Hindawi
Mathematical Problems in Engineering
Volume 2018, Article ID 3719564, 10 pages
https://doi.org/10.1155/2018/3719564

Research Article

Hot Spot Data Prediction Model Based on Wavelet Neural Network

Ming Zhang [1,2] **and Wei Chen** [2]

[1] *School of Information Science and Engineering, Linyi University, Linyi, Shandong 276005, China*
[2] *School of Information Engineering, Wuhan University of Technology, Wuhan, Hubei 430070, China*

Correspondence should be addressed to Ming Zhang; zhangming2000225@163.com

Received 24 August 2018; Accepted 16 October 2018; Published 30 October 2018

Guest Editor: Weihai Zhang

The novel hybrid multilevel storage system will be popular with SSD being integrated into traditional storage systems. To improve the performance of data migration between solid-state hard disk and hard disk according to the characteristics of each storage device, identifying the hot data block is significant issue. The hot data block prediction model based on wavelet neural network is built and trained by using historical data. This prediction model can overcome the cumulative effect of traditional statistical methods and has strong sensitivity to I/O loads with random variations. The experimental results show that the proposed model has better accuracy and faster learning speed than BP neural network model. In addition, it has less dependence on sample data and has better generalization ability and robustness. This model can be applied to the data migration of distributed hybrid storage systems to improve performance.

1. Introduction

The hybrid storage system with traditional hard disk drive (HDD) and solid-state drive (SSD) has become a significant storage system in large-scale data processing [1]. HDD have been broadly used as the primary storage medium in the past decades due to their relatively low cost per gigabyte and large storage capacities. However, the access performance of HDD has been improving slowly and the performance gap between HDD and CPU has been becoming bigger. The HDD has become the performance bottleneck of the computer system, even in the datacenter [2, 3]. The SSD is flash-based electronic storage device with low read delay and high writing delay [4]. So the SSDs are installed into the traditional HDD-based storage systems to build the HDD/SSD hybrid storage system [5]. To exploit the high random read performance owned by SSD, the hybrid storage systems focus on identifying hot data from cold data accurately. These systems perform data migration operation to copy hot data from HDD to SSD and cold ones from SSD into HDD for improving the access performance. Several authors analyzed the convergence and oscillatory properties of data using differential equations and dynamic equations on time scales; see, e.g., [6–9]. So the identification of data status in hybrid storage is a significant issue. Here we proposed a prediction model to identify the hot data in hybrid storage system and this model can be used in data migration in cloud storage system or distribution storage systems [10].

To identify the hot data in storage system, the current researches focus on statistical method that logs and counts the number of accessing for each data block. Kgil [11] and Koltsidas [12] proposed data block classification model based on data page I/O statistics to guide data migration strategy. Yang put forward to a novel method for increasing intermediate state to reduce occasional read tendency page classification error and the caused false migration in order to reduce the error in the classification of data pages [1]. Chai et al. used data block temperature to qualify the data block status and identify hot data block by sorting data block temperature [13]. However, the above mentioned models belong to statistical method that has a congenital deficiency. This is because these statistical models calculate the accessing requests from the data block creation. It has a certain time accumulation effect, which will cause classification model to be insensitive to the

change of random access load. There is a large number of I/O random access loads in complex application environments where large data is stored and calculated. This statistical classification model cannot respond quickly to changes in storage load and inevitably affects the accuracy and cost of data migration in hybrid storage systems. To solve the cumulative effect of statistical method and improve the response for random access workloads, we build a prediction model based on wavelet neural network to identify the hot and cold data block. Our contributions can be summarized as follows.

(1) We predict the data block status based on the wavelet neural network that has better nonlinear learning and generalization capacity. This model is very suitable for the small training data set and can learn with high resolution in sparse training data set. Additionally, the prediction model generates the thermal degree of data block, which is a continuous value within $[0, 1]$. So it is more flexible for data migration mechanism than previous models.

(2) The prediction model can solve the shortcoming of statistical method and response of the variation of hot data faster than those statistical models. This model is more suitable to deal with the uncertain workloads in distribution storage system.

This paper is organized as follows: in Section 2 we briefly describe the problem of identifying hot data. Section 3 introduces the features and qualifying model. Section 4 provides the detailed prediction model. Section 5 we evaluate the prediction model by experiments and analyze the results. Finally, Section 6 concludes this paper with discussion of future work.

2. Problem Description

In this section, we introduce the problem firstly. Accessing Data block rules are very complex in large-scale data distributed storage environment. According to historical data access trends, the current data block read operation heat degree is predicted to provide basic basis for data migration in hybrid storage system. Distributed big data storage system is a multitenant computing platform. On the platform a large number of users launch random computing tasks, so a lot of data's reading and writing requests will be produced. For the convenience of later discussion, some of the important concepts are defined as follows.

Definition 1. Data block B. A data block is a triple: $B = \langle F, S, P \rangle$, where F is the file that the data block belongs to. S is the size of the attached file. P is the set of data block access rule attributes.

Definition 2. Data block access attribute set P. Data block access attribute set $P = (\delta, \theta, T, \varsigma)$, where δ is the data block unit time access frequency. θ is the access density in statistical period. T is the data block inertia; ς is the data block concurrency.

Definition 3. The access frequency per unit timeδ is the access frequency per unit time in a statistical cycle. For example, the statistical period for a data block is 30 days and 240 times of access happens in this statistical cycle. They happens in 3 days, respectively, and then $\delta = 240/3$.

Definition 4. θ is the access density in statistical period. The ratio of the number of unit times of data access in a statistical period to the length of the statistical period. It reflects the distribution of data block access. For example, a month is calculated by 30 days, a total of 240 times of access happen, they happen in different 3 days respectively, then $\theta = 3/30$; if happen in 18 days, then $\theta = 18/30$.

Definition 5. The recent data block inertia T. In distributed storage system, from the beginning of the creation of data block, the set of time for each access is $\langle t_1, t_2, \ldots, t_n \rangle$. The time interval for each access is t_2-t_1, t_3-t_2, t_n-t_{n-1}. For easy measurement and use, this paper uses the most recent access interval of data block as the inertia value of data block. For example, if current time is t_i, data inertia is $T = t_i$-t_{i-1}.

Definition 6. Data block concurrency ς is the number of processes which access data block. This index reflects the degree of concurrent access to data block to a certain extent.

Definition 7. Life of data blocks L. Time interval between current time and the time of data block creation. Usually the longer the life of the data block is, the lower the probability of being accessed is.

In this case, the future access probability of data block is predicted according to the access characteristics and laws of large number of data blocks. The data blocks with high access probability are hot data. On the contrary, the data blocks with low access probability are cold data. Then whether the data state is hot or cold is determined to prepare for further data migration.

3. Data Heat and Eigenvector

First, create an index to quantify data block heat degree, and this index can express the frequency that a data block is accessed.

Definition 8. Data heat degree h. Using data block access frequency and access density, two important access attributes and weighting them to build data heat degree index, as shown in

$$h = w\delta + (1 - w)\theta, \quad h \in [0, 1]. \tag{1}$$

In Equation (1), δ and θ are the two dimensionless indexes with same direction. The larger the value is, the greater the heat of data block is. In addition, the maximum method is used to normalize δ, that is $\delta = \delta_{cur}/\delta_{\max}$, $\delta \in [0, 1]$. θ is a real number in interval $[0, 1]$. This method of describing data block heat mainly focuses on the main indexes of data block access. It has advantages of simpleness and intuitiveness. Because the weight is assigned value with subjective method, $w = 0.7$, this weight identification has a certain subjective shortcomings.

In order to establish the prediction model of thermal data, we need to construct training data set from large number of data block access history data. First, we select eigenvalues which can reflect access rule and is easy to measure from sample data, and construct eigenvector for each data sample. The following lists the specific eigenvectors $\langle T, \varsigma, S, L\rangle$: the recent data block inertia T, concurrency ς, data block life, etc. They are indexes which have more significant correlation with data block heat. Data block inertia is normalized by maximum value method, that is, $T = T_{cur}/T_{\max}$, $T \in [0, 1]$. Then the method is used to perform co-oriented treatment for characteristic indexes, $T = T_{\max}/T_{cur}$. Data block concurrency also needs to be normalized with maximum value, that is, $\varsigma = \varsigma_{cur}/\varsigma_{\max}$, $\varsigma \in [0, 1]$. S the size of the file associated by data block is looked as a supplementary eigenvalue, after normalization, $S = S_{cur}/S_{\max}$, $\varsigma \in [0, 1]$. Data block life value is also normalized with maximum value. $L = L_{cur}/L_{\max}$, $L \in [0, 1]$. Then this method is used to co-orient index value. According to the observation on training data set, the maximum value of the above eigenvalues is selected by artificial experience for normalization, where $\delta_{\max} = 500$, $T_{\max} = 200$ statistical time points, $\varsigma_{\max} = 30$ concurrent users, $S_{\max} = 500$MB, and $T_{\max} = 1$ year.

4. Data Block Heat Degree Prediction Model

Data heat degree prediction model is a nonlinear function relationship $f(\mathrm{T}, \varsigma, S, L) = h$. The input of the prediction model is a quarternary real number vector. The model output is the heat state of data block $h \in [0, 1]$. By training the prediction model, the nonlinear function relation between the features vector of data block access and the heat degree is fitted. We choose wavelet neural network (WNN) to construct the prediction model. The wavelet neural network integrates the advantages of wavelet analysis and artificial neural network, and has excellent nonlinear mapping ability and generalization ability [14]. In addition, the experimental observation data show that the training sample data has obvious sparsity. A large number of sample data show similar access characteristics. A small number of sample characteristics has diversity. Therefore, the wavelet neural network has obvious advantages in learning. Due to the multiscale time-frequency analysis capability of the wavelet function in wavelet neural network, the local singularity function can be studied with high accuracy by adjusting the expansion and translation of the wavelet function in training data-intensive region. In training data sparse area, low-scale parameters are used to learn smooth function. The characteristics of training data are the main reason to select wavelet neural network for fitting nonlinear prediction model. WNN has better adaptability and better prediction accuracy than traditional BP neural network [15].

4.1. Fuzzy Preprocessing Eigenvalues. In feature vector $\langle T, \varsigma, S, L\rangle$, relevant file size S and data block life L have obvious uncertainty. The value difference of various sample data is larger. So if using the means of standard quantification, the value of two indexes will affect the proportion of other indexes and reduce the accuracy of training model. In order to solve this problem, these two indexes are quantified by fuzzy evaluation method, and the output of fuzzy processing is taken as the input of wavelet neural network [16].

The concrete method is as follows: Firstly, we define the fuzzy set $A = \{(a, \mu_A(a)), a \in R, \mu_A(a) \in [0, 1])\}$, where $\mu_A(a)$ is the appurtenant degree function of the fuzzy set, which reflects the degree of the element a belonging to the fuzzy set A. For example, the data block-related file size is set to three levels: small, medium, and large. The appurtenant degree function of large file is $\mu_{great}(e(s))$, $\{(e(s), \mu_{great}(e(s))), e(s) \in [0, 1], \mu_{great}(e(s)) \in [0, 1])\}$, and it is expressed as the extent that the data block size e (s) belongs to the fuzzy set large file. In order to blur file size index, we use fuzzy appurtenant function as in

$$\mu_A(a) = \begin{cases} 0, & a < l \\ \dfrac{a-l}{b-l}, & l \le a < b \\ 1, & b \le a < c \\ \dfrac{h-a}{h-c}, & c \le a < h. \end{cases} \tag{2}$$

First normalize file size with maximum method, normalize it into [0, 1]. Then use Equation (2) to process. For example, if $e(s) = 0.47$, then $\mu_{small}(0.47) = 0.12$, $\mu_{median}(0.47) = 0.72$, $\mu_{great}(0.47) = 0.16$; therefore E (s) = 0.47 represents the fuzzy set $\{(0.47, \mu_{small}(0.47)), (0.47, \mu_{median}(0.47)), (0.47, \mu_{great}(0.47))\} = \{(0.47, 0.12), (0.47, 0.72), (0.47, 0.16)\}$. Design three fuzzy sets $\mathrm{U_s}$, $\mathrm{U_m}$, and $\mathrm{U_g}$ for data block association file size index; they represent small files, medium files, and large files, respectively (denoted as Small, Median, and Great). Then, the same method is used to define three fuzzy sets (old data, middle-aged data, and new data) for data block life index, so as to solve the fuzzy processing of this index.

4.2. Wavelet Neural Network Structure. The main idea of wavelet neural network is to combine wavelet function with traditional BP neural network, replace the sigmoid function in BP neural network with nonlinear wavelet basis function, and use the linear superposition of nonlinear wavelet basis to fit the nonlinear function [14]. In this paper, the wavelet neural network structure is shown in Figure 1.

The wavelet neural network in Figure 1 consists of input layer, fuzzy layer, inference layer, wavelet layer, and output layer. The number of neuron nodes in each layer is n, n × M, M, M, and 1, respectively.

(1) Input layer (Layer 1): Each node of this layer is directly connected to each input component xj of the eigenvector, and passes the input value $\mathrm{X}=[\mathrm{x_1},\mathrm{x_2},\mathrm{x_3},\ldots,\mathrm{x_n}]$ to the next layer. Here the eigenvector is $\langle T, \varsigma, S, L\rangle$, n=4;

(2) Fuzzy layer (Layer 2): Through the fuzzy rules, the input vector is transformed into fuzzy values. Here Gaussian function is used to complete the work of the fuzzy appurtenant function, as shown in

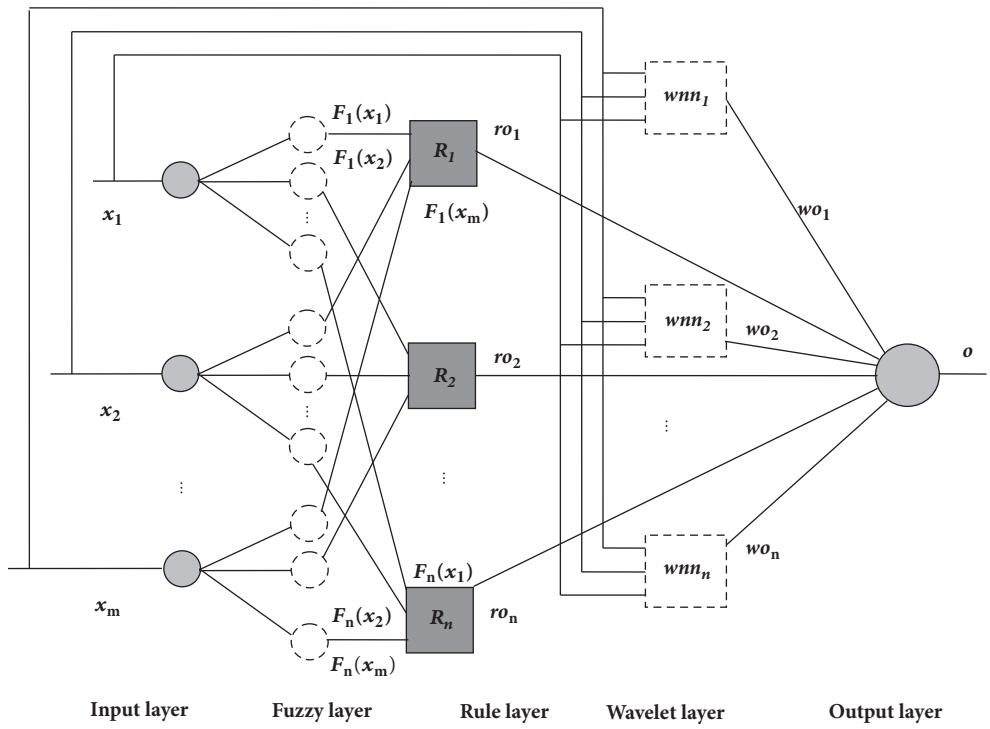

Figure 1: Architecture of wavelet neural networks.

$$F_j(x_i) = \exp\left(-\frac{\left(x_i - c_{ij}\right)^2}{2{\sigma_{ij}}^2}\right) \quad \left(j = 1, 2, \ldots, n,\ i = 1, 2, \ldots, m\right), \tag{3}$$

where i is the component of input vector, j is the ordinal number of fuzzy rule, c_{ij} and σ_{ij} are the center position and distribution parameter of Gaussian function, respectively. In addition, the fuzzy rule can be described as Ri: if x_1 is A_{i1} then $y_i = B_{i1}$.

(3) Rule layer (Layer 3): Rule layer is also known as inference layer. Each neuron corresponds to a fuzzy rule. Then we use the following formula to calculate the fitness degree of fuzzy rules.

$$\mu_j(x) = \prod_{j=1}^{n} F_j(x_i) \quad (i = 1, \ldots, m). \tag{4}$$

In (4), n is the rule ordinal number and $\mu_j(x)$ is the input value of next layer.

(4) Wavelet layer (Layer 4): The wavelet network layer output is mainly used for output compensation. The output of the neurons of this layer is calculated by using

$$wo_j = w_j \Psi_j(z)$$
$$\Psi_j(z) = \sum_{i=1}^{m} \frac{1}{\sqrt{\left|a_{ij}\right|}} \left(1 - {z_{ij}}^2\right) e^{-{z_{ij}}^2/2}, \tag{5}$$

where $z_{ij} = (x_i - b_{ij})/a_{ij}$ is the weight of the connection from the ith node of input layer to the jth node of wavelet layer; w_j is the weight of the connection from the jth node of wavelet layer to output layer; a_{ij} is the extension and contraction parameter of wavelet function; b_{ij} is the translation parameter of wavelet function.

(5) Output layer (Layer 5): This layer is the final output layer of wavelet neural network. It will produce prediction results of predictive model. As a result of using fuzzy rules to quantify eigenvector index, this layer also perform defuzzy calculation, so it is also known as anti-fuzzy layer.

4.3. Training Algorithm. The wavelet neural network model in the previous section can be expressed as follows:

$$Q(X) = \sum_{j=1}^{N} W_{ij}^{(J)} \Psi_j(z) + \overline{\theta}, \tag{6}$$

where $X = [x_1, x_2, x_3, \ldots, x_M]^T$ is input vector; as shown in (5) and $\Psi_j(z)$ is the neuron wavelet activation function of wavelet layer; W_{ij} is the weight of the connection from the jth node of wavelet layer to output layer and $\overline{\theta}$ is the average estimated value of output sequence; Q(X) is the output value of wavelet neural network. The purpose of wavelet neural network training is to use sample data to determine the important parameters. This model parameters to be trained are Z_{ij}, W_j, a_j, b_j, and $\overline{\theta}$. We do not use the batch approach of traditional BP for training. Here we use genetic algorithm for repeated approximation to actual measurement data in sample space to obtain. Firstly, the parameters to be trained in model are constructed into parameter vector with string structure by their order. Each vector is a chromosome for

genetic operation. Each chromosome is encoded with a real number. The initial value of the parameter is determined by the following method:

(1) Determination of Stretching and Translation Parameters. According to the nature of wavelet function, the window center position and width of the wavelet function are fixed values. Given that the initial center of the wavelet window is x_{0j} and the window width is Δx_{0j}, then the scaling factor a_j is given by the following formula:

$$a_j = \frac{\sum_{j=1}^{M} x_{j\max} - \sum_{j=1}^{M} x_{j\min}}{\Delta x_{0j}}. \quad (7)$$

The translation factor b_j is determined by Equation (8):

$$b_j = 0.5 \times \left(\sum_{j=1}^{M} x_{j\max} + \sum_{j=1}^{M} x_{j\min} \right) - a_j \times x_{0j}, \quad (8)$$

M in (7) and (8) are the number of input vectors. $x_{j\max}$ and $x_{j\min}$ are the maximum and minimum sample values of the j-th neuron of the input layer, respectively.

(2) Determination of Network Weights. The initial value of the weight from input layer to wavelet layer Z_{ij} and the weight from wavelet layer to output layer W_j is to select a uniformly distributed random number in $[-1, 1]$ and to ensure that the various values are not zero.

(3) Determination of the Parameter $\overline{\theta}$. It is obtained according to the calculated mean of partial sample data.Then it is constantly updated and corrected during calculation.

Before the start of wavelet neural network training, the genetic population size was set to $L = 200$, the maximum number of iterations was $J = 300$, the network convergence accuracy was $= \varepsilon = 3\times10^{-3}$, the probability of selection was $p_s = 0.65$, the crossover probability was $p_c = 0.8$, and the probability of variation is $p_m = 0.03$. The specific training algorithm is as follows [17]:

Step 1. Set the initial value of the iteration variable, $J = 0$, and then base on the determination method of parameter initial value to create L initial parent classes $T_1^{(0)}, T_2^{(0)}, \ldots, T_L^{(0)}$;

Step 2. Calculate fitness function, as shown in (9). When the smaller the value of the adaptive function is, it means the better the network training effect is [18].

$$E\left(T_l^{(J)}\right) = \frac{1}{2\sum_{k=1}^{K}\left[Q(X_k, T_l^{(J)} - Q'_k\right]^2}, \quad (9)$$

where $Q(X_k, T_l^{(J)}$ is the wavelet neural network output value calculated by (6), Q'_k is the expected output value of prediction model, and K is the number of elements in training sample set.

Step 3. Cross and mutate the jth generation of chromosomes, and select N individuals to enter the next generation of evolution.

Step 4 (determine convergence). When the condition $(E_{\max} - E_{\min})/E_{avg} \leq \varepsilon$ or $J > J_{\max}$ is satisfied, the training algorithm ends; otherwise update the variable J, $J = J + 1$ and then return to Step 3. $E_{\max}$, $E_{\min}$, and E_{avg} represent the maximum, minimum, and average value of the calculated fitness function, respectively.

Step 5. Select the best combination of parameters that has reached best fitness accuracy in the previous step and then perform real-time prediction.

By now, the data block hotspot prediction model with wavelet neural network as core has been completed. The training and learning of the predictive model is described in a concrete example. First, the eigenvector $d_1 = \langle T, \varsigma, S, L\rangle$ of a data block is obtained by measuring, $d_1 = (0.47,0.38,0.6,0.3)$ after normalization. Then the heat degree of data block is calculated by the heat degree equation, $h = 0.45$. The calculated heat degree value is the expected output value of the sample. The sample data is denoted as $d_1 = \langle(0.47, 0.38, 0.6, 0.3), 0.45\rangle$. After selecting 1000 such sample data, the training algorithm is used to obtain the main parameters of wavelet neural network. The constructed wavelet neural network model is used to predict. For example, the input vector of the data block to be predicted is $d_2 = (0.37,0.18,0.3,0.3)$; then the model output value $\widetilde{h} = 0.39$ is the predicted heat degree of the data block.

5. Experiment and Analysis

Since the existing disk load dataset does not provide detailed I/O information, we use the disk access data in actual production environment to train the prediction model and perform performance analysis. The experimental environment is Linux operating system, and the file system is ext2, each data block is 4KB. The following methods are used to measure the access character data of data blocks in disk: Blktrace is used to collect I/O data of disk data blocks in Linux system. Blktrace can monitor the I/O events of a particular disk block device, capture and record events such as reading, writing, and synchronous operations [19]. Then blkparse is used to analyze blktrace log data, and we can obtain the attributes such as the processes which accessed the data block, the associated file node, and timestamp. Writing script analysis program based on this tool software, access feature attributes of any one data block of the monitored disk such as access frequency, access density, associated file size, and concurrent program number can be obtained.

5.1. Performance Analysis. In this section, we mainly validate the advantages of the prediction model constructed by wavelet neural network. In this paper, we use traditional BP neural network as the benchmark model to compare. First observe the distribution of access attributes of sample data block, as shown in Figure 2.

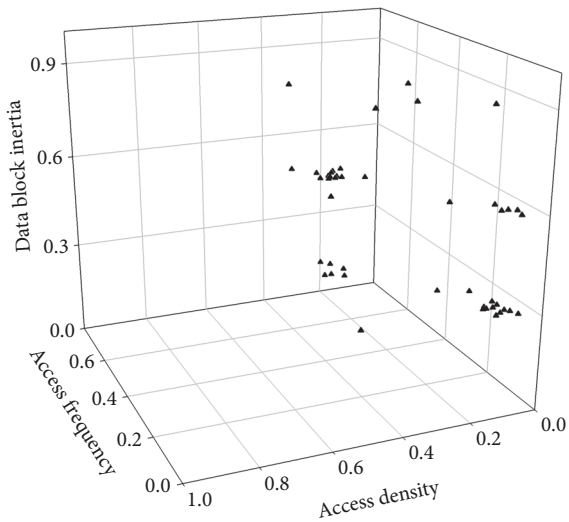

Figure 2: Sample data attribute distribution diagram.

In Figure 2, x, y and z denotes the access frequency, access density and data block inertia within unit time in data block respectively, and the data range is $[0, 1]$ after normalization. It can be seen from the figure that the access property has a certain extent of regularity. There is obvious aggregation phenomenon. In BP model, the excitation function of neuron is S function. The common sample data of two models is used to train the batch model. The weights and coefficients in BP network are adjusted adaptively. The minimum mean square error energy function is used to optimize network parameters. The error is as follows:

$$E = \frac{1}{2\sum_{l=1}^{L}\left(y_l - \widehat{y}_l\right)^2}, \tag{10}$$

where l is the number of samples and y_l and $\widehat{y}_l$ are the actual value and the predicted value, respectively. The gradient descent method is used to calculate the instantaneous gradient vector. The learning rate parameter in the reverse propagation process $\eta = 0.05$. In order to facilitate distinction, the proposed model is denoted as WNN and the benchmark model is BP neural network. The two models use same sample data to train and test. The training speed of the two predictive models is compared. Firstly, 300 sample data are randomly selected to train the two prediction models. The data preprocessing modes are same. The error of WNN and BP model is shown in Figure 3. The error accuracy of WNN is lower than the pre-set value after 57 iterations. BP is not lower than the preset value until after 217 iterations.

In addition to adjusting network weights, the WNN model can adjust the scaling factor and translation factor of wavelet function. It has better adaptability and elasticity than BP model and has more sensitive nonlinear fitting ability.

5.2. Comparison of Prediction Accuracy. In this section, we use the trained prediction model in previous section for identifying hot data block The main purpose is to compare prediction accuracy of the two models. The accuracy of wavelet neural network and BP is compared by selecting 10 actual measured data, as shown in Table 1. Table 1 lists the predicted values (average values) and the variance of predicted values generated by the two models.

The experimental results show that the prediction error of WNN is 1.6%, while the prediction error of BP is 3.1%. This indicates that WNN has better nonlinear fitting ability. In addition, the variance of the predicted value of WNN is 3.3%, while the variance of BP is 7.1%. This indicates that WNN prediction model has better fault tolerance capacity. This is because that BP neural network excitation function uses conventional S function. This excitation function is relatively smooth. It cannot quickly respond to those sample data which changes rapidly from wave peak to wave trough, resulting in that the trained model is not stable enough in real-time prediction.

5.3. Model Robustness. In this section, we can verify that WNN has multi-dimensional learning ability, train the learning with high frequency by adjusting scaling and translation factors in data dense area and train the learning with low frequency in data sparse area. So it has the ability to automatically adapt to sample data. Here 300 samples of sample data intensive area are selected to train the two prediction models. Then 10 testing data of previous section are used to verify prediction accuracy. Training and comparison is shown in Figure 4.

As shown in Figure 4, the WNN model reaches the preset error value after 56 iterations. The BP model is lower than the pre-set error after 75th iteration. Comparing Figure 4 and Figure 3, it can be found that the convergence rate of BP model is greatly improved. This is mainly due to the better

Table 1: Comparison of predicted results.

ID	measured value	WNN		BP	
		predicted	variance	predicted	variance
1	0.592	0.583	0.019	0.610	0.044
2	0.081	0.082	0.003	0.010	0.001
3	0.453	0.446	0.015	0.467	0.034
4	0.403	0.409	0.013	0.415	0.03
5	0.102	0.103	0.003	0.098	0.007
6	0.073	0.071	0.002	0.075	0.005
7	0.415	0.421	0.014	0.402	0.029
8	0.524	0.515	0.017	0.540	0.039
9	0.173	0.175	0.006	0.167	0.012
10	0.393	0.386	0.013	0.380	0.027

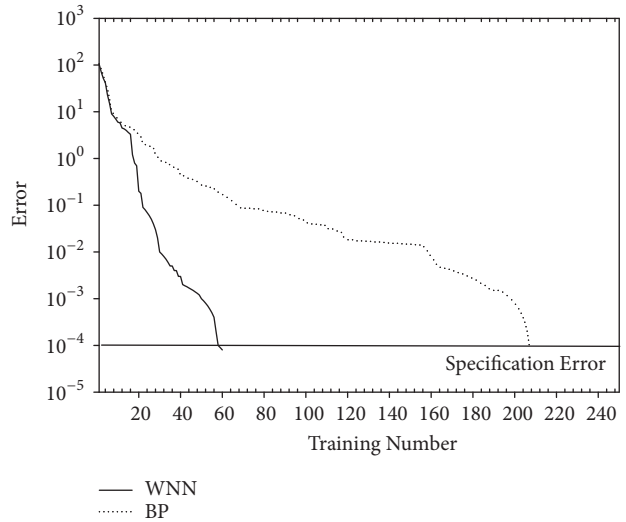

Figure 3: Training comparisons of WNN and BP.

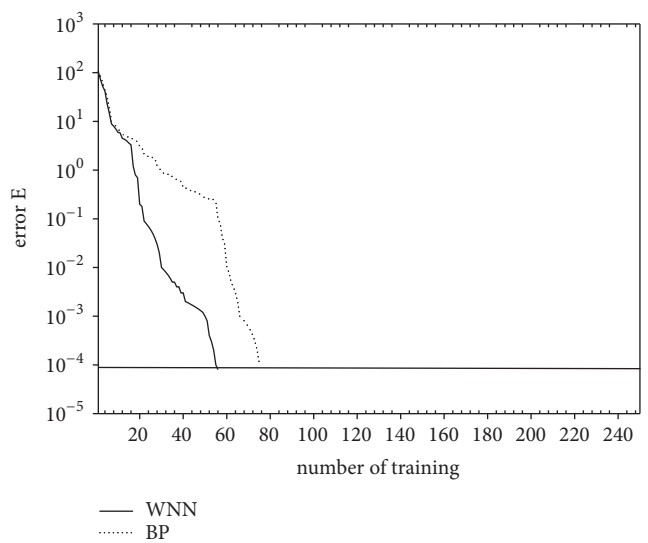

Figure 4: Training comparisons of dense sample data of WNN and BP.

Table 2: Comparison of predictive results of dense area data after training.

ID	Measured value	WNN		BP	
		predicted	variance	predicted	variance
1	0.592	0.581	0.019	0.580	0.020
2	0.081	0.083	0.003	0.079	0.003
3	0.453	0.447	0.015	0.463	0.016
4	0.403	0.406	0.013	0.395	0.014
5	0.102	0.104	0.003	0.104	0.004
6	0.073	0.074	0.002	0.071	0.002
7	0.415	0.422	0.014	0.424	0.015
8	0.524	0.517	0.017	0.513	0.018
9	0.173	0.174	0.006	0.177	0.006
10	0.393	0.387	0.013	0.385	0.013

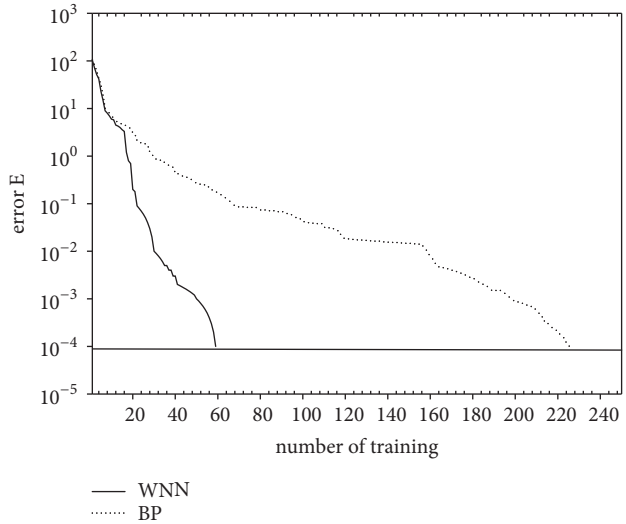

Figure 5: Training comparisons of sparse sample data of WNN and BP.

training effect of dense sample. The prediction accuracy comparison is shown in Table 2.

As shown in Table 2, the prediction error of WNN is basically the same, while the prediction error of BP is reduced to 2.1%. This indicates that BP model is better trained in the dense data area. So the prediction accuracy is improved. In addition, the predicted variance of BP model decreases from 7.1% in previous section to 3.5%, which is basically similar to the predicted variance of WNN. The above results show that the training effect of BP is improved and the prediction ability is enhanced in the area with dense sample data. In order to highlight the advantages that WNN has better adaptive ability to sample data, another 100 data are selected from the sparse area of sample data to retrain the two prediction models. Finally, the verification data is used to compare the prediction accuracy and stability.

As shown in Figure 5, when using sparse area sample data to train the prediction model, WNN converges into the pre-set error range at the 61st iteration, while BP model is just lower than the pre-set error at 225th iteration. This indicates that BP model has a poorer adaptive ability to sparse area sample data. The comparison of prediction accuracy is shown in Table 3.

In prediction accuracy, the predicted mean value and variance of WNN are kept stable. The prediction accuracy of BP is reduced from 3.1% to 3.3%. Especially, the variance of the predicted value is reduced to 9.7%. This indicates that the prediction robustness of BP is lower than that of WNN. The dependence on sample data is stronger. It indicates that the generalized learning ability of WNN prediction model is better than that of the prediction model constructed with BP neural network.

5.4. Comparison with Statistical Method. In literature [13], the identifying hot data block is used as a Data-block temperature model. This model called EESDC was applied to reducing data migration overhead and improving energy-efficient of large-scale streaming media storage systems. Note that we used the identifying hot data method as benchmark model, which belongs to the statistical model. To compare with EESDC, we further demonstrate that our proposed model is better than the traditional statistical models. This experiment

Table 3: Comparison of predicted results of sparse area data after training.

ID	Measured value	WNN		BP	
		predicted	variance	predicted	variance
1	0.592	0.584	0.019	0.612	0.057
2	0.081	0.083	0.003	0.078	0.007
3	0.453	0.444	0.015	0.468	0.044
4	0.403	0.408	0.013	0.390	0.036
5	0.102	0.104	0.003	0.105	0.010
6	0.073	0.070	0.002	0.071	0.007
7	0.415	0.422	0.014	0.429	0.040
8	0.524	0.514	0.017	0.507	0.047
9	0.173	0.175	0.006	0.179	0.017
10	0.393	0.385	0.013	0.380	0.035

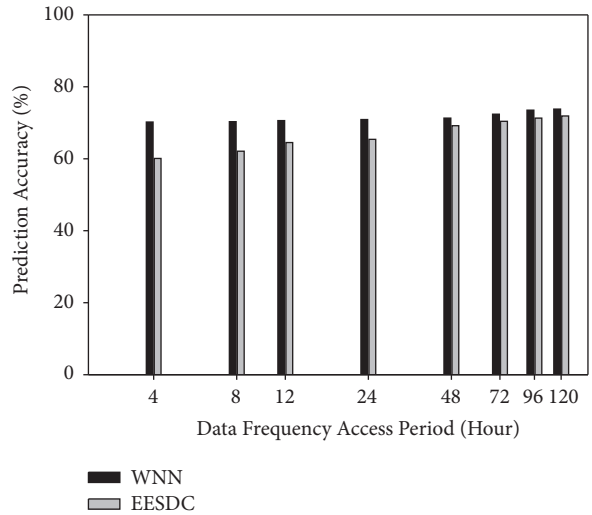

Figure 6: Prediction accuracy comparison of WNN and traditional statistical model.

is designed on a video learning website of campus. The hybrid storage system of website has large-scale stream video files for students. These video files have different hot degree for users. Such as some videos may be popular for one or two hours, some videos even keep hot for many days. Here we select different kind's videos to verify our model and benchmark model.

As shown in Figure 6, x axis represents the data block frequency period. As x equals 4, this means that this kind video files are accessed frequently in four hours. Here 100 file of each kind of stream video is selected and used two prediction models to identify the data block status. y axis shows the prediction accuracy. The WNN prediction model is stable for different kinds of video files. However, EESDC is not well in handling video files, whose hot degree change fast in short period. The prediction accuracy of EESDC gradually increases while increasing the frequency access period. This comparison shows that our proposed model overcomes the cumulative effect of traditional statistical methods and is suitable to prediction short-term variation data block.

6. Conclusion

To handle the large-scale data migration in distribution storage system and improve the performance of migration, a novel prediction model of data block status was provided to judge the heat degree of data block. We extracted the access features of data block and used those features as input vector to train the prediction model. The kernel of prediction model is wavelet neural network that has better capacity than other models. Additionally, we used the fuzzy rule to deal with the uncertain of sampling data. Compared with BP neural network, our proposed prediction model has better nonlinear fitting capacity due to the fact that wavelet neural network can learn with low frequency in sparse sample area and with high frequency in dense sample area. The experimental results show that our proposed prediction model has better generalization capacity and robustness than BP model and relevant model. In the future, we will integrate this prediction model into the data migration model and verify the prediction accuracy in the production environment.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

Ming Zhang (1983-), male, is with a Ph.D. degree and is a lecturer. His research interests include distributed intelligent computing and distributed multisensor data fusion system. Wei Chen (1963-), male, is with a Ph.D and is a Professor and a Doctoral supervisor. His major research interests include mobile communication, signal processing, and distributed system.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This project is supported by the Shandong Provincial Natural Science Foundation, China (nos. ZR2017MF050 and ZR2014FL008), Shandong Province Key Research and Development Program of China (nos. 2018GGX101005, 2017CXGC0701, and 2016GGX109001), and Project of Shandong Province Higher Educational Science and Technology Program (no. J17KA049).

References

[1] P.-Y. Yang, P.-Q. Jin, and L.-H. Yue, "A time-sensitive and efficient hybrid storage model involving SSD and HDD," *Jisuanji Xuebao/Chinese Journal of Computers*, vol. 35, no. 11, pp. 2294–2305, 2012.

[2] Y. Yamato, "Cloud storage application area of HDD–SSD hybrid storage, distributed storage, and HDD storage," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 11, no. 5, pp. 674-675, 2016.

[3] H. Dongmei, D. Yanling, and H. Qi, "Research on Ocean Large Data Migration Algorithm in Hybrid Cloud Storage," *Journal of Computer Research and Development*, vol. 51, no. 1, pp. 199–205, 2014.

[4] C. Zhiguang, N. Xiao et al., "A high-performance and reliable storage system for backing up SSDs with disks," *Journal of Computer Research and Development*, vol. 50, no. 1, pp. 80–89, 2013.

[5] S. Zhenlong, L. Qiong et al., "Design and Implementation of Hybrid Storage System for High Performance Computing," *Journal of Shanghai Jiaotong University*, vol. 47, no. 1, pp. 113–118, 2013.

[6] P. Wang and X. Liu, "Rapid convergence for telegraph systems with periodic boundary conditions," *Journal of Function Spaces*, vol. 2017, Article ID 1982568, 10 pages, 2017.

[7] L. H. Liu and Y. Z. Bai, "New oscillation criteria for second-order nonlinear neutral delay differential equations," *Journal of Computational and Applied Mathematics*, vol. 231, no. 2, pp. 657–663, 2009.

[8] M. Bohner, T. S. Hassan, and T. Li, "Fite-Hille-Wintner-type oscillation criteria for second-order half-linear dynamic equations with deviating arguments," *Indagationes Mathematicae*, vol. 29, no. 2, pp. 548–560, 2018.

[9] T. Li and Y. V. Rogovchenko, "Oscillation criteria for second-order superlinear Emden-Fowler neutral differential equations," *Monatshefte für Mathematik*, vol. 184, no. 3, pp. 489–500, 2017.

[10] J. Luo, L. Fan, Z. Li, and C. Tsu, "A new big data storage architecture with intrinsic search engines," *Neurocomputing*, vol. 181, pp. 147–152, 2016.

[11] T. Kgil, D. Roberts, and T. Mudge, "Improving NAND flash based disk caches," in *Proceedings of the ISCA 2008, 35th International Symposium on Computer Architecture*, pp. 327–338, China, June 2008.

[12] I. Koltsidas and S. D. Viglas, "Designing a Flash-Aware Two-Level Cache," in *Advances in Databases and Information Systems*, vol. 6909 of *Lecture Notes in Computer Science*, pp. 153–169, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[13] Y. Chai, Z. Du, D. A. Bader, and X. Qin, "Efficient data migration to conserve energy in streaming media storage systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 11, pp. 2081–2093, 2012.

[14] F.-J. Lin, Y.-C. Hung, and K.-C. Ruan, "An intelligent second-order sliding-mode control for an electric power steering system using a wavelet fuzzy neural network," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 6, pp. 1598–1611, 2014.

[15] T. Hieu, X. Zhang et al., "Passive-Islanding detection method using the wavelet packet transform in grid-connected photovoltaic system," *IEEE Transactions on Power Electronics*, vol. 31, no. 10, pp. 6955–6967, 2016.

[16] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 194–207, 2014.

[17] L. Quan, W. Xiaoyan et al., "Double Elite Coevolutionary Genetic Algorithm," *Journal of Software*, vol. 23, no. 4, pp. 765–775, 2014.

[18] Y. Liang and C. Nie, "The Optimization of Genetic Algorithm Configuration Parameter Generated by Coverage Table," *Chinese Journal of Computers*, vol. 35, no. 7, pp. 1522–1537, 2012.

[19] G. Casale, S. Kraft, and D. Krishnamurthy, "A Model of Storage I/O Performance Interference in Virtualized Systems," in *Proceedings of the 2011 31st International Conference on Distributed Computing Systems Workshops (ICDCS Workshops)*, pp. 34–39, Minneapolis, MN, USA, June 2011.