Traffic Classification of Power Communication Network Based on Improved Hidden Naive Bayes Algorithm

Lei lv，Qing Zhang，Shilun Zeng，Hong Wu，Liang Chen

(State Grid Meishan Electric Power Supply Company, sichuan 620010, China);

Keywords: Traffic classification; Traffic identification; Hidden Naive Bayes Algorithm

Abstract. With the deepening of power grid construction, power information and communication network bears more and more business. The introduction of network traffic classification is conducive to the rational distribution of network resources, improve service quality. This paper firstly introduces the requirements of traffic classification in electric power communication network, uses flow model based communication network traffic collection technology and machine learning based recognition method, and finally uses the supervised learning algorithm based on machine learning to determine the use of hidden naive Bayes algorithm And classifies the traffic of the power communication network.

Introduction

With the development of China's power grid, the structure of electric power communication network will be changed from traditional point-to-multipoint to dynamic network structure. There is a trend of traditional service change, service flow increasing and business category becoming more and more complicated. Network traffic classification, traffic analysis and evaluation, a reasonable allocation of limited network resources[1] .

Combining the characteristics of power information communication network and its specific requirements for network traffic classification, it is concluded that the method of traffic recognition based on machine learning is a new type of traffic recognition method with high accuracy and wide application range. For power communication Network such a network environment for this point is easy to do.

Demand for Traffic Classification in Power Information Communication Network

With the development of power system, the comprehensive information system of power system becomes more and more urgent. The construction of electric power information communication network is an important part of power system information construction.From the power enterprise management network often encountered problems, the need for a solution that allows network administrators to understand the details of the network use, network management personnel in a timely manner insight into the network operating conditions, to keep abreast of network applications Implementation. This requires that the traffic classification method of the electric power information communication network meet the requirements of high recognition rate and accuracy, high security, and can recognize and classify the port service and support the sensing and recognition technology of various applications[2].

Flow acquisition technology

Aiming at the traffic collection of power network information a new method of traffic collection is proposed, which is based on object and application communication network traffic, flow characteristic statistics. This method is based on the statistics of the flow statistics of each monitoring object in different time periods and the basic description unit of the traffic information. This information is derived periodically to describe the traffic characteristics of the network, complete the collection of traffic information and display[3].

A flow consists of two parts: a part is a five-tuple, that is, the source address, destination address, source port, destination port, protocol type, used to identify the basic attributes and information of each packet in the flow; Based on the flow model of traffic collection method, through the acquisition of the flow records and statistical processing to get the required link flow characteristics of the information to complete the flow of the collection and analysis. The sampling method has a good applicability, which can predict the general characteristics of the original network traffic from the basic characteristics of the collected samples, and greatly reduce the overhead and load of the system.

Recognition Method Based on Machine Learning

Network traffic identification needs to consider the feasibility and effectiveness of identification, that is, selection algorithm should have a lower computational complexity, efficient and fast processing of a large number of network traffic, but also has a good nonlinear division capability, can correctly identify Complex, multi-category network data.Supervised learning algorithms need to select attributes, and then through the training set to build a classifier model, the use of this classification model for classification testing, record the results of classification.

Therefore, it is the best choice to apply traffic recognition based on machine learning in electric power information communication network. It is possible to use the traffic collected in previous networks as a training sample. Using supervised learning in machine learning is more effective than unsupervised learning More efficient flow of knowledge. Figure 1 is a flow chart of the classification method for supervised learning[4-5].

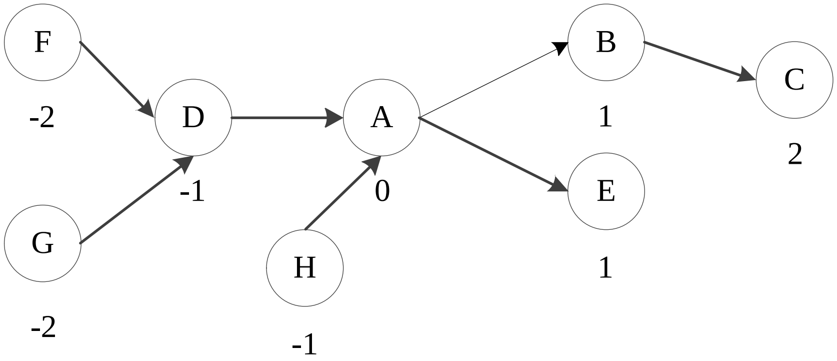

Figure.1. Flow chart of supervised learning classification method

Hidden Naive Bayes Algorithm

The Bayesian approach provides a probabilistic approach to reasoning. It assumes that the variables to be tested follow certain probability distributions and can reason from these probabilities and observed data to make optimal decisions. Bayesian methods can not only compute explicit hypothesis probabilities, but also provide an effective means for understanding most other methods.

Hidden Naive Bayesian Model on the basis of the Naive Bayes model, we introduce a hidden parent node for each attribute node to express the dependency relationship between the attribute and other attributes. It not only avoids the complicated calculation of the optimal Bayesian network structure, but also reduces the independence of the attributes, and makes full use of the interdependence between the attribute nodes. Hidden Bayesian Bayes contain three types of nodes: class nodes, attribute nodes, and hidden parent nodes.Respectively, with C, A, said the structure shown in Figure 2:

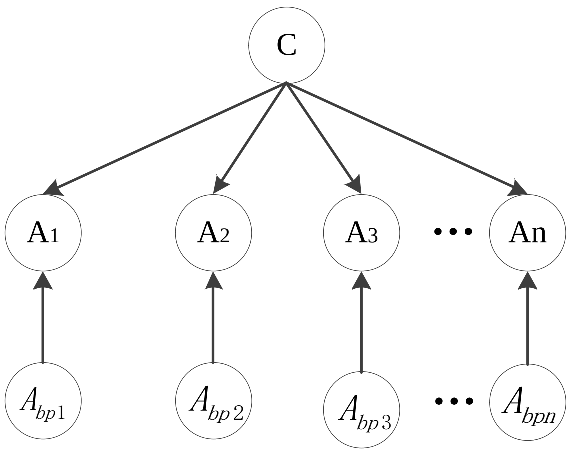

Fig.2. Block diagram of hidden Bayesian classifier

Its joint distribution is defined as follows:

$$P(A_1, A_2, \cdots, A_m \mid C) = P(C)\prod_{i=1}^{n} P(A_i \mid A_{hpi}, C) \tag{4-1}$$

among them:

$$P(A_i \mid A_{hpi}, c) = \sum_{j=1, j\neq i}^{n} W_{ij} * P(A_i \mid A_j, C) \tag{4-2}$$

From the above equation,Hides the parent node A_{hpi} and weighted sum of all other dependencies between individual attributes. In the HNB the dependencies between attributes are represented by the hidden parent of the attribute.Hiding the parent node determines the method determines the dependency between the attributes of the strength.In (4-2) use $P(A_i \mid A_j, C)$ to define hidden parent nodes.Compared with the TAN model, the dependence between attributes can be used more fully, which improves the classification accuracy and reduces the time complexity.

From (4-1)and(4-2) It is very important to determine the weight in the construction of the implicit Naive Bayesian model. We use W_{ij} conditional mutual information to define the weight:

$$W_{ij} = \frac{I_p(A_i; A_j \mid C)}{\sum_{j=1, j\neq i}^{n} I_p(A_i; A_j \mid C)} \tag{4-3}$$

This weight represents the size of the conditional dependency between two attributes, that is, the greater the degree of dependence, the greater the weight, and vice versa smaller. Among them, the attribute variables A_i and attribute variables between the conditions of mutual information, the specific calculation according to the formula

$$P(a_j \mid c) = \frac{\sum_{i=1}^{n} \delta(a_{ij}, a_j)\delta(c_i, c)}{\sum_{i=1}^{n} \delta(c_i, c)}.$$

$$I_p(A_i; A_j \mid C) = \sum_{a_i, a_j, c} P(a_i, a_j, c) \log \frac{P(a_i, a_j, c)P(c)}{P(a_i, c)P(a_j, c)} \tag{4-4}$$

In addition, a new probabilistic estimation method is proposed to calculate the probability P (c) and the conditional probability $P(a_i \mid a_j, c)$. The specific formula is as follows:

$$P(c) = \frac{F(c) + 1.0 / n_c}{n + 1.0} \tag{4-5}$$

$$P(a_i \mid a_j, c) = \frac{F(a_i, a_j, c) + 1.0 / n_i}{F(a_i, c) + 1.0} \tag{4-6}$$

Obviously, the above probability estimation method can be regarded as a special case of m-estimation. Ie, the m-estimate of the parameter m, which is called the constant of the equivalent sample size, is 1, and it is assumed that the prior estimate P of the probability is to be determined as the uniform distribution, the m-estimate becomes the probability estimate proposed in this paper Test method.

The implicit Naive Bayes model is an extension of the structure of Naive Bayes model. By introducing a hidden parent node for the Naive Bayesian model, the independence of the constraints is relaxed. The structure is simpler than the Naive Bayesian model Complex, similar to the tree-extended Bayesian model[6].

Traffic classification implementation steps

According to the traffic flow of electric power communication network, we mainly analyze the business in the integrated data network. Through the steps of collecting, identifying and classifying the traffic of the power network, the classification of power information network traffic is finished. The detailed steps are:

Step 1: Collect traffic. Using Wire shark integrated data network in a core router network port packet capture (packet).

Step 2: Construct the flow. To analyze the traffic packets collected in Step 1, a packet with the same quintuple (source IP, source port, destination IP, destination port, and transmission protocol) is constructed as a stream.

Step 3: Feature extraction and discretization. Based on the Bayesian algorithm based on the flow of electricity traffic classification is based on the flow of different attributes of the characteristics of classification, to do before the classification of the flow characteristics.

Step 4: Classification algorithm. Including the training process and the testing process. The labeled power network traffic is taken as the training data set, and the algorithm parameters are obtained through the training algorithm. Then the test data stream is processed to obtain the final classification result. The classification algorithm adopts the hidden Naive Bayes algorithm.

Summary

Through the simulation of the model, it can be seen that the machine learning method is the best choice in the electric power information communication network. Through training the collected traffic as a sample, using supervised learning in machine learning can be more effective than unsupervised learning Effective traffic identification. After the simulation of the classification algorithm, the hiding Naive Bayesian classification algorithm has a better performance than the traditional Naive Bayes algorithm, thus verifying the validity of the implicit Naive Bayesian classification algorithm.

Reference

[1] W. Yong hao, L. Cong, W. Jin, Z. Gui ping, “One New Research on Method of Intelligent Substation Network Traffic Prediction,” IEEE Fifth International Conference on Intelligent Systems Design and Engineering Applications (ISDEA), pp.683-687, June 2014.

[2] L. M. Dong, G. Yang, X. M. Cao, “Methods of Network Traffic Prediction,”Computer Engineering, August 2011, vol.37, no.16, pp.98-100.

[3] Z. Bin, L. Hong bing, “Network traffic prediction modeling and simulation,”Computer Simulation, 2011, vol.28, no.12, pp.84-87.

[4] D. Xiao bo, H. Min qiang, “Digital Substation Communication network performance simulation ,” Power System Technology, 2008, vol.9, no.17,pp.57-60.

[5] A.Q. Huang, M.L. Crow, G.T. Heydt, etc all. “The Future Renewable Electric Energy Delivery and Management (FREEDM) System: The Energy Internet,” Proceedings of the IEEE, vol. 99, pp.133–148, Jan.2011.

[6] Scott, Steven L., and Smyth, Padhraic. "The Markov Modulated Poisson Process and Markov Poisson Cascade with Applications to Web Traffic Modeling." Bayesian Statistics, pp.671-680, 2003.

[7] Victor S. Frost; Benjamin Melamed. "Traffic Modeling for Telecommunications Networks". IEEE Communications, vol.32 pp.70-81, 1994.