A Novel Method for Predicting Network Traffic Based on Maximum Entropy Principle

Jingyu Wang and Yang Zhao[2*]

[1]*Science College, Inner Mongolia University of Technology, Hohhot 010051, Inner Mongolia, China*
[2]*Department of Electronic and Information Technology, Jiangmen Polytechnic, Jiangmen 529090, China Jiangmen Polytechnic*
corresponding author e-mail: zhaoyang19781023@gmail.com

Abstract

The network of application service is becoming more and more increasingly complex, with the development of network communication technology, which puts forward higher requirements on network behavior characteristics, the network management and traffic control, therefore, network traffic analysis and prediction is more and more important significance. This paper presents a novel network traffic prediction model, which is based on maximum entropy algorithm. The simulation results show that the algorithm proposed in this paper has higher prediction accuracy than the traditional methods, and improves the prediction accuracy of network traffic.

Keywords: Network traffic, Maximum entropy principle, Prediction

1. Introduction

Nowadays, the network has become very large and complex, with the size of rapid growth, due to the fast development of computer network. The web-based applications in networking environment is more and more abundant, which means that the greater chance of network encountering service problems, and the performance of the network more vulnerable. Therefore, real-time network monitoring is a fundamental part of the network management, network monitoring purpose is to collect information about network status and behavior. Network traffic is a very important aspect of network monitoring, and accurate flow model can gain real network traffic statistical characteristic. However, if you can't get a suitable model to describe the statistical characteristics of the actual traffic, the network performance will become very poor and difficult to control, because they either overestimated or underestimated the network performance.

Network traffic accurate prediction has very important value for the accurate network performance analysis, for congestion management and traffic equilibrium, improve service quality and so on all. The network traffic prediction can provide technical support for network administrators, help them adjust the network control strategy, improve the quality of network service, so the network traffic modeling and forecasting has been the focus of the research on network [1]. As we all know that, there are a lot of uncertainty in network traffic, if simply with historical data fitting degree as the weight of the single forecasting model are combined, the precision of the comprehensive prediction result is not ideal.

In this paper, the principle of maximum entropy in information theory is adopted, to build a comprehensive model of network traffic prediction. It is helpful to deal with a lot of uncertain information in network traffic prediction, so as to further improve the accuracy of traffic forecast. The rest of this paper is organized as follows. In Section II, network traffic prediction related works are introduced in detail. Section III gives the

methodology proposed in this paper, which includes the network traffic prediction principle and maximum entropy principle (MEP). Experiments and analysis are given in Section IV. Finally, in Section IV, we summarize our work and give the direction of further job.

2. Related Works

Recently, network traffic prediction problem has attracted many attentions of the scholars at home and abroad, and an extensive and in-depth research, obtain good effect. Network traffic prediction method can generally be divided into traditional Markov method, time series analysis method and neural network method, etc. The principle of the autoregressive or autoregressive moving average forecasting model of the algorithm is simple, and the prediction accuracy for short-term forecasting is high, but the result of long-term forecast is bad. Literatures [16, 17] proposed forecasting model based on autoregressive or autoregressive moving average (AR), paper [18] proposed autoregressive integrated moving average (ARMA). In paper [19], the researchers assumed that the network traffic appeared a smooth change of data, and proposed AR model to deal with the network traffic prediction. However, network traffic is the result of comprehensive shaping of many factors, randomness, time-varying non-stationary characteristics, therefore, the traditional model is difficult to do an accurately long-term projections for network flow.

Some scholars started to import the theory of nonlinear modeling and forecasting network traffic, such as support vector machine (SVM), grey theory and neural network prediction model of network traffic and so on. Papers [2-3] established the network traffic reliability prediction method based on the neural network model. In papers [4-7], neural network was used to study the nonparametric, nonlinear classification and prediction problems of network traffic. Paper [8] presents a time series forecasting algorithm, which adopts neural network puts forward a nonlinear mapping relationship based on the input set and expectation model.

Due to network traffic is an affected by multiple factors combination of complex system, a single prediction algorithm can only describe the part or section information, it is difficult to comprehensively and accurately predict its change rule. Some scholars have established the combination of network traffic prediction model based on the famous M-competition theory, the results showed that combination model can forecast larger limit use a variety of sample information, system for more than a single prediction model considering problems, more comprehensive [9-10]. Due to the network traffic has multi-scale features at the same time, some scholars put forward a kind of wavelet analysis and integration of ARIMA network traffic prediction model, has obtained the good prediction effect [11]. However, this model adopts the single ARIMA model for high frequency and low frequency part modeling, and then it is difficult to accurately describe the network traffic complex changes, which means that, the forecast needs further improve [12-13].

At present, the most concern is how to further improve the accuracy of the network traffic prediction, in view of the above algorithm in the literature. This paper proposes a new network traffic prediction model. It is through the principle of maximum entropy estimation which is the record of model according to the historical forecasting error of various trade-offs, the prediction model. The simulation experiment result is acquired through practice to, in the actual network, the projections for traffic data obtained from experiments show that compared with a traditional centralized network traffic prediction algorithm, the prediction accuracy of this method are much higher.

3. Methodology

3.1. Network Traffic Prediction Principle

Network flow changes by the people behavior over the weekend and holidays and other factors affect. Besides that, the economy has a long-term influence on changes in network traffic, and then people behavior of network traffic is cyclical, weekends and holidays lead to changes random and sudden. According to the above factors, the total capacity of network flow at some point X (t) could be described as available the following model

$$X(t) = N(t) + W(t) + U(t) \quad (1)$$

Where, N (t) is the basic normal flow of network traffic in a moment; W (t) shows the effect of people behavior of network traffic; U (t) is the influence of the network traffic over the weekend and holidays.

From equation (1), the network traffic is a complex nonlinear system, which indicates that, it can't fully reflect the variety regulation of the network traffic by using a single model and it is difficult to obtain a higher prediction precision. Fig.1 shows the flow chart of predicting network traffic.
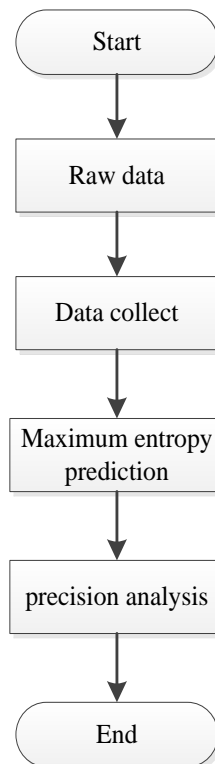

Figure 1. Flowchart of Network Traffic Prediction Principle

3.2 Maximum Entropy Principle

In 1948, C.E. Shannon proposed the concept of information entropy [13], the entropy as a random uncertainty event or a measure of the amount of information, which laid the scientific foundation of the modern information theory. After that, E.T. Jayne [14, 15] presented a criterion in 1957. It should have maximum entropy to comply with all the known information when reasoning according to some information. Furthermore, the criterion was called maximum entropy principle. The principle of maximum entropy can

be expressed as the following optimization problem in math:

$$\max H(X) = -\sum_{i=1}^{n} p_i \ln p_i \qquad (2)$$

$$st \sum_{i=1}^{n} f_k(x_i) p_i = F_k \quad k = 1, 2, \cdots, m (m < n) \qquad (3)$$

$$\sum_{i=1}^{n} p_i = 1 \quad P_i \geq 0 \qquad (4)$$

Where, $H(x)$ is the entropy of random variable X; P_i is the probability of X when x_i appears; $f_k(x_i)$ is a function of X; F_k is the mean value of some $f_k(x_i)$. Here, we import two lagrangian multiplier variables, namelay, α and β_k, in order to calculate the maximum of entropy in the constraint condition equations (3, 4).

$$E = H + \alpha \left(\sum_{i=1}^{n} P_i - 1 \right) + \beta \left(\sum_{i=1}^{n} f_k(x_i) - F_k \right) \qquad (5)$$

The equation (5) could be changed into as equation(6):

$$H \leq \sum_{i=1}^{n} p_i \left\{ \frac{1}{p_i} \exp\left[-a - \sum_{k=1}^{m} \beta_k f_k(x_i) \right] - 1 \right\}$$
$$\alpha + \sum_{k=1}^{m} \beta_k F_k \qquad (6)$$

It must take the equal sign in order to get the maximum value of variable H, and then we could obtain the p_i should be satisfied as follows:

$$p_i = \exp\left[-\alpha - \sum_{k=1}^{m} \beta_k f_k(x_i) \right] \quad i = 1, 2, \cdots, n \qquad (7)$$

The equation is the relevant distribution of entropy when constraint condition equations (3) and (4) are satisfied at the same time.

4. Experimental Results and Analyses

4.1. Network Traffic Data Collect

In general, the more data, the result of learning and training he can correctly reflect the relationship between input and output, that is to say, the higher the accuracy of the prediction. In practice, impossible unlimited increase the sample data, this kind of circumstance should try to select a representative sample, namely on system performance score good sample. In this paper, the network traffic experiment data is from the Bell-core laboratory (BC-pAug89), which collected 100 million data packet in 3 142 seconds when 8, 29, 1989.

We have done data preprocessing before experiment, and took 1s as the sampling period, got 3 142 sample points, which thus ensuring in did not lose the original characteristics of the data and improved the efficiency of the algorithm. The network traffic data is as shown in Figure 2. It can be seen that, the original network traffic data is non-linear.
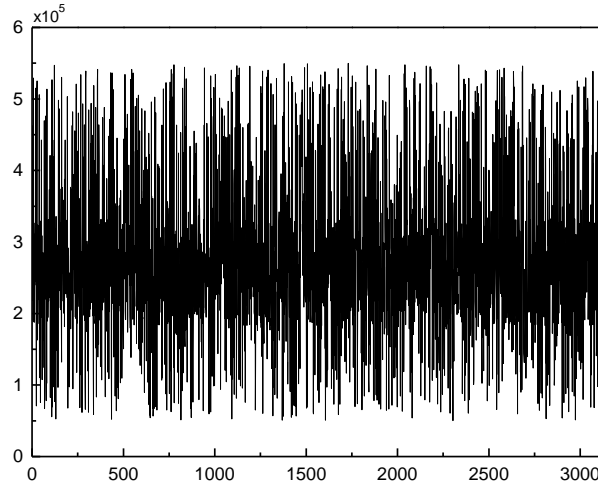
Figure 2. Flowchart of Network Traffic Prediction Principle

4.2 Network Traffic Prediction Based on MEP

Actually, there are many uncertain factors influencing network traffic, which has brought fuzzy characteristics of it. It will inevitably cause greater error, if we still adopt single forecasting model to forecast. Aiming at this situation, this paper presents a network traffic prediction model, which is established on the principle of maximum entropy algorithm. The constraint condition is according to the previous record of forecasting error of various prediction models for the prediction results of normalized processing, and in all kinds of prediction model prediction results and their distribution of maximum entropy approach to predict the results similar to the results of the real.

This paper choose history day for predicting network traffic, which can adopt the single forecasting model to forecast, in order to obtain prediction statistics. Besides that, we also adopted several historic day for the virtual prediction, and then according to the actual results to obtain characteristic value.

Assume that, N is the reference day of the predicting process, and T is the flow mapping to the predicting time point of every reference day. Furthermore, set y_{nt} (t = 1,2, …… , T; n = 1, 2, …… , N), and y_{ntr} (r = 1,2, …… , R) is the prediction value of network traffic.

Suppose the second order center distance [20] of prediction of network traffic is as follows:

$$E = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{y_{tn} - \hat{y}_{tnr}}{\hat{y}_{tnr}} \right)^2 \tag{8}$$

The average value of prediction could be expressed as:

$$\delta_{tr} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{y}_{tnr} - y_{tn}}{y_{tn}} \right)^2 \tag{9}$$

Now, we could build the prediction model according to the principle of maximum entropy.

$$\max H(x) = -\int p(x) \ln p(x) dx \tag{10}$$

$$st \begin{cases} \int p(x) \left[\dfrac{x - \hat{y}_{tr}}{\hat{y}_{tr}} \right] dx = E \\ \int p(x) dx = 1 \end{cases} \quad (11)$$

Now, we could solve the parameter value of α and β according to the equations of (9-11). Finally, the probability density function of network traffic could be expressed as in equation(12), furthermore, we could get the prediction value of network traffic of the day by repeating the above steps,

$$p(x) = \exp \left[-\alpha - \sum_{r=1}^{R} \beta_r \left(\frac{x - \hat{y}_{tr}}{\hat{y}_{tr}} \right)^2 \right] \quad (12)$$

4.3 Prediction Results and Analysis

Figure 3, Figure 4 and Figure 5 show the network traffic prediction result of MEP algorithm. Fig.3 shows the overall prediction result of network traffic, and we intercept 50 points to view in detail, as shown in Figure 4. From Figure 4, it is clearly seen that, the predicted results of the MEP are very close to the actual network flow. The MEP method could describe the relevant performances of network traffic flow, such as random, time-varying non-stationary changing trends.
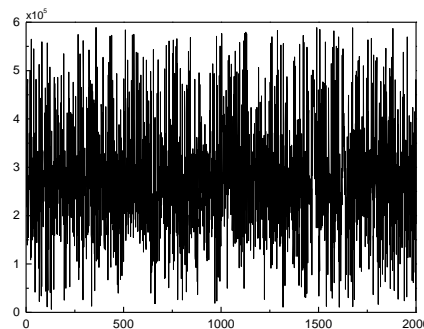

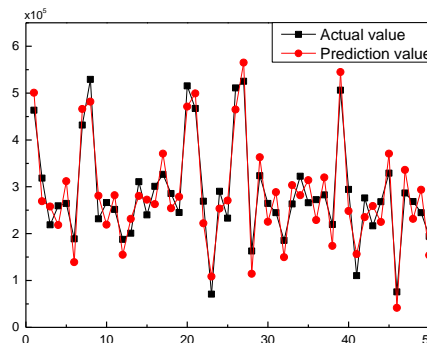
Figure 3. Overall Preidction of Network Traffic

Figure 4. Local Prdiction of Network Traffic

Fig.5 shows the residual plot of MEP predication results, it is can be seen that, most of the points are distributed around zero, only few point reach the extremum.
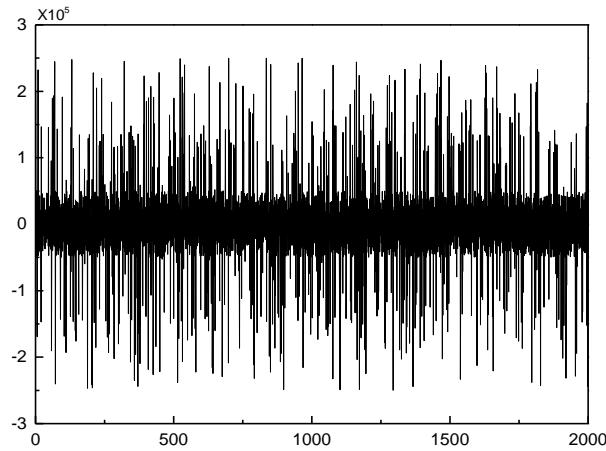

Figure 5. Residual Plot of Network Prediction Algorithm

In order to verify the prediction performance of proposed MEP method, we also import the other forecasting method, namely, statistical approach based on auto regression moving average model (ARMA). Fig.6 shows the prediction comparison of MEP and ARMA algorithm. It is can be clearly seen that, the corresponding relative prediction error of maximum entropy principle method better than network traffic prediction model based on autoregressive average ARMA from the plot. Furthermore, we can also obtain the result that, the MEP method has a good stability, because the original prediction accuracy measurement based on the maximum entropy is not increasing as the prediction step length and subsequently
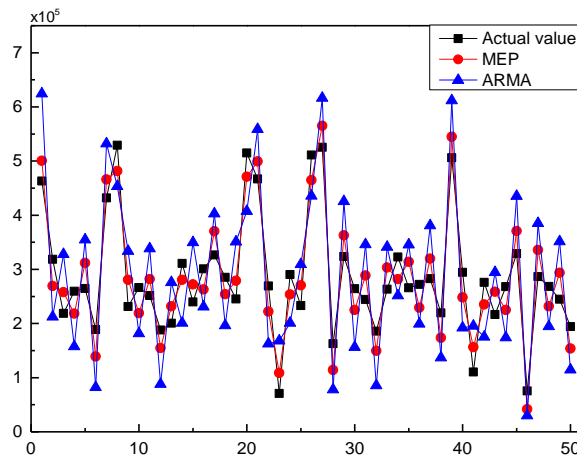

Figure 6. Performance Comparison of Different Network Peridction Algorithms

Here we use MAE and RMSE to make a quantitative evaluation of the three algorithms based on equations (13-14). The MAE and RMSE values of MEP method are 0.9555 and 11.42, respectively. On the other hand, the MAE and RMSE values of ARMA method are bigger than MEP, which means that the prediction accuracy of MEP method is better than other forecast algorithm.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \qquad (13)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} |x_i - y_i|^2} \qquad (14)$$

5. Conclusion

Optimize network control strategy is very meaningful to study the characteristics of network traffic, and then to establish the forecast model of conform to the actual flow characteristics, which not only can explore the new way to network management, but also for the construction of a new generation of information network and management provides a theoretical basis and technical support.

Due to the high degree of correlation, nonlinear and chaotic characteristics of network traffic flow, it is very difficult to establish a very accurate mathematical model and as a result, making the prediction of network traffic is more difficult. In order to solve this problem, this paper proposes the principle of maximum entropy algorithm to forecast the network traffic, which has higher prediction accuracy than the exiting prediction methods.

As future work, it would be of importance to investigate the effectiveness of MEP algorithm and make it suitable for real-time condition. Besides that, to model network traffic as combination format, and introduce some new techniques to the flow model and continuous improvement may be still hot spot in our future research.

Acknowledgments

References

[1] T.T Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning", IEEE Communications Surveys and Tutorials, vol. 10, no. 4, (**2008**), pp. 56-76.

[2] S. Li, Q. Yin and P. Guo, "A hierarchical mixture of software reliability model for prediction",. Applied Mathematics and Computation, vol. 185, no. 2, (**2007**), pp. 1120-1130.

[3] Q. Hu, M. Xie and S. Ng, "Robust recurrent neural network modeling for software fault detection and correct prediction", Reliability Engineering and System Safety, vol. 92, no. 3, (**2007**), pp. 332-340.

[4] B. Ozkaya, A. Demir and M.S. Bilgili, "Neural network prediction model for the methane fraction in biogas from field-scale landfill bioreactors", Environmental Modeling & Software, vol. 22, no. 6, (**2007**), pp. 815-822.

[5] H. Farsi and F. Gobal, "Artificial neural network simulator for supercapacitor performance prediction", Computational Materials Science, vol. 39, no. 3, (**2007**), pp. 678-683.

[6] K. Byungwhan, and B. Jungki, "Prediction of plasma processes using neural network and genetic algorithm", Solid-State Electronics, vol. 49, (**2005**), pp. 1576-1580.

[7] L. Yang, C.W. Dawson and M.R. Brown, "Neural network and GA approaches for dwelling fire occurrence prediction", Knowledge-Based Systems, vol. 19, no. 4, (**2006**), pp. 213-219.

[8] G.C. Liao, "Using immune genetic algorithm based hybrid techniques for short-term unit commitment problem", Electrical Engineering, vol. 87, (**2005**), pp. 267-279.

[9] B.R. Chang and H.F. Tsai, "Improving network traffic analysis by foreseeing data-packet-flow with hybrid fuzzy based model prediction", Expert Systems with Applications, vol. 36, no. 3 (**2009**), pp. 6960-6965.

[10] A. Este, F. Gringoli and L. Salgarelli, "Support vector machines for TCP traffic classification Computer Networks", vol. 53, no. 14, (**2009**), pp. 2476-2490.

[11] Y. Chen and B. Yang, "Small-time scale network traffic prediction based on flexible neural tree", Applied Soft Computing Journal, vol. 12, no. 1, (**2012**), pp. 274-279.

[12] D.C. Park, "Prediction of network traffic using dynamic bilinear recurrent neural network", Proceedings of the 5th International Conference on Natural Computation, (**2009,** pp. 419-423, Tianjin, China.

[13] C.E. Shannon, "A Mathematical Theory of Communication", The Bell System Technical Journal, vol. 27, (**1948**), pp. 379-423, 623-656.

[14] E.T. Jaynes, "Information theory and statistical mechanics", Physical Review, vol. 106, no. 4, (**1957**), pp. 620–630.

[15] E.T. Jaynes, "Information theory and statistical mechanics. II", Physical Review, vol. 108, no. 2, (**1957**), pp. 171–190.

[16] J. Jiallg and S. Papavassiliou, "Enhancing network traffic prediction and anomaly detection via statistical network traffic separation and combination strategies", Computer Communications, vol. 29, no. 10, (**2006**), pp. 1627-1638.

[17] B.X. Zou and Q. Liu, "Network traffic prediction based on ARMA model", Journal of Computer Research and Development, vol.39, no. 12, (**2002**), pp. 1645-1652.

[18] Z.L. Deng, "Optimal Estimation Theory with Applications—Modeling, Filtering and Information Fusion Estimation", HARBIN Institute of Technology Press, (**2005**).

[19] M. Corradi, G. Garroppo and S. Giordano, "Analysis of F-ARIMA processes in the modeling of broadband traffic", Proceedings of the International Conference on Communications, (**2001**), pp. 964-968, Helsinki.

[20] L.S. Wei, Y.K. Ching, C.S. Chieh and L.Z. Jung, "Particle swarm optimization for parameter determination and feature selection of support vector machines", Expert Systems with Applications, vol. 35, no. 4, (**2008**), pp. 1817－1824.