feature selection is about 95%, it can be seen the importance of feature selection for machine learning. Deep learning can automatically combine the low-order features of the input, transform, arrange the combination, get high-order features, eliminating the need for manual construction of high-order features of the workload[8].

In a broad sense, the deep learning network structure is also a multi-layer neural network, multi-layer neural network mainly includes input layer, hidden layer and output layer:

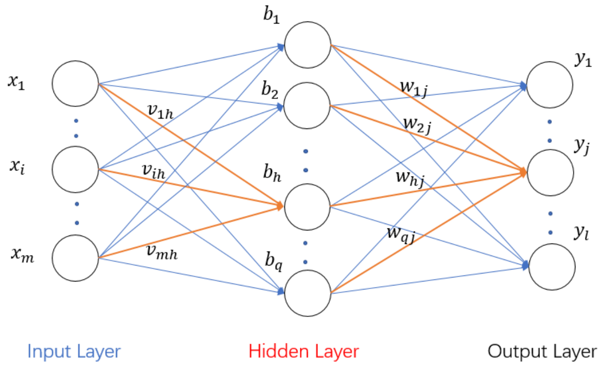

Figure 1 Deep Learning Network Structure

The hidden layer can have multiple layers. The parameters in the network are randomly initialized, and the weighting parameters in the network are adjusted by the Back Propagation (BP) algorithm. Assuming the network result is shown in Fig. 1, the input of the j-th output neuron:

$$\beta_j = \sum_{h=1}^{q} w_{hj}\, b_h \qquad (3)$$

The input of the h-th hidden neurons:

$$\alpha_h = \sum_{i=1}^{d} v_{ih}\, x_i \qquad (4)$$

For a training sample (x_k, y_k), it is assumed that the output of the neural network is $\hat{y}_k = (\hat{y}_1^k, \hat{y}_2^k, \cdots, \hat{y}_l^k)$, ie $\hat{y}_j^k = f(\beta_j - \varepsilon_j)$, Where ε_j is the offset term, the mean square error of the network on (x_k, y_k) is:

$$E_k = \frac{1}{2}\sum_{j=1}^{l}(\hat{y}_j^k - y_j^k)^2 \qquad (5)$$

BP algorithm is based on gradient descent strategy, the goal is to minimize the mean square error E_k:

$$\Delta w_{hj} = -\theta \frac{\partial E_k}{\partial w_{hj}} \qquad (6)$$

Application of chain rules for derivation, can eventually get:

$$\Delta w_{hj} = \theta \cdot \hat{y}_j^k (1 - \hat{y}_j^k)(y_j^k - \hat{y}_j^k) \cdot b_h \qquad (7)$$

Through the BP algorithm can be based on the input training samples automatically adjust the weight of the entire network parameters to reduce the mean square error.

The traditional multi-layer neural network increases the number of layers, resulting in the gradient in the reverse propagation, the deeper the spread of the gradient closer to 0, resulting in the gradient disappeared. And with the increase in the number of layers, the number of weight parameters will be correspondingly increased, the more complex the model, and thus easier to overfit[9,10].

Deep learning by using the new activation function ReLU, the new weight initialization method, the new loss function, the new anti-fitting method (Dropout, regularization, etc.) to solve the traditional multilayer Disadvantages in the network.

5. Experimental environment and experimental data:

5.1. Experimental data:

This article uses the experimental dataset used by Moore et al. There are 10 types of data, each of which is abstracted from a complete TCP bidirectional stream containing 249 network flow attributes. Since the number of samples of some traffic types such as GAMES is too small, we only use eight of these types, each of which is shown in the following table:

Table 1 Number of each type

Type	WWW	MAIL	FTPDATA	DATABASE	FTPCONTROL	FTPPASV	P2P	SERVICES
Number	328092	28567	5797	2648	3054	2688	2094	2099

The same number of samples are taken from each flow type as experimental data, and when a certain type of sample is insufficient, all samples of this type are taken.

5.2. Experimental environment:

The deep learning framework used in this article is Google's TensorFlow, relative to other deep learning framework, TensorFlow has four following advantages : usability、flexibility、efficiency and support . TensorFlow is supported by Google, and Google has invested a lot of effort in developing TensorFlow, which expect TensorFlow to become a general framework for machine learning researchers and developers .

6. Experimental results and analysis:

This experiment uses a random forest, naive Bayesian and deep learning neural network algorithm for classification, the maximum tree depth of the random forest is 3, and the neural network has two hidden layers, and the number of neurons is 500 and 100 respectively. First, 100 samples of each type were extracted, the experimental data were randomly divided into two parts, one was trained as a training set and the other as a test set. The accuracy of the test set was obtained. Then gradually increase the number of samples of each type, the final results as shown below:

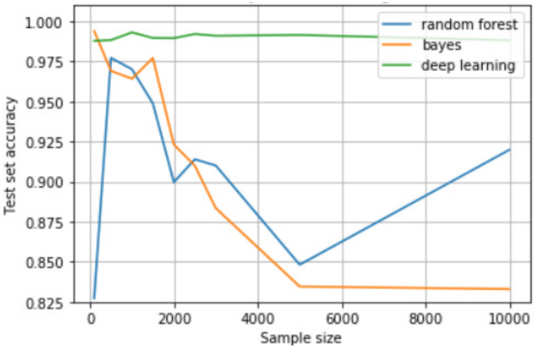

Figure 2 The Accuracy Of The Three Algorithms

It can be seen from the results that the Naive Bayesian method can not accurately fit the distribution of network flow attributes due to the direct use of the independence hypothesis, resulting in poor classification results. While the random forest can maintain a high accuracy, but the accuracy of the model with the increasing number of samples fluctuate. Deep learning neural network algorithm accuracy rate can reach 98.5% or more, and compared to the other two classification algorithms, the deep learning neural network algorithm accuracy is not only higher, but also more stable.

7. Concluding

Using the machine learning method to deal with the traffic classification problem is a hotspot of network traffic classification in recent years. However, with the rise of Big Data and the improvement of hardware computing ability, the deep learning has quickly become the new darling. In this paper, we compare the differences between depth learning and naive Bayesian and random forest algorithms in network traffic classification. The deep learning neural network has higher accuracy and better stability than other machine learning algorithms. It's accuracy rate can reach more than 98.5%. The deep learning neural network has a great advantage in the classification of network traffic without the need of characteristic selection and extremely powerful fitting ability.

References:

[1] Hong Zhi Wang,Li Hui Yan. A New Network Traffic Classification Method Based on Optimized Hadamard Matrix and ECOC-SVM[J]. Advanced Materials Research,2014,3326(989):.

[2] HaiTao He,XiaoNan Luo,FeiTeng Ma,ChunHui Che,JianMin Wang. Network traffic classification based on ensemble learning and co-training[J]. Science in China Series F: Information Sciences,2009,52(2):.

[3] Mikhail Dashevskiy,Zhiyuan Luo. Two methods for reliable classification of network traffic[J]. Progress in Artificial Intelligence,2012,1(3):.

[4] Haitao He,Chunhui Che,Feiteng Ma,Xiaonan Luo,Jianmin Wang. Improve Flow Accuracy and Byte Accuracy in Network Traffic Classification[M].Springer Berlin Heidelberg:2008.

[5] Alberto Dainotti,Antonio Pescapé,Carlo Sansone. Early Classification of Network Traffic through Multi-classification[M].Springer Berlin Heidelberg:2011.

[6] Xiaoling Tao,Yong Wang,Yi Wei,Ye Long. Network Traffic Classification Based on Multi-Classifier Selective Ensemble[J]. Recent Advances in Electrical &amp; Electronic Engineering,2015,8(2):.

[7] Aiqing Zhu. A P2P Network Traffic Classification Method Based on C4.5 Decision Tree Algorithm[M].Springer Berlin Heidelberg:2014.

[8] Sweta Keshapagu,Shan Suthaharan. Analysis of Datasets for Network Traffic Classification[M].Springer New York:2013.

[9] Lei Ding,Fei Yu,Sheng Peng,Chen Xu. A Classification Algorithm for Network Traffic based on Improved Support Vector Machine[J]. Journal of Computers,2013,8(4):.

[10] Auld Tom,Moore Andrew W,Gull Stephen F. Bayesian neural networks for internet traffic classification.[J]. IEEE Transactions on Neural Networks,2007,18(1):.