Hindawi
Mobile Information Systems
Volume 2022, Article ID 4511409, 7 pages
https://doi.org/10.1155/2022/4511409

Research Article

A Network Traffic Network Prediction Model with K-Means Optimization Algorithm

Zhen Wei and Jingwei Sun

School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, Anhui, China

Correspondence should be addressed to Jingwei Sun; 2014010059@mail.hfut.edu.cn

Received 17 June 2022; Accepted 16 July 2022; Published 11 August 2022

Academic Editor: Le Sun

Data storage and computation on cloud servers can handle many gigabytes of data, but their network traffic will also be heavy. Researchers have developed several models for predicting network traffic in order to reduce the communication pressure on cloud servers. However, the existing models are not accurate enough to be applied to the cloud server. To deal with this problem, a network traffic prediction model (NTPM) with the K-means optimization algorithm is presented in this paper, which clusters network traffic data from cloud servers by the K-means optimization algorithm, and then SVM is used to train the model. Our study shows that compared with a recent NTPM that predicted network traffic accurately, the proposed model provides better network traffic prediction and is better suited to cloud servers.

1. Introduction

With cutting-edge computing and storage capabilities, cloud servers are a new type of network infrastructure. In recent years, for industries ranging from big data to cloud computing and Internet of things, it plays a critical role [1]. The increased number of applications on cloud server results in an increased number of users, and the related network traffic is also growing, which brings great communication pressure to the cloud server. To deal with this problem, a feasible way is to predict network traffic. By analyzing a large number of network traffic data, an efficient NTPM and algorithm are designed. Thus, a practical network traffic scheduling scheme is constructed [2]. Using the model to predict network traffic also has many advantages. For example, according to the prediction of the model, it can help cloud security applications detect malicious network traffic attacks from hackers and ensure that cloud servers provide more stable services for related applications. At present, there are two ways to design the network traffic model. The first is to design the corresponding algorithm according to the traditional statistical theory. However, this method cannot achieve high accuracy and is not suitable for cloud servers. The second way is to use computer-aided NTPM. By introducing big data and artificial intelligence technology, it can achieve high prediction accuracy and is also very suitable for running on cloud servers [3].

In computer-aided NTPM, there are usually two design methods. The first is to design linear models, such as autoregressive model and differential autoregressive moving average model. However, with the network communication of cloud server becoming more and more complex, these linear models cannot achieve accurate prediction. The second way is to design nonlinear models. In addition to SVM and neural networks (NN), machine learning technologies are increasingly used to improve network traffic prediction accuracy. Liu et al. [4] formulated a model using the chaos theory and SVM to predict network traffic in short-term. By introducing CTSA, the time series is modeled and predicted, and together, CTSA and SVM are used. In addition, using MLP, MLPWD, and SVM, Nikravesh et al. [5] presented an NTPM. Recently, Dong [6] proposed a NTPM based on the multiclass SVM algorithm and active learning, which can address the problem of imbalance in the identification of NTPM. Dong's [6] NTPM uses machine learning technology, and compared to the linear model, the

model increases prediction accuracy to some extent. However, NN are more likely to be trapped in local minima and decreased the accuracy.

To handle these problems, a new NTPM with K-means optimization algorithm are formulated in this paper. In the new NTPM, firstly, K-means optimization algorithm is used to cluster the historical traffic data in the data set, which helps to deal with the problem that the prediction is easy to focus on the minimum point. Secondly, the prediction regression model is established by using SVM, and the specific training of the model uses literature [7] where a cloud server network traffic data set in. Finally, through the analysis of literature and simulation results, compared with the scheme in [8], our scheme has advantages in several indexes such as average absolute percentage error (MAPE) and goodness of fit (R).

2. Related Work

SVM is able to help classification research, which plays a good auxiliary role in the classification of all kinds of data. Laref et al. [9] proposed a parallel global optimization model to optimize the hyperparameters of support vector machine (SVM) regression, which can be trained to provide accurate water demand prediction in a short period, which proves that its selection has advantages in generalization. In addition, by continuous innovation of technology, SVM has been applied to regression and prediction tasks and achieved good results. The principle of SVM is different from that of SVM. In the problem of regression fitting, the final realization is to linearize the nonlinear problem, in order to achieve the desired goal. The specific processing method can be described as follows: nonlinear transformation maps the input data with high dimension spaces. A kernel function theory replaces the point product operation in the high-dimensional space when solving the optimal decision-making function. A series of complex calculation steps are avoided effectively. In order to solve the nonlinear regression problem in the sample space, linear regression is applied in the feature space. Duan [10] proposed an NTPM with SVM technology. By using the groups to select the corresponding learning parameters, the particle swarm optimization algorithm (PSOA) is introduced to Duan et al.'s NTPM, which avoids the local optimal solution and provides better generalization ability.

Tang [11] applied SVM to the simple classification of network traffic prediction, where to improve prediction accuracy, they propose applying a hybrid model of denoising and SVM. A hybrid NTPM with SVM technology was presented by Chen et al. [12]. In this model, the empirical mode decomposition of SVM is carried out to reduce a number of noise signal. The parameters of SVM in Chen et al.'s NTPM are integrated with PSOA. An NN model with quantum genetic technology was proposed by Tian et al. [13]. An optimization NTPM with the efficiency of the quantum genetic algorithm is designed to optimize traffic data based on the chaotic characteristics of traffic data, which is then verified and can predict the network traffic more accurately. An NTPM with SVM technology as well as fuzzy analytic hierarchy process (FAHP) was proposed by Wang et al. [14]. With FAHP, the parameters of Wang et al.'s NTPM are constructed. Wang et al.'s NTOM is able to track the traffic change lines, and it can also make accurate predictions under the conditions of small fluctuations in prediction error, improving the accuracy and stability of the model. Tian and Li [15] formulated an NTPM for local area network. In Tian and Li NTPM, the parameters of SVM are optimized by using the improved free search algorithm, which achieves faster convergence speed and better fitness value, improves the prediction performance, and does not increase the complexity of the algorithm. A model for small time scale traffic prediction was presented by Meng et al. [16]. For the SVM regression model, the nearest number of adjacent points is selected using the BIC neighbor selection method and achieves effective network traffic prediction. Recently, an NTPM with optimization algorithm was presented by Bishnu and Bhattacherjee [17]. In Bishnu et al.'s NTPM, a program predicts modules, and methods were formulated. An NTPM with combination algorithm and using long-term k-means memory neural network to construct the NTPM for short-term 2019 coronavirus disease cases was presented by Vadyala et al. [18]. For comparing the prediction to the past few days, Vadyala et al.'s model uses a weighted k-means algorithm based on extreme gradient enhancement.

3. NTPM with K-Means Optimization Algorithm

An efficient NTPM with K-means optimization algorithm as well as SVM is described in this section. For constructing a specific model, we process the data as follows:

In the actual research results, there are many factors that cause data problems, such as too much redundant information, the difference of value range between attributes, the accuracy of information provided, and manual input error. Therefore, it is necessary to deal with the original data sets in a scientific, reasonable, and all-round way before studying and digging data. In order to explore future data more efficiently. The final research of mining data depends on the quality or level of preprocessing. If the preprocessing is good, it can reduce the error rate of searching data. If the preprocessing is not good enough, a lot of noise will be generated in data retrieval, and the prediction accuracy will be reduced. So, preprocessing data set is an essential step before studying and analyzing data. The preprocessing of this paper covers data feature extraction and cleaning. Verifying and checking the data after it has been gathered is known as "data cleaning." The sample data of mobile network traffic includes abnormal value and invalid value (null value, no corresponding value in dictionary, etc.). The integrity of the data can be guaranteed directly. Then, extract and arrange the feature data, and combine the traditional general missing value meaning that the value of one or some attributes in the existing data set is essentially incomplete, such as missing data and noise data. In order to find incomplete data in the dataset, we fill in the missing values in the following ways.

(i) Ignore tuple: the specific meaning of "ignore tuple" is to remove incomplete data in order to retain complete data. This advantage is that the retained data is the initial data, which shows that the research conclusion is closer to reality. The disadvantage is that if the incomplete value element accounts for a large proportion of the total element, it will be eliminated. This complete original data may affect the correctness of the conclusion.

(ii) Fill the missing values of column attributes with the average value of each column attribute: the advantage of this method is that it can be used to fill the missing values of column attributes, and the number of missing values can be filled into the values of column attributes. The entire property may affect the accuracy of the final information and result in some errors; however, it is a very popular and easy to operate method to supplement missing values.

(iii) Fill in the missing part with the predicted maximum probability value: when inputting the missing value, Bayesian and regression methods should be used, and the inference and decision tree should be referred to when establishing the estimated value. Even if the results obtained by this reasoning method are not satisfactory, the value of the lack of integrity is better, but the method itself is more complex, so the practicability of the practical process is relatively weak.

The specific structure of the new NTPM is as follows:

First of all, in the process of clustering, the most important thing is to set a specific standard for dividing a data set into different clusters. The sample data density in these data sets is generally greater than the defined threshold, so the clustering of data needs to be reflected through the sample density. These clustering algorithms usually only consider the relationship between the size of the data set values and have no good explanation in the trend of the base station traffic change, and they do not take into account the horizontal time relationship of the data set values. There is a certain change trend in the data value before and after the time point, so the base station change trend is introduced into the clustering standard K. In our scheme, the original algorithm of selecting K is improved, where we use two different K areas to set the regulation of K. The number of clusters and the combination of the sample is as follows: G_1, G_2, where $G_1 = \{z_1, z_2, \ldots z_{G_1}\}$, $G_2 = \{z_1, z_2, \ldots z_{G_2}\}$, and G_1, G_2 are inputted into K optimal disjoint clusters. Let $C = \{C_1, C_2, \ldots, C_k\}$. The data set disjoint formula is as follows (Figure 1):

$$\mathrm{SE} = \sum_{i=1}^{K} \|z - m_i\|^2. \quad (1)$$

Among them, SE represents the sum of squares of errors, m_i denotes the sample particle, and $\|z - m_i\|$ is Euclidean distance, which used to detect the distance between the sample and the point from the set C. For the purpose of solving the derivative of a particle SSE, let the derivative be 0 and get the clustering mean value m_i.

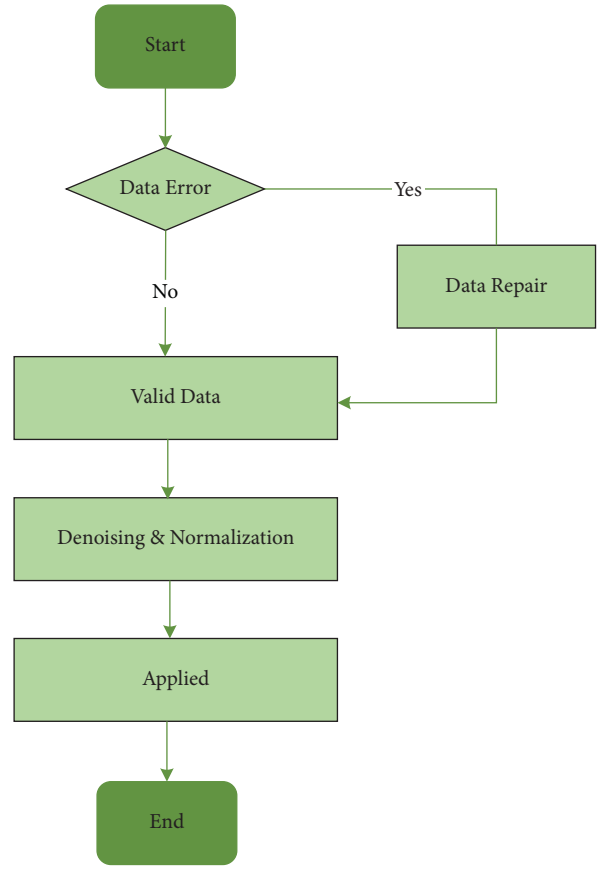

Figure 1: The diagram of preprocessing.

The core idea of clustering algorithm is feature selection and extraction. High value features are extracted from all the features of the data set, where a couple of vectors composed of the selected features are encoded to output new effective features. Then, the appropriate distance function is selected according to the feature form to measure the approximation degree, and then the clustering is carried out through the approximation degree. After completion, the clustering algorithm is evaluated by external effectiveness evaluation or correlation test evaluation.

Secondly, considering the multiscale nature of network traffic, a multiscale network traffic representation method is designed to construct sample sets of different layers. The scale of network traffic is from top to bottom, from fine to coarse, forming an inverted pyramid shape, as shown in Figure 2. This is helpful to deal with and analyze network traffic from different levels and angles and to prepare for the subsequent network traffic prediction. The network traffic represented by multiscale is progressive and can be transformed into each other. The coarse scale network traffic of the lower layer is constructed according to the fine scale network traffic of the higher layer. At the same time, the coarse scale network traffic of the lower layer can be refined into the fine scale network traffic of the higher layer. The multiscale network traffic representation method refers to

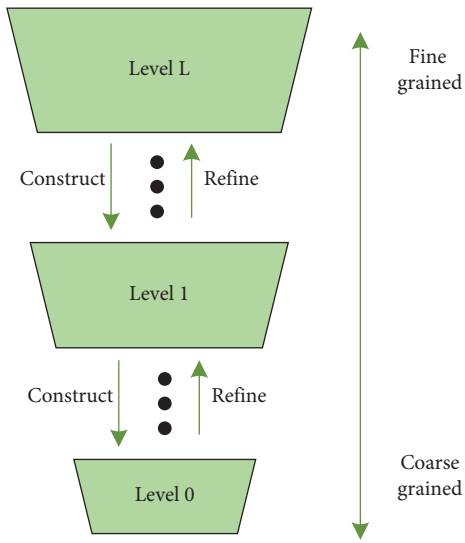

Figure 2: Multiscale network traffic representation.

the hierarchical granularity of each layer to construct different scales. The hierarchical granularity is the time interval.

In the SVM algorithm, the training set is assumed to be $\langle (x_i, y_i), i = 1, 2, \ldots, l \rangle$, in which $x_i (x_i \in R^d)$, $x_i = [x_i^1, x_i^2, \ldots, x_i^d]^T$. x_i is the i_{th} input of column vector of training samples, and $y_i \in R$ is the corresponding output value. The algorithm of SVM is defined as follows.

$$f(x) = w * \phi(x) + b. \tag{2}$$

Among them, w represents the weight of support vector input to output only, $\phi(x)$ reflects the nonlinear mapping between input and high-dimensional features, and b is the distance between input and output. Let ε be the linear insensitive loss function.

$$L(f(x), y, \varepsilon) = \begin{cases} 0, & |f(x) - y| \le \varepsilon, \\ |f(x) - y| - \varepsilon, & |f(x) - y| > \varepsilon. \end{cases} \tag{3}$$

Among them, $f(x)$ denotes the function for prediction, which the output value returned by the regression function, and y is the corresponding true value.

To calculate the value of embedding dimension and delay variable required in NTPM, a reconstruction of the phase space of the original data of the network traffic is necessary because the data is one-dimensional and chaotic. To achieve the purpose of higher prediction accuracy, a new phase space is constructed.

Figure 3 describes the flow of NTPM with optimization algorithm. Specific 7 algorithms are as follows:

(i) Input the number of clusters K, penalty factor C, insensitive loss factor ε, and the kernel constant σ

(ii) Input historical data set F_0, where the flow element f_i consists of two attribute values: traffic t_i and network traffic bandwidth b_i

(iii) Expand F_0. The attribute of the element F_1 is: $f_i = \{t_i, b_i, v_i, p_i, m_i\}$ ($v_i \in \{0, 1, -1\}$, when $v_i = 0$ means no change in flow rate, $v_i = -1$ means the bandwidth decreases, $v_i = 1$ indicates that the bandwidth increases), p_i represents the time period, which is denoted by t_i, and m_i expresses the maximum bandwidth of p_i

(iv) Execute the optimization algorithm. Input F_1, this algorithm is applied to F_1, and add the category label c_i to F_1 to get F_2

(v) Partition training set F_2' and test set F_2''

(vi) Input F_2' to SVM for training

(vii) Input F_2' and F_2'' to SVM for testing

4. Model Simulation and Analysis

4.1. Experimental Environment. In this chapter, the new NTPM is simulated and compared with the BIRCH model. The experimental platform is Intel 3.3 GHz processor, 32 GB memory, Win11 system, and Python 3.7. The data set adopted is a cloud server network traffic data set [7], including 1000 data, each data including bandwidth and time characteristics, the interval is 5 minutes, the division ratio is 6 : 4 in the aspect of training set and test set, using the first 600 network traffic data for the propose of training, and the last 400 network traffic data for the propose of testing. Specific parameters are shown in Table 1.

The comparison indexes are the actual flow forecast results and the average absolute percentage error. Finally, the formulation of MAPE and R are like the following:

MAPE is defined as follows:

$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{|x_{\text{forecast}i} - x_{\text{actual}i}|}{x_{\text{actual}i}} \right). \tag{4}$$

From the definition of MAPE, it shows that the more accurate the prediction effect of the model is, the smaller the value of MAPE.

R is defined as follows:

$$R = \frac{1 - \sum_{i=1}^{n} \left(x_{\text{forecast}_i} - x_{\text{actual}_i} \right)^2}{\sum_{i=1}^{n} \left(x_{\text{forecast}_i} - \ddot{x} \right)^2}. \tag{5}$$

Among them, $\ddot{x} = (1/n) \sum_{i=1}^{n} x_{\text{forecast}_i}$ denotes to the i_{th} prediction results of the flow rate, and x_{actual_i} refers to the actual flow.

Table 2 shows the comparison between the predicted results of different models and the actual flow in ten consecutive data sets.

From Table 1, we can calculate the MAPE of the new NTPM is 11.32%, while that of BRICH model is 15.32%. The MAPE value of the new NTPM is 4% lower than that of BRICH model, so the new NTPM has better prediction accuracy.

The predicted data flows are denoted by the horizontal line. The predicted data flows are denoted by the vertical line. The new NTPM is denoted by the solid, where BRICH model is denoted by the dotted line. From the result, the new NTPM's minimum MAPE value was 4.47%, and the maximum value was 9.25%, where in BRICH model the

Table 1: Specific parameters.

Parameter	Setting
The number of initialization clusters	$K = 5$
Penalty factor	$C = 0.7$
The kernel constant	$\sigma = 4$

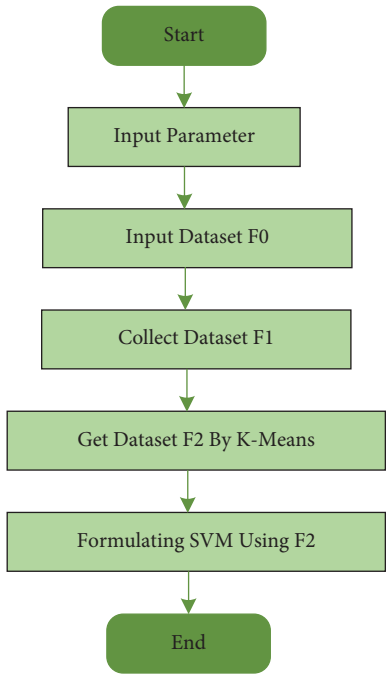

Figure 3: Scheme flow.

Table 2: Comparison between predicted results and actual flow.

Number	Actual traffic (gbps)	The new NTPM	BRICH model
1	4.3187	4.3221	4.3712
2	4.2341	4.2411	4.3785
3	4.5213	4.5631	4.6121
4	5.1201	5.2401	5.3012
5	5.1371	5.1231	5.0101
6	5.1934	5.1761	5.2489
7	5.1206	5.1301	5.3012
8	5.3416	5.4118	5.3376
9	6.3592	6.4118	6.9128
10	6.7121	6.6129	6.5125

minimum value is 4.89% and the maximum value is 15.13%. Therefore, the error of the new NTPM is less than the BRICH model.

The predicted data flows are denoted by the horizontal line. The predicted data flows are denoted by the vertical line. The new NTPM is denoted by the solid, where BRICH model is denoted by the dotted line. From the result, the new NTPM's minimum R value is 83.31%, while the maximum value is 99.87%. However, BRICH model's minimum R value is 6.61%, while the maximum value is 95.12%. Therefore, the fitting degree of the new NTPM is better when compared with the BRICH model, and the accuracy of network traffic prediction is better.

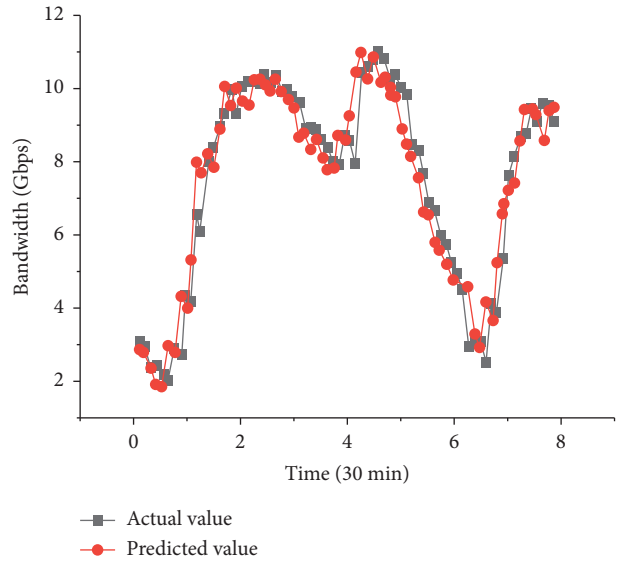

Figure 4: Forecast result and actual result.

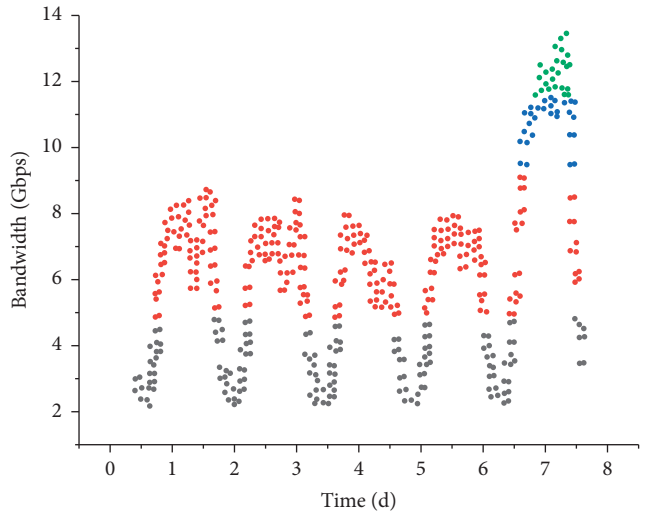

Figure 5: Clustering of the proposed model.

Moreover, from the result of clustering, the new NTPM divides network traffic into more pleasant categories, and the clustering effect is better than that of BRICH model, which shows the new NTPM has better prediction accuracy.

Figure 4 shows the performance of the new NTPM in the prediction results. In Figure 4, the timeline is denoted by the horizontal line. The bandwidth of network traffic is denoted by the vertical line. The triangle marker is the volume predicted by the new NTPM. The actual traffic volume is denoted by the round mark. From Figure 4, the new NTPM realizes a better performance than that of BRICH model, which means the new NTPM has better prediction accuracy.

To sum up, compared with the BRICH model in the aspect of accuracy and clustering, the new NTPM can achieve fine-grained clustering effect and have certain

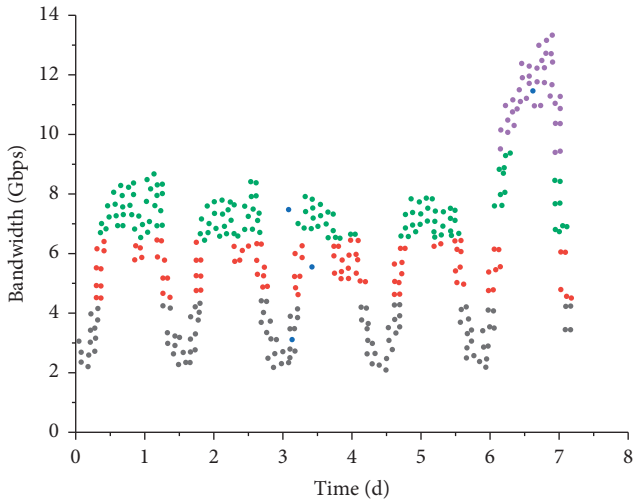

Figure 6: Clustering of the BRICH model.

advantages in the aspect of predicting results, MAPE and R. Therefore, the new NTPM has better performance in NTP.

Figure 5 demonstrates the clustering results of different models, in which the circular markers denote different traffic and the color represents the clustering category of traffic. Similarly, Figure 6 shows the performance of BRICH model in network traffic clustering. As can be seen from the figure, the new NTPM divides network traffic into more surprising categories, and the clustering effect is better than the BRICH model, with better prediction accuracy.

In conclusion, compared with the BRICH model, the new NTPM can achieve fine-grained clustering effect, so it can perform better in network traffic prediction.

5. Conclusion

In the area of many technologies such as big data, cloud computing, and Internet of things, the cloud server plays a critical role, and designing NTPM is very important. To handle the insufficient accuracy of previous NTPMs, a new NTPM with K-means optimization algorithm is formulated in this paper, which can extract the clustering network traffic data by the optimization algorithm and using SVM to handle the model. The experimental results indicate that compared with the BRICH model, the accuracy of NTP of the new NTPM is increased. Also, the new NTPM have reduced communication pressure of cloud server applications, so it can improve the efficiency of cloud server, ensure the quality of service, save computing and communication resources, and reduce the energy cost.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] A. Sari, "A review of anomaly detection systems in cloud networks and survey of cloud security measures in cloud storage applications," *Journal of Information Security*, vol. 6, no. 2, pp. 142–154, 2015.

[2] M. Ala'Anzy and M. Othman, "Load balancing and server consolidation in cloud computing environments: a meta-study," *IEEE Access*, vol. 7, pp. 141868–141887, 2019.

[3] M. Rahman, S. Iqbal, and J. Gao, "Load balancer as a service in cloud computing," in *Proceedings of the 2014 IEEE 8th International Symposium on Service Oriented System Engineering*, pp. 204–211, IEEE, Oxford, UK, April 2014.

[4] X. Liu, X. Fang, Z. Qin, C. Ye, and M. Xie, "A short-term forecasting algorithm for network traffic based on Chaos theory and SVM," *Journal of Network and Systems Management*, vol. 19, no. 4, pp. 427–447, 2011.

[5] A. Y. Nikravesh, S. A. Ajila, C. H. Lung, and D. Wayne, "Mobile network traffic prediction using mlp, mlpwd, and svm," in *Proceedings of the 2016 IEEE International Congress on Big Data (BigData Congress)*, pp. 402–409, IEEE, Washington, DC, USA, December 2016.

[6] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Expert Systems with Applications*, vol. 176, Article ID 114885, 2021.

[7] B. Heller, S. Seetharaman, P. Mahadevan, and Y Yiannis, "Elastictree: saving energy in data center networks[C]//Nsdi," vol. 10, pp. 249–264, 2010.

[8] Z. A. Yilmaz, S. Wang, W. Yang, and Z Haotian, "Applying bert to document retrieval with birch," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 19–24, Hong Kong, China, November 2019.

[9] R. Laref, E. Losson, A. Sava, and M. Siadat, "On the optimization of the support vector machine regression hyperparameters setting for gas sensors array applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 184, pp. 22–27, 2019.

[10] M. Duan, "Short-time prediction of traffic flow based on pso optimized svm," in *Proceedings of the 2018 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pp. 41–45, IEEE, Xiamen, China, January 2018.

[11] J. Tang, X. Chen, Z. Hu, F. Zong, C. Han, and L. Li, "Traffic flow prediction based on combination of support vector machine and data denoising schemes," *Physica A: Statistical Mechanics and Its Applications*, vol. 534, Article ID 120642, 2019.

[12] W. Chen, Z. Shang, and Y. Chen, "A novel hybrid network traffic prediction approach based on support vector machines," *Journal of Computer Networks and Communications*, vol. 2019, Article ID 2182803, 10 pages, 2019.

[13] H. Tian, X. Zhou, and J. Liu, "A hybrid ntpm based on optimized neural network," in *Proceedings of the 2017 18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pp. 284–287, IEEE, Washington, DC, USA, December 2017.

[14] Q. Wang, A. Fan, and H. Shi, "Network traffic prediction based on improved support vector machine," *International Journal of System Assurance Engineering and Management*, vol. 8, no. S3, pp. 1976–1980, 2017.

[15] Z. Tian and S. Li, "A network traffic prediction method based on IFS algorithm optimised LSSVM," *International Journal of*

Engineering Systems Modelling and Simulation, vol. 9, no. 4, pp. 200–213, 2017.

[16] M. Qing-Fang, C. Yue-Hui, and P. Yu-Hua, "Small-time scale network traffic prediction based on a local support vector machine regression model," *Chinese Physics B*, vol. 18, no. 6, pp. 2194–2199, 2009.

[17] P. S. Bishnu and V. Bhattacherjee, "Software fault prediction using quad tree-based K-means clustering algorithm," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 1146–1150, 2012.

[18] S. R. Vadyala, S. N. Betgeri, E. A. Sherer, and A. Amritphale, "Prediction of the number of COVID-19 confirmed cases based on K-means-LSTM," *Array*, vol. 11, Article ID 100085, 2021.