Research on Traffic Classification Technology of Electric Power Communication Network Based on Hidden Markov Model

Nanfang Li, Zongrong Li, Lei Zhao
National Network Qinghai Electric Power Company Electric Power Science Research Institute. 810008, China

Abstract. With the penetration of information and communication technology into power production and operation, the dependence of power grid on power communication network becomes stronger and stronger, and the reliability and transmission performance of communication network become the key factors affecting the safe and stable operation of power grid. Based on the analysis and comparison of the existing network traffic models, a hidden Markov model with both long-correlation and short-correlation characteristics is selected to fit and predict the traffic characteristics of the power communication network itself. The article focuses on the development and changes of the production and management of power grid companies, and analyzes the company's business structure and the bearing relationship of various types of power communication networks. The traffic classification technology of power communication network based on hidden Markov model proposed in this paper has higher accuracy and better stability, and the scores of each evaluation index are higher.

1. Introduction
In order to ensure the safe and stable operation of power system, power communication network emerges as the times require. The purpose of power communication network $W$ is to meet the needs of power production. It mainly transmits dispatch, protection, load control and enterprise management information. How to plan the cable configuration, route selection, bandwidth allocation, how to reduce the heavy losses caused by emergencies, how to effectively improve the speed and utilization of network operation, network traffic analysis and prediction technology plays a key role [1]. Hidden Markov Model (HMM) can provide a good probability description for the change of time series, but the static classification ability of HMM is weak. Therefore, in recent years, people have begun to combine ANN and HMM to construct some hybrid classifiers [2]. For the network traffic imbalance caused by different types of network service deployment modes and different service flows, it is required to support accurate prediction and intelligent control of network traffic, alleviate the unreasonable consumption of uncontrollable traffic to network resources, thereby improving the network. Consumption capacity [3].

In recent years, the forms and types of application layer protocols have become more complex than in the past, and the specifications of these new protocols are often not disclosed and do not comply with the default fixed port conventions. Therefore, how to calculate the large amount of network traffic and make the classification and classification method feasible is the primary problem faced by traffic classification technology [4].

At present, the State Grid has completed the construction of an IP-based network for information communication networks, and has realized the evolution of existing networks to IP networks. However, the research on intelligent pipelines of power communication networks is still in its infancy. Online
traffic monitoring of network traffic in real-time, rapid scanning of the entire network, timely response to early warning network anomalies or virus attacks, and real-time, accurate network traffic flow and traffic composition analysis for daily network maintenance [5-6]. There are many kinds of classifiers based on the hybrid of neural network and hidden Markov model. However, the MLP (Multilayer Perceptron) is directly embedded in the HMM for global optimization, which leads to a sharp increase in the training volume [7]. Active queue management mechanism is introduced to predict network traffic, so that it can predict the future trend of network traffic, and accordingly modify the probability of packet dropout, so that the queue length can be better controlled [8]. This paper studies the network traffic identification technology applied in the power information communication network environment, mainly introduces the basic concepts and development status of traffic classification technology. Through the perception and classification of network traffic, we can effectively understand the network traffic of all kinds of applications in power communication network, strengthen the perspective ability of network carrying business, and provide a solid foundation for more skillful realization of differentiated business guarantee; if we can predict traffic overload, we can analyze and solve problems before traffic overload occurs, we can significantly improve the availability of the network.

2. Methodology

With the development of power system, the comprehensive informatization of power system is becoming more and more urgent. The construction of power information communication network will be an important part of power system information construction. In practice, if the phenomenon is caused by one (or all) of several random factors, we can build a mixed model to study the complex phenomenon [9]. Practice shows that the finite mixture model is a powerful and effective tool for probabilistic analysis of univariate or multivariate data. Thus, different application types can be identified [10]. The process of classification is that the classifier located in the network finds the handshake signal packet when TCP in a TCP connection establishes the connection, and finds the destination port number from the SYN packet, and then judges the application type according to the port number. The process of UDP classification is similar to [11]. Through traffic classification, you can know the proportion of various types of network services, which will help to carry out traffic prediction work in the future, improve the utilization efficiency of existing network resources, and balance the distribution of network traffic. The scheduling data network and the integrated service data network are The two core components that make up the power data communication network. They are divided into two separate physically isolated networks by function and security level. At the beginning of setting an initial number of clusters, then by splitting and merging some branches to change the number of clusters, and finally converge to a reasonable value.

According to the statistics of the integrated service data network of the power communication network, the vast majority of the current integrated service data networkIt is carried by the SDH network, with 2M bandwidth as the basic expansion unit. The details are shown in Table 1.

Table.1 Traffic Analysis of Power Communication Integrated Service Data Network

| Business type       | Conference TV | remote viewing | Image monitoring | Production MIS | Central control information |
|---------------------|---------------|----------------|------------------|---------------|-----------------------------|
| Peak flow rate      | 1.8M          | 1.3M           | 1.7M             | 800K          | 780K                        |
| 2 M100 (%)          | 85            | 77             | 79               | 3.6           | 4.2                         |

ECN Operation Model This paper analyzes the ECN operation model from the perspective of energy flow. In ECN, there are two types of energy flows: power flow and communication flow, as shown in Figure 1. In EN, the trend changes with time through buses or lines. In contrast, since communication points send and receive messages on a regular basis, communication flows are sent on a discrete time basis.
The payload-based network traffic identification method is proposed to avoid the excessive dependence of the traditional classification method on the port number. Through the abnormality of the traffic volume of each network class, problems such as device faults and link congestion can be discovered in time, and the control capability and security of the network are improved. A resource sharing, safe and fast data path has been constructed in power plants, substations and dispatch centers. The real-time monitoring service has a data transmission period of seconds, is connection-oriented, and has a variable bit rate, and requires high service quality. At this time, the classification effect of this model is better than other models. But correspondingly, when a Gaussian distribution falls on the boundary of two areas where only slightly overlapping class data belongs, the performance of the generic component model is obviously not as good as that of the independent hybrid model. The advantage of this method is that the classification accuracy is very high. The disadvantage is also obvious: only those known non-encrypted traffic can be identified, but there is no way to deal with other unknown traffic; because different types of business usually adopt different protocols, each application layer protocol has its own characteristics, which can be specific port number, specific character or specific Bit sequence. And some services (such as work flow services, instant messaging, etc.) need to be visited and processed in collaboration, which leads to the sudden, decentralized and interactive characteristics of these services, and the transmission frequency is random.

Consider the cascading failure diagram of the interaction between ECNs as shown in Figure 2. When the cascading failure occurs in the EN, if the branch is reloaded, the fault packet is generated according to the control policy, and the corresponding communication point sends the data packet to the control center through the communication network. Thereafter, the Control Center sends the command back to the inside.
The power communication transmission delay varies according to the important level and requirements. It is roughly divided into 3 levels, as shown in Table 2 below.

| Level | Requirement | Transmission delay (s) |
|-------|-------------|-----------------------|
| A     | The requirement of real-time transmission is the highest and priority is given to transmission. | <4 |
| B     | High real-time transmission requirements, followed by priority | <11 |
| C     | There is no strict real-time requirement for transmission, and the priority is the lowest. | <55 |

In practical research, cumulative distribution function is generally used to represent the ratio of nodes whose degree is not less than R in the network. The cumulative degree distribution function can effectively control the problem that some boundary nodes in large-scale complex networks are susceptible to noise interference.

\[ S(r_k) = \sum_{r_i \neq r_j} w(r_i) D(r_k, r_i) \]  

(1)

Then the ratio of the number of edges M actually existing between the k nodes to the maximum number of edges that may exist, that is, the clustering coefficient Mk defined as the node r, is:

\[ M_k = c_k^d M_{k-1} \]  

(2)

The clustering coefficient N of the network represents the clustering degree of the whole network. It is defined as the arithmetic average of clustering coefficients of all nodes in the network. The expression is as follows:

\[ T_r = \frac{1}{N} \sum_{i=1}^{N} r_i r_i^T \]  

(3)

The average path length T of the network is defined as the average of the shortest paths between all pairs of nodes in the network.

\[ PR(k) = T_r(k) - V_r(k) \]  

(4)

In practical application, the concept of cumulative median distribution function is usually adopted. The formula for calculating cumulative median distribution function of nodes or edges is as follows:

\[ R_{ij} = M_j \sum_{i=1}^{N} \hat{R}_{ij} \]  

(5)

The following is a discussion of the relevance of VoIP traffic data. The Hurst coefficient of VoIP traffic data is calculated by the R/S method. If H>0.5, there is a long correlation between the time series of VoIP traffic. As shown in Table 3, after fractional difference, the value of D tends to zero, and the longer the sequence length, the more obvious the effect. Therefore, the fractional difference speech traffic has successfully removed the long correlation.

| Sequence length | Differential front | Post differential |
|-----------------|--------------------|-------------------|
|                 | H      | d   | H    | d    |
| 400             | 0.63   | 0.33| 0.52 | 0.05 |
| 800             | 0.81   | 0.34| 0.59 | 0.15 |
| 1200            | 0.81   | 0.36| 0.51 | 0.03 |
Due to the obvious shortcomings of port number-based classification method and payload-based classification method, a network traffic classification method based on host behavior is proposed. Classification methods based on load characteristics have high performance in accuracy, robustness and scalability. However, because of its need to detect the content of the package or even the whole stream, the computational overhead is huge, and there are serious performance bottlenecks. In order to realize the transmission of this kind of service, VPLS channel is constructed in MPLS network of integrated service data network, which is solved by two-layer VPN scheme. Because of the difference of protocol layer, the attributes of transmission stream will be partly affected. The differences in communication traffic between different components of the communication network determine their basic flow variation characteristics and different responses to various influencing factors, showing different response characteristics, with linear and periodic variations. Therefore, the service bandwidth requirement of the transport network should consider the networking channel requirements of the bearer service network. It is more accurate to use multi-fractal measures or dimensions to describe. The so-called multi-fractal refers to the uneven distribution of network traffic, and the traffic at different time scales exhibits different traffic characteristics. The performance requirements of the transmission are also classified into extremely fast, medium speed, low speed, etc. depending on the type of the message. The different communication links used make the transmission delays also different.

3. Result Analysis and Discussion
As the types of services on the network continue to increase and the traffic data continues to increase, it is too difficult to analyze these data directly. Data mining technology can extract hidden information with specific use from a large amount of data. Due to the fixed cycle of power production and operation, such as equipment online monitoring, business hall and construction site video collection, network video conferencing, etc., the power video service has obvious periodicity. Traffic analysis of power communication services, combined with different business characteristics, can select different business models for approximation and simulation. With the need to support the operation of the power system and the construction of smart cities, the implementation of the network's traffic time-sharing price affects the distribution of the load, which is a problem that needs to be considered in the future after the mature development of the power market. Especially the prediction of peak flow is an important content of flow forecasting. In practical systems, it is unlikely to occur. Its introduction only serves as a reference object. That is, the larger the deviation of observation behavior from the model, the greater the possibility that the observation behavior is an intrusion behavior. The structural risk minimization principle based on statistical learning theory not only minimizes the empirical risk, but also minimizes the empirical risk and confidence range by finding the maximum classification interval, thus avoiding the phenomenon of over-fitting.

Based on the network model, the statistical data of cumulative degree distribution of nodes and cumulative median distribution of networks are calculated, and the statistical curves are drawn in linear coordinates, single logarithmic coordinates and double logarithmic coordinates respectively. According to the fitting results, the nodal degree distribution is analyzed as follows:

Fitted to a straight line in linear coordinates, the cumulative distribution function is:

\[ B(\tilde{X}) = \prod_{i=1}^{m} \left( f_i(\tilde{X}) - f_i(\tilde{X}_w) \right) \]  \hspace{1cm} (6)

The degree distribution of the network obeys Poisson distribution, and the network is uniform, which is similar to stochastic network. The cumulative distribution function is as follows:

\[ B(\tilde{X}) = \prod_{i=1}^{m} \left( f_i(\tilde{X}) - f_i(\tilde{X}_w) \right)^{w_i} \]  \hspace{1cm} (7)

Fitted to a straight line in double logarithmic coordinates, the cumulative distribution function is:
In order to ensure the quality of service transmitted by the substation communication network, under the premise of a certain network bandwidth, the substation communication network divides the information transmission priority according to the information update time interval requirements of various applications. It is usually a two-layer pre-transmission network. It can be regarded as a feed forward neural network composed of a hidden layer. The role of the hidden layer is equivalent to transforming the input pattern, transforming the input data of the low-dimensional mode into a high-dimensional space, so that the classification problem becomes more likely to be linearly separable. Assuming that the observations are independent of each other, a distribution function can be used to represent their observations. If the model assumption is correct and the outliers in the observations have less influence, the entire model can be inferred from the estimates of several unknowns to obtain a distribution representation of the observed data. All training samples need to be trained. The training samples are too many and the training speed is slow. The training data is used to build a classification model, and then a classifier is generated based on the model. Then the classifier is applied to classify the unknown data sets. In practical applications, channel bandwidth allocation based on single particle is often considered for the same kind of service with the same cross section. In addition, precise traffic model and prediction methods can also promote the design of new generation network protocols, network management and diagnosis, design of High-performance Routers and other network hardware devices, and improve the quality of service of the network.

According to the analysis of static topological characteristics, network G has 99 nodes and 130 edges, which belongs to the small world network. In this paper, the random fault simulation is performed for the nodes in G, that is, the Nr nodes are randomly removed from the network G, and the edges connected to the Nr are simultaneously invalidated. At this time, the subgraphs with the number of nodes N-Nr can be obtained. Figure 3 below shows the network connectivity robustness curve under point random failure.

\[ B(\vec{X}) = \prod_{j=1}^{n} u_j(x_j)^{\nu_j} \]  

(8)

The performance function curve in Figure 4 below shows statistical characteristics similar to the connectivity robustness. As the number of random exit nodes increases, the performance function of the network decreases gradually, but the change is generally flat.
Fig. 4 Network performance function under point random failure

Through the analysis of the service types and transmission characteristics of the power communication supporting the distribution network, it is found that the service flows carried by different subnets partially overlap with the traffic fractal level. Therefore, the sampling and shaping analysis of the same data flow of the existing actual power communication can still be adopted. Complete the model building. When the assumed distribution model deviates significantly from the actual distribution model, the estimated value will not be destructively affected, and the goodness can still be guaranteed. In general, the higher the accuracy of the service that conforms to the larger traffic, the lower the accuracy of the traffic with lower traffic. It can be seen from the confusion matrix that there are still cases where the classification result is misled by the large traffic service. However, since the classification accuracy of each service is high, this situation has little effect on the overall classification performance. The classification method based on payload has a high accuracy, but the relative cost is large, and the need to interpret data packets will cause privacy violation and security problems. For the network environment with high security requirements such as power information communication network, this method is not appropriate. The advantage of Markov model is that it is easy to deal with time series with queuing performance. The disadvantage of Markov model is that the complexity of analytical solution increases with the increase of analytical parameters. According to the total traffic carried by the communication network at this level, according to the overall structure of the current communication network, the traffic is divided into different integrated data networks, scheduling data networks, transmission networks, and the related planning and decision-making conclusions of optimization, transformation and expansion of different systems are formed.

4. Conclusion

Network traffic forecasting is the key issue of network service management, congestion control and flow control. With the construction of a strong smart grid, power communication network has undergone profound changes. As the strong support of power grid informatization, automation and interaction, the reliability and transmission performance of power communication network become the key factors affecting the safe and stable operation of power grid. The application of network traffic identification technology in power information communication network can improve the visualization ability of network services, and is conducive to more effective allocation of resources and security protection. This paper analyses the existing common traffic classification and recognition technologies: traffic identification based on port mapping, net load characteristics, traffic statistics and machine learning, and then focuses on the hidden Markov model algorithm, which aggregates the actual scenario requirements of power information communication network. The actual network traffic characteristics and the metrics that characterize these traffic characteristics are the basis and basis for traffic modeling. Currently, the most important statistical characteristics of network traffic are self-similarity on large time scales and multi-fractal on small time scales. Considering the difficulty of data processing, the processing and classification of traffic data in this paper are offline. However, in some cases, traffic classification has
strong requirements for real-time performance, and offline traffic classification systems cannot achieve some practical applications. Requirements. Therefore, the next step can improve the time performance of the classification system and further enhance the real-time operation of the classification system.

References

[1] Zhang Y, Cai Z, Li X, et al. Analytical Modeling of Traffic Flow in the Substation Communication Network[J]. IEEE Transactions on Power Delivery, 2015, 30(5):2119-2127.

[2] Wu P, Zhang G, Wu J M, et al. Research on the Application of Software Defined Network in Electric Power Data Communication Network[J]. Advanced Materials Research, 2014, 971-973:1735-1738.

[3] Han S, Peng Z, Wang S. The maximum flow problem of uncertain network[J]. Information Sciences, 2014, 265:167-175.

[4] Morstyn T, Hredzak B, Agelidis V. Network Topology Independent Multi-Agent Dynamic Optimal Power Flow for Microgrids with Distributed Energy Storage Systems[J]. IEEE Transactions on Smart Grid, 2016:1-1.

[5] Luo C, Guo S, Guo S, et al. Green Communication in Energy Renewable Wireless Mesh Networks: Routing, Rate Control, and Power Allocation[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(12):3211-3220.

[6] Qian J, Zhou Z, Gu T, et al. Optimal Reconfiguration of High-Performance VLSI Subarrays with Network Flow[J]. IEEE Transactions on Parallel and Distributed Systems, 2016:1-1.

[7] Ruiwen H, Jianhua D, Lai LL. Reliability Evaluation of Communication-Constrained Protection Systems using Stochastic-flow Network Models[J]. IEEE Transactions on Smart Grid, 2017:1-1.

[8] Low S H. Convex Relaxation of Optimal Power Flow, Part II: Exactness[J]. IEEE Transactions on Control of Network Systems, 2014, 1(2):177-189.

[9] Kong P Y. Optimal Configuration of Interdependence between Communication Network and Power Grid[J]. IEEE Transactions on Industrial Informatics, 2019, PP(99):1-1.

[10] Wang X Z, Liu Z G. Network Seepage Flow Research Based on Complex Network[J]. Applied Mechanics and Materials, 2014, 608-609:392-396.

[11] Wang X, Pi Y, Mao W, et al. Network Coordinated Power Point Tracking for Grid-Connected Photovoltaic Systems[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(7):1425-1440.