Time Series Data Reconstruction Method Based on Probability Statistics and Machine Learning

Haiying Chen[1(✉)] and Yinghua Liu[2]

[1] Xianning Vocational Technical College, College of Humanities and Arts, Xianning, China
chenhaiying5454@163.com
[2] Wuhan Institute of Design and Sciences, Wuhan, China

Abstract. In order to improve the reconstruction ability of time series data under probability statistical model, a time series data reconstruction method based on machine learning is proposed. The time series data distribution structure model under probability statistical model is constructed. The spatial multi-sensor information sampling method is used to sample the time series data information flow under the probability statistical model, and the phase space reconstruction method is combined to reconstruct the time series data information structure under the probability statistical model. The probability statistical model is established to decompose the time series data, and the distributed grid computing method is used to extract the big data association features of the time series data under the probability statistical model. Combined with the adaptive weight learning method, the optimal control of the scheduling is carried out. The big data cross-domain scheduling of the time series data under the probabilistic statistical model is realized under the support vector machine learning mode. The simulation results show that the method has good adaptability to time series data cross-domain scheduling under the probability and statistics model, and the load balance of data output is strong.

Keywords: Probability statistics · Machine learning · Time series · Reconstruction

1 Introduction

With the development of cloud computing technology, a large number of block chain slicing resource data are distributed in cloud computing storage space, so it is necessary to construct time series data scheduling model under probability and statistics model, combine cloud computing and cloud storage technology to schedule and retrieve time series data across domains, and improve the ability of time series data cross-domain retrieval and query. The research of related time series data cross-domain scheduling method is of great significance in optimizing cloud storage space design and cloud resource information retrieval optimization [1].

In embedded environment, block chain slicing storage system design and resource optimization scheduling under probability and statistics model are carried out, cloud

Y.-D. Zhang et al. (Eds.): ICMTEL 2020, LNICST 327, pp. 149–159, 2020.
https://doi.org/10.1007/978-3-030-51103-6_13

computing storage structure model is established, and block chain slicing storage and cross-domain scheduling under probability and statistics model are carried out by using link structure reorganization method [2, 3]. In the traditional method, the clustering energy consumption scheduling algorithm is mainly used in the cloud resource scheduling algorithm. Combined with the priority list control method, the big data partition scheduling of time series data under the probability statistical model is carried out. In reference [4], a time series data scheduling method based on adaptive priority list control is proposed, combined with the optimal storage and distributed retrieval model of cloud resources, the block chain slicing storage scheduling is carried out. However, the reconstruction ability of time series data under probability statistical model is not good. In reference [5], a resource scheduling algorithm for block chain slicing storage system under the probability and statistical model of double threshold equilibrium control is proposed. The time axis of cloud resource scheduling is divided into uniformly distributed time windows, and the time series data scheduling method is combined with the block structure reorganization method, but the computational overhead of this method is large and the real-time performance is not good.

In order to solve the above problems, a time series data reconstruction method based on machine learning is proposed. Firstly, the time series data distribution structure model under the probability statistical model is constructed, the spatial multi-sensor information sampling method is used to sample the time series data information flow under the probability statistical model, and the phase space reconstruction method is used to reconstruct the time series data information structure under the probability statistical model [6]. Then the feature decomposition model of time series data under probability and statistics model is established, and the big data association feature extraction of time series data under probability statistical model is carried out by combining distributed grid computing method, and the optimal control of scheduling is carried out by combining adaptive weight learning method. Big data cross-domain scheduling of time series data under probability and statistics model is realized under support vector machine learning mode. Finally, the simulation results show the superior performance of this method in improving the reconstruction ability of time series data.

2 Time Series Data Sampling and Information Structure Reorganization

2.1 Time-Series Data Sampling Under Probabilistic Statistical Model

In order to realize big data fusion and cross-domain scheduling of time series data under probability statistical model, it is necessary to construct time series data distribution structure model under probability statistical model, and to sample time series data information flow under probability statistical model by using spatial multi-sensor information sampling method, in order to realize time series data scheduling under probability statistical model. Firstly, the kernel structure and resource storage structure model of block chain slicing storage system under probabilistic statistical model are analyzed, and the information flow model and time series analysis of time series data under probabilistic statistical model are carried out. Under the probabilistic statistical

model, cloud computing grid space realizes memory management, process management and spatial distributed structure storage management through resource cross-domain scheduling [7]. The distribution structure model of time series data under probability and statistics model is shown in Fig. 1.

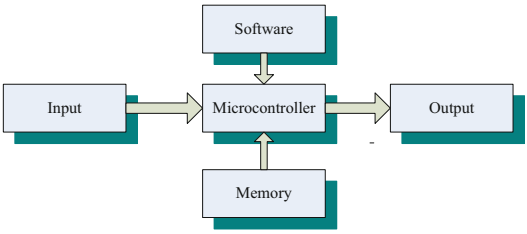

Fig. 1. Model of time series data distribution structure under probabilistic statistical model

In that whole probability statistic model, a block chain fragment storage system adopts a grid form, and various scattered time sequence data are reconstructed and linearly reformed, and through the analysis, a block chain fragment storage structure model of a block chain under a probability statistic model is obtained. in the method, a collaborative distribution method is adopted for designing a resource input interface [8], a time series data large data scheduling in a probability statistical model is adopted, and a cloud computing time series data large data characteristic distribution set of a resource searching module is set as follows:

$$P = \{p_1, p_2, \cdots p_m\}, m \in N \tag{1}$$

Wherein, the m represents the dimension of the data large data distribution space of the time series data under the probability statistical model, the p_m is a fuzzy degree function of the slice storage node, and the data fusion is carried out on the data large data stream of the time series data under the multiple probability statistical models, under the embedded platform, the fragment storage and the structure matching are carried out, and the time series data large data stream under the probability statistical model to be distributed is as follows:

$$flow_k = \{n_1, n_2, \cdots, n_q\}, q \in N \tag{2}$$

In the above formula, q represents the storage depth of the large data stream set of time series data under multiple probabilistic statistical models, n_q represents the data sequence of big data information flow of time series data under probabilistic statistical model, and N represents the total number of big data symbols of time series data under probabilistic statistical model. According to the above analysis, combined with the phase space reconstruction method, the time series data information structure is reorganized under the probability and statistical model [9].

2.2 Time Series Data Information Structure Reorganization

The characteristic decomposition model of time series data under probability statistical model is established, and the time series data structure under probability statistical model is reorganized with distributed grid computing method. The tasks in the client side of time series data scheduling under probability statistical model are submitted to the server, and a distribution set of time series data with N input time series data is obtained. the random number series components of time series data output are as follows:

$$x_{n+1} = 4x_n(1 - x_n)\ n = 1, 2, \cdots, NP \tag{3}$$

Wherein, NP is the distributed bandwidth of time series data in time period T, n represents the number of tasks in process management. The phase space reconstruction method is used to reconstruct the time series of time series data scheduling under probability and statistics model, and the phase space reconstruction equation is obtained as follows:

$$\begin{cases} x = (x_1, x_2, \ldots, x_n) \\ y = F(x) = (f_1(x), f_1(x), \ldots, f_m(x))^T \end{cases} \tag{4}$$

Wherein, the $x = (x_1, x_2, \ldots, x_n)$ is a set of interference data scheduled for the time series data under the probability statistical model; the $y = F(x)$ represents a delay function of the block chain fragment storage in the probability statistical model; and the type of the characteristic vector set n_i of the time series data scheduling is n_i, and the association rule mining method is adopted, the error of the inter-layer prediction of the time series data is $P(n_i) = \{p_k | pr_{kj} = 1, k = 1, 2, \cdots, m\}$, the priority attribute of the time series data scheduling under the probability statistical model is constructed by using the false nearest neighbor method, the reconstruction problem of the time series data under the probability statistical model is converted into a detection problem of a multivariate unknown parameter, The expression of the detected statistical feature quantity is as follows:

$$x_{\min,j} = \max\left\{x_{\min,j},\ x_{g,j} - \rho(x_{\max,j} - x_{\min,j})\right\} \tag{5}$$

$$x_{\max,j} = \min\left\{x_{\max,j},\ x_{g,j} + \rho(x_{\max,j} - x_{\min,j})\right\} \tag{6}$$

In the above formula, ρ is the transmission loss coefficient of time series data reconstruction, and the weight adjustment of the time series data scheduling is constructed in the interval $[x_{\min,j}, x_{\max,j}]$, and the bandwidth of the resource scheduling is SW. Based on the analysis, the information entropy of the time series data scheduling node in the probability statistical model is obtained:

$$H_i(x) = \sum_{k=1}^{K} p_k \ln \frac{1}{p_k} = -\sum_{k=1}^{K} p_k \ln p_k \tag{7}$$

The method for controlling the output of the chain-fragment storage resource to obtain the spatial distribution feature quantity meets the $\Phi : \mathrm{M} \rightarrow R^{2d+1}$, so that the integration degree distribution of the time sequence data scheduling is as follows:

$$\Phi(z) = (h(z), h(\varphi_1(z)), \ldots, h(\varphi_{2d}(z)))^T \tag{8}$$

In the block chain slicing storage system based on probability and statistics model, the adaptive weighted control of time series data is carried out, and the ambiguity function is obtained as follows:

$$x_n = [x(0), x(1), \cdots, x(N-1)]^T \tag{9}$$

The third order statistical feature analysis method is used to decompose the time series data under the probability statistical model [10], and the eigenvalues are defined as follows:

$$cum(\lambda_1 x_1, \lambda_2 x_2, \cdots, \lambda_k x_k) = \left(\prod_{i=1}^{k} \lambda_i\right) cum(x_1, x_2, \cdots, x_k) \tag{10}$$

Among them, the k-order cumulant of time series data under probability statistical model is $c_{kx}(\tau_1, \tau_2, \cdots, \tau_{k-1})$, resource load balanced transmission information flow $\{x(n), x(n+\tau_1), \cdots, x(n+\tau_{k-1})\}$, according to the above analysis, the time series data information structure reorganization model is constructed, combined with subspace fusion method, the time series data reconstruction is carried out [11].

3 Time Series Data Reconstruction and Optimization

3.1 Association Feature Extraction

On the basis of sampling the information flow of time series data under probability statistical model by using spatial multi-sensor information sampling method, the feature decomposition model of time series data under probability statistical model is established, and the time series data structure of time series data under probability statistical model is reorganized with distributed grid computing method [12]. The optimal control model is obtained, and the output characteristic quantity of time series data cross-domain scheduling is described as follows:

$$x(t) = \mathrm{Re}\{a_n(t) e^{-j2\pi f_c \tau_n(t)} s_l(t - \tau_n(t)) e^{-j2\pi f_c t}\} \tag{11}$$

In the support vector machine learning mode, the load balancing characteristics of resource scheduling can be described as follows:

$$c(\tau, t) = \sum_n a_n(t) e^{-j2\pi f_c \tau_n(t)} \delta(t - \tau_n(t)) \tag{12}$$

In the above formula, $a_n(t)$ is the output information flow of block chain slicing scheduling on n channels, $\tau_n(t)$ is the delay on n transmission channels, f_c is the modulation frequency of block chain slicing scheduling under probabilistic statistical model, the adaptive weighting method is used to schedule the time series data in real time under probabilistic statistical model [13], and the attribute distribution quantification function of time series data in probabilistic statistical model is as follows:

$$\mathrm{S(i,j)} = \frac{\sum_{u \in U_{\mathrm{ij}}} (V_{u,i} - 3)(V_{u,j} - 3)}{\sqrt{\sum_{u \in U_{ij}} (V_{u,i} - \overline{V_{.i}})^2} \sqrt{\sum_{u \in U_{ij}} (V_{u,j} - \overline{V_{.j}})^2}} \tag{13}$$

In which, the time series data information resource evaluation matrix of the probability statistical model is an optimized characteristic solution of the time series database distribution set of the $R = \left(r_{\mathrm{ij}}, a_{\mathrm{ij}}\right)_{m \times n}$ and the probability statistical model:

$$\Phi = diag[e^{j\phi_1}, \cdots, e^{j\phi_P}] \tag{14}$$

According to the analysis, the sample set distribution model of the time series data under the probability statistical model is met:

$$T_{i,j}(t) = \frac{\left|p_{i,j}(t) - \Delta p(t)\right|}{p_{i,j}(t)} \tag{15}$$

Where, $U_{i,j}(t)$ is used to represent the association rule items of time series data under the probabilistic statistical model. according to the above analysis, the association features are extracted [14].

3.2 Weight Analysis and Time Series Data Reconstruction Output

Combined with adaptive weight learning method to optimize scheduling, the time series data of time series data in probabilistic statistical model is scheduled across domains under support vector machine learning mode. 4 tuples $\left(E_i, E_j, d, t\right)$ is used to represent the main feature decision tree of time series data sharing scheduling under probabilistic statistical model. E_i, E_j is the bifurcation node of time series data in directed graph under probabilistic statistical model [15]. The differential fusion feature quantity of time series data sharing under probability and statistics model is obtained.

$$J_m(U, V) = \sum_{k=1}^{n} \sum_{i=1}^{c} \mu_{ik}^m (d_{ik})^2 \tag{16}$$

In the formula, m is a finite data set of time series data distribution under a probability statistical model, and $(d_{ik})^2$ is a similarity distribution map, and the decision-making independent variable of the time series data under the probability statistical model is as follows:

$$(d_{ik})^2 = \|x_k - V_i\|^2 \tag{17}$$

And

$$\sum_{i=1}^{c} \mu_{ik} = 1, k = 1, 2, \cdots, n \tag{18}$$

The optimal scheduling and mining of time series data under probabilistic statistical model is carried out [16], and the priority clustering model of time series data sharing scheduling under probabilistic statistical model is obtained as follows:

$$V_{u,i} = \frac{D_i^-}{D_i^+ + D_i^-}, \overline{V_{.j}} = \frac{R_i^+}{R_i^+ + R_i^-} \tag{19}$$

Based on the analysis, the adaptive fusion clustering model of time series data under probabilistic statistical model is constructed, and the reconstruction of time series data under probabilistic statistical model is realized by using time series data fusion method [17–20].

4 Simulation Experiment and Result Analysis

In order to test the application performance of this method in time series data scheduling under probabilistic statistical model, Matlab is used to carry out simulation experiment, and embedded Linux technology is used to design the platform of time series data scheduling under probabilistic statistical model. The sampling time length of time series data under probabilistic statistical model is 60 s, the characteristic sampling

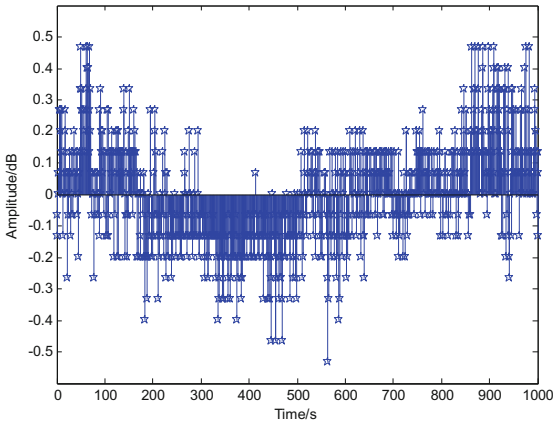

Fig. 2. Time domain waveform of time series data distribution under probability statistical model

frequency is 80 kHz, and the carrier frequency of block chain slicing storage is 12 kHz. A frequency component of 250 Hz is set between 400 and 600 sampling points for distributed adjustment of time series data, and the sampling of time series data information under probability and statistics model is shown in Fig. 2.

Taking the resource data of Fig. 2 as the research object, the time series data scheduling under the probability statistical model is carried out, and the reconstruction output is shown in Fig. 3.

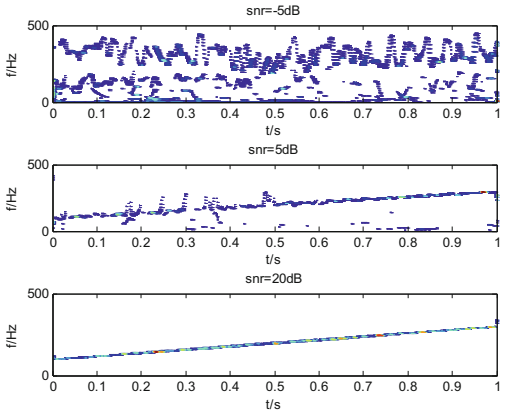

Fig. 3. Time series data to reconstruction output under probabilistic statistical model.

The analysis Fig. 3 shows that the proposed method can effectively realize the time series data scheduling under the probability statistical model, and test the equilibrium degree of the block chain storage resource scheduling under the probability statistical model. The comparative results are shown in Table 1.

Table 1. Equalization comparison of block chain reconstruction under probability and statistical model

Iterations	Proposed method	Reference [4]	Reference [5]
100	0.913	0.845	0.845
200	0.924	0.876	0.864
300	0.954	0.914	0.912
400	0.987	0.926	0.934
500	0.998	0.943	0.967

As shown in Table 1, compared with the traditional method, the proposed method has higher equilibrium of block chain and higher practical application. The analysis Table 1 shows that the proposed method has a better balance in the open source scheduling of the block chain time series data under the probability and statistics model.

In order to verify the complexity of the proposed algorithm and the running time of the algorithm, the experimental results are shown in Fig. 4.

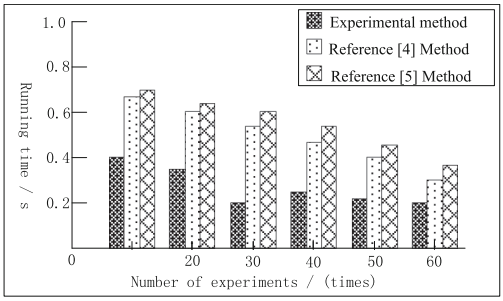

Fig. 4. Comparison of running time of different methods

As shown in Fig. 4, under the same conditions, the proposed method has the shortest running time, indicating that the proposed method has the lowest complexity and high operating efficiency.

5 Conclusions

In this paper, a time series data reconstruction method based on machine learning is proposed. The time series data distribution structure model under probability statistical model is constructed. The spatial multi-sensor information sampling method is used to sample the time series data information flow under the probability statistical model, and the phase space reconstruction method is combined to reconstruct the time series data information structure under the probability statistical model. The probability statistical model is established to decompose the time series data, and the distributed grid computing method is used to extract the big data association features of the time series data under the probability statistical model. Combined with the adaptive weight learning method, the optimal control of the scheduling is carried out. The big data cross-domain scheduling of the time series data under the probabilistic statistical model is realized under the support vector machine learning mode. The simulation results show that the method has good adaptability to time series data cross-domain scheduling under the probability and statistics model, and the load balance of data output is strong. The method has good application value in time series data reconstruction. However, due to the limited time, the efficiency of reconstruction of time series data needs to be improved, which is also my future research direction.

6 Acknowledgment

The 2018 annual scientific research project of Hubei Provincial Department of education. Based on modern statistical theory and machine learning theory, economic time series analysis is carried out (B2018371)

References

1. Bi, A., Dong, A., Wang, S.: A dynamic data stream clustering algorithm based on probability and exemplar. J. Comput. Res. Dev. **53**(5), 1029–1042 (2016)
2. Yang, J., Wei, C.: Testing serial correlation in partially linear additive models. Acta Mathematicae Applicatae Sinica, English Serie **35**(2), 401–411 (2019)
3. Zhang, X., He, Y.-H.: Modifid interpolatory projection method for weakly singular integral equation eigenvalue problems. Acta Mathematicae Applicatae Sinica English Serie **35**(2), 327–339 (2019)
4. Zhou, Z., Rahman Siddiquee, Md.M., Tajbakhsh, N., Liang, J.: UNet ++: a nested u-net architecture for medical image segmentation. In: Stoyanov, D., et al. (eds.) DLMIA/ML-CDS -2018. LNCS, vol. 11045, pp. 3–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00889-5_1
5. Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Jorge Cardoso, M.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Cardoso, M.J., et al. (eds.) DLMIA/ML-CDS -2017. LNCS, vol. 10553, pp. 240–248. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67558-9_28
6. Huang, G., Liu, Z., Laurens, V.D.M., et al.: Densely connected convolutional networks. In: CVPR 2017 Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2261–2269. IEEE Computer Society, Washington, DC (2017)
7. Pan, C., Jia, Y., Cai, R., Yang, L.: Routing strategy for spatial information network based on MPLS. Comput. Eng. **45**(3), 85–90 (2019)
8. Lee, G.M., Lee, J.H.: On nonsmooth optimality theorems for robust multiobjective optimization problems. J. Nonlinear Convex Anal. **16**(10), 2039–2052 (2015)
9. Ma, M.Y., Chen, S.L., Zuo, Y.: Research on private set intersection cardinality protocol based on Goldwasser-Micali encryption system. Appl. Res. Comput. **35**(9), 2748–2751 (2018)
10. Patricia, N., Caputo, B.: Learning to learn, from transfer learning to domain adaptation: a unifying perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1442–1449. Columbus, OH, USA (2014)
11. Sun, L., Guo, C.H.: Incremental affinity propagation clustering based on message passing. IEEE Trans. Knowl. Data Eng. **26**(11), 2731–2744 (2014)
12. Yu, Q., Tu, G., Li, N., Zhou, T.: Multi-hop multi-policy attributed-based fully homomorphic encryption scheme. J. Comput. Appl. **39**(8), 2326–2332 (2019)
13. Goyal, V., Pandey, O., Sahai, A., et al.: Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM Conference on Computer and Communications Security, pp. 89–98. ACM, New York (2006)
14. Mernik, M., Liu, S.H., Karaboga, M.D., et al.: On clarifying misconceptions when comparing variants of the Artificial Bee Colony Algorithm by offering a new implementation. Inf. Sci. **291**(10), 115–127 (2015)
15. Hsieh, T.J.: A bacterial gene recombination algorithm for solving constrained optimization problems. Appl. Math. Comput. **231**(15), 187–204 (2014)

16. Gentry, C., Sahai, A., Waters, B.: Homomorphic encryption from learning with errors: conceptually-simpler, asymptotically-faster, attribute-based. In: Canetti, R., Garay, J.A. (eds.) CRYPTO 2013. LNCS, vol. 8042, pp. 75–92. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40041-4_5
17. Brakerski, Z., Cash, D., Tsabary, R., Wee, H.: Targeted homomorphic attribute-based encryption. In: Hirt, M., Smith, A. (eds.) TCC 2016. LNCS, vol. 9986, pp. 330–360. Springer, Heidelberg (2016). https://doi.org/10.1007/978-3-662-53644-5_13
18. Gentry, C., Peikert, C., Vaikuntanathan, V.: Trapdoors for hard lattices and new cryptographic constructions. In: Proceedings of the 40th Annual ACM Symposium on Theory of Computing, pp. 197–206. ACM, New York (2008)
19. Yan, X.X., Ye, Q., Liu, Y.: Attribute-based encryption scheme supporting privacy preserving and user revocation in the cloud environment. Netinfo Secur. **17**(6), 14–21 (2017)
20. Zhou, Y.H., Shi, W.M., Yang, Y.G.: A quantum protocol for millionaire problem with continuous variables. Commun. Theoretical Phys. **61**(4), 452–456 (2014)