Research on Difference Elimination Method Between Small Sample Databases Based on Feature Extraction

Jin-hua Liu[1] and Fu-lian Zhong[2(✉)]

[1] Department of Professional and Continuing Education, XinYu University, XinYu 338031, China
[2] School of Mathematics and Computer Science, XinYu University, XinYu 338031, China
zhongfulian682@163.com

Abstract. The traditional method of eliminating the differences between small sample databases takes a long time and has a low accuracy. Therefore, a method of eliminating the differences between small sample databases based on feature extraction is designed. In order to realize the data communication between small sample databases, we construct the data retention mechanism of small sample databases, store the sample data safely, discretize the data attributes, sort the primary and secondary relationship of the sample data, select the optimal integration and sharing path of the sample data, cluster the sample data, and select the cluster head and relay node, Eliminate the differences between small sample databases. The experimental results show that compared with the traditional method, this design method shortens the time of eliminating the differences between small sample databases, and improves the accuracy of eliminating the differences.

Keyword: Small sample database · Data clustering · Data fusion · Difference elimination

1 Introduction

With the rapid development of Internet, a large number of users access the data and information resources in small sample database through Internet. The characteristics of small sample database are that there are differences in logic and physics. Using different information resources to operate each other in small sample database has become a hot research topic in database field. The difference elimination of small sample database is to access the database transparently in the computer network environment, support the operation between databases, and update, merge and sort the other database based on the support of database application tools. In order to eliminate the differences between small sample databases, it is necessary to shield the differences between databases distributed in different sites. In the network environment, the differences between databases have two aspects: on the one hand, the database hardware and software environment is different; On the other hand, there are differences in the database itself, including the differences caused by data semantics.

S. Wang et al. (Eds.): IoTCare 2021, LNICST 414, pp. 226–243, 2022.
https://doi.org/10.1007/978-3-030-94185-7_16

In the anti underwater cooperative communication system, nodes can send data, and nodes can share the uplink and downlink frequencies, so as to realize the reuse of database resources, Eliminate differences between databases. The domestic research on the methods of eliminating the differences between databases has also made great progress. This paper proposes a method of eliminating the differences between small sample databases based on KL divergence. KL divergence is used to analyze the degree of differences between databases, and the training data is used to generate a classifier to classify the data in the database, and the results are combined to form a data space, which is measured by complementary information entropy. This paper classifies the uncertain information in the database, judges the differences between databases, and uses the basic criteria of spatial differences to eliminate the differences between databases by simulating data sets. However, the traditional method of eliminating the differences between small sample databases takes a long time and can not eliminate the differences between databases correctly.

In order to solve the shortcomings of traditional methods, this paper designs a method to eliminate the differences between small sample databases based on feature extraction to ensure the efficiency and accuracy of eliminating the differences between databases. The main work of this paper is as follows:

(1) The data retention mechanism of small sample database is constructed to store the sample data safely, and the data attributes are discretized to order the primary and secondary relationship of the sample data.
(2) Select the optimal integration and sharing path of the sample data. On the basis of clustering the sample data, select the cluster head and the relay node to realize the data communication among small sample databases.
(3) Extract model training data by features, and eliminate differences among small sample databases by fusion of communication sample data.

2 Design of Difference Elimination Method Between Small Sample Databases Based on Feature Extraction

2.1 Building Data Retention Mechanism of Small Sample Database

Based on the principle of cryptography, the sample data is encoded, and the data retention mechanism of small sample database is constructed. In the sample data storage process, the random dynamic time-varying factors are added to control the data storage state in real time. The data retention mechanism is divided into two parts: upload and download, and pseudo transformation. In the upload and download process, the implementation of cryptography coding based on cryptography principle is used to dynamically repair the network data, and tolerate some invalid network nodes [1]. For the data upload stage, the sample data is segmented to get multiple data blocks of the same size. The random linear coding method is used to store the coded data blocks to nodes. For the data read stage, the coding blocks are downloaded randomly to merge the sample data. For the stage of mimicry transformation, the less functional storage regeneration code is used to screen out the network nodes that need to be repaired,

control the repair bandwidth, code randomly and construct new data blocks to obtain the data nodes after mimicry transformation, and replace the original data content to ensure the integrity of sample data. The calculation formula of repair bandwidth Q of storage regeneration code is as follows:

$$Q = \frac{\xi\beta(K-1)}{L(K-L)} \tag{1}$$

Among them, L is the number of fixed size original blocks, β is the number of coding blocks to store the regeneration code, K is the number of network data nodes, and ξ is the repair bandwidth cost [2]. The amount of data stored in the network node is taken as the minimum storage extremum point of the regeneration code. After the node is repaired, the pseudo transformation parameters are negotiated to update the storage content of the sample data, so as to ensure that the stored data is always in the coding state, so as to download all the data on the network node. The data coding network structure is shown in Fig. 1.

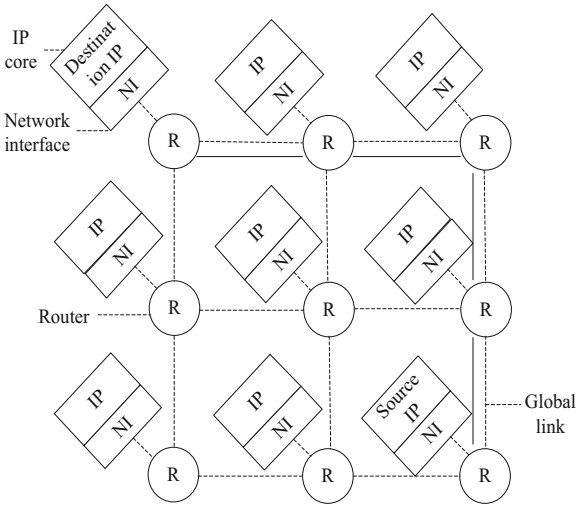

Fig. 1. Data coding network structure

As shown in Fig. 1, the data coding communication structure is composed of link, IP core, router and IP core. The network interface is used to connect the router and IP core, transmit the source IP packets to the destination IP, and use the router to determine the transmission direction. The router is composed of control logic, buffer and crossbar. The control logic is used to arbitrate the channel, calculate the route, and determine the transmission direction The routing decision supports the authorization of the sample data, which is then transmitted to the next router through the crossbar until it reaches the packet destination.

In order to reduce the security threat of the database network environment, a data retention mechanism is constructed after data access and storage by encoding. On the basis of analyzing the repeated organization form of encoded data, the encoded data files are copied for many times, and the file copies are allocated so that they can be transferred to different storage nodes. At the same time, the file is divided into several subfiles, the file is encoded into redundant data blocks, and the low-density check code is configured for each data block. The original file of the database is reconstructed through any data block in the redundant data, so as to avoid the leakage of sample data and part of the original file when there is network attack, so as to ensure the security of the sample data [3]. The implementation process of the retention mechanism is shown in Fig. 2.

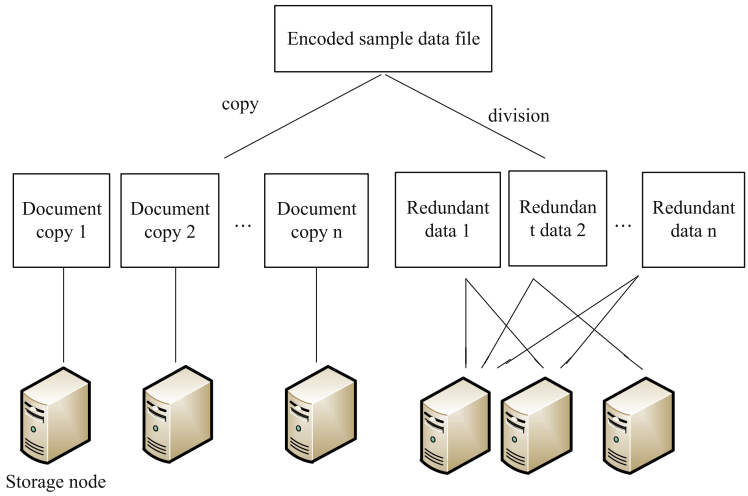

Fig. 2. Implementation process of data retention mechanism

As shown in Fig. 2, combined with storage technology and network communication technology, the storage nodes distributed in small sample databases are connected to establish a logical unified storage server. The expression of redundancy process of coded data is as follows

$$G_{n\times l} = \begin{bmatrix} I_{k\times k} \\ \phi_{k\times(n-k)} \end{bmatrix} O_{k\times l} \tag{2}$$

Among them, $G_{n\times l}$ is the n redundant data blocks divided by k block file, $O_{k\times l}$ is the original data column vector composed of k original data blocks, l is the number of original data block columns, $I_{k\times k}$ is the $k \times k$ order identity matrix, and $\phi_{k\times(n-k)}$ is the

$k \times (n - k)$ order Cauchy matrix. In the segmented redundant data, the formation vector $G'_{k\times k}$ of k redundant data blocks is arbitrarily extracted:

$$G'_{k\times l} = U_{k\times k} G_{n\times l} \tag{3}$$

where $U_{k\times k}$ is a Vandermonde matrix of order $k \times k$. Then the formula for obtaining the original data of the data retention mechanism is as follows:

$$O_{k\times l} = U^{-1}_{k\times k} G'_{k\times l} \tag{4}$$

Through the formula (4) to complete the decoding and decryption operation, restore the original file of the small sample database, restore the original data content. So far, the construction of data retention mechanism of small sample database is completed.

2.2 Preprocessing Sample Data of Small Sample Database

Preprocess the sample data in the database, discretize the continuous numerical attributes, and sort the primary and secondary relationship of the sample data. The maximum minimum normalization formula is used to transform the original data linearly

$$V = \beta \frac{(L-M)}{(M-N)} \tag{5}$$

Among them, L is the data value, M and N are the maximum and minimum values of the original data of the same attribute, β is the mapping interval, and V is the mapped value of the original data. Using the clustering function of neural network, using the network center instead of the continuous value of the original data, clustering the continuous attributes of the mapping data, transforming the data attributes into discrete values, ensuring the relative relationship of the attributes, displaying regular rules, reducing the number of values of the same attribute data [4]. The radial basis function of Multivariable Interpolation is used in the neural network, and the three-layer forward neural network is selected as the typical structure of the neural network. In the middle layer, the attribute features extracted from the input layer are transformed to make the data category closer to the center of the network. The specific training process is shown in Fig. 3.

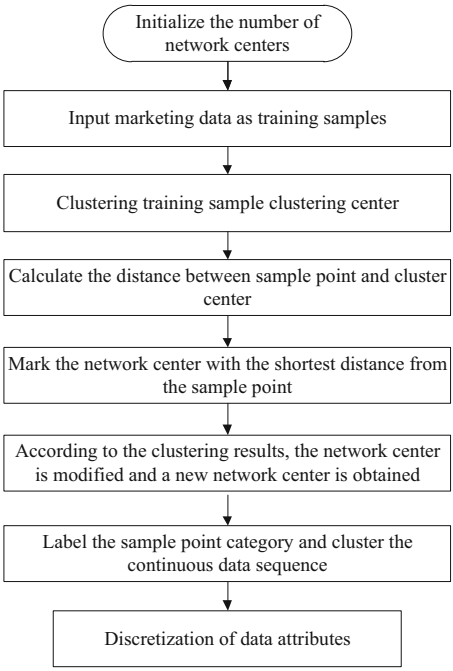

Fig. 3. Training process of database sample data

Suppose that the output value of the i neuron is x_i, the sample point of the j network center is G_j, and the middle layer neuron of the j network center is T_j, then the modified new network center B is:

$$B = \frac{\sum x_i}{G_j} \quad \forall x_i \in T_j \tag{6}$$

The mapping data sample points are divided to the new network center B, and the network center set is used as the value range instead of the sample point value, so as to eliminate the different dimensions of each dimension data, so as to find the change rule between the database sample data. Using sprint classification algorithm, sort the primary and secondary relationship of database sample data. In the sample data, select the attribute with the highest priority as the root, provide the preprocessed attribute set, search the commonness from the massive sample data, make a series of decisions, classify the sample data, split the decision tree node, and then split the sample data attribute, so that the attribute is accurately associated with the child node, and get the attribute value segmentation training set [5]. If the number of data set categories is m and the number of data set categories is the number of leaf node categories, the calculation formula of splitting parameter F is as follows:

$$F = 1 - \sum_{I}^{m} p_I^2 \tag{7}$$

where p_I is the relative frequency of data set category I. In the training set, a data node is selected to represent the data of the training set as the internal node of the decision tree. The branch structure of logical judgment is used as the edge of the decision tree, and the data attributes are associated with the root node of the decision tree to construct a multi tree decision tree. When all marketing data belong to the same category, the class label is used to define the leaf node. When the sample data does not belong to the same category, the data attribute is measured according to the information entropy, and the data in the original attribute set is deleted. When the candidate set is empty, the leaf node is returned and marked as a common category [6–8]. For different types of sample data, the calculation formula of information entropy Q is as follows:

$$Q = -\sum_{I=1}^{\mathrm{m}} \frac{|C_I|}{|\xi|} \log_2 \left(\frac{|C_I|}{|\xi|} \right) \tag{8}$$

where ξ is the given training set of decision tree and C_I is the set of data set belonging to class I objects. The training set ξ is classified according to the characteristics of attributes to obtain multiple different objects. The weighted sum of the information entropy Q is carried out through the partition entropy to calculate the information gain attributes of marketing data

$$A = -\sum_{I=1}^{\phi} \left(\frac{|\zeta|}{|D|} \times Q \right) \tag{9}$$

Among them, A is the gain of marketing data information, ϕ is the attribute characteristic quantity of training set, D is the information quantity required for training set classification, and ζ is the information quantity required by training set partition. In the attribute set, select the attribute with the highest information gain A, mark the leaf node, get a score of the highest information gain attribute, and make the training set subset elements meet the score value. When the categories are the same at the node, the remaining attributes cannot be subdivided, or the given score value has no data, the class label is created, the decision tree partition is terminated, and the classification of database sample data is completed. At this point, the preprocessing of database sample data is completed.

2.3 Choose the Best Integration and Sharing Path of Sample Data

For the preprocessed sample data, select the optimal integration and sharing path, that is, the link path of the sample data, and add and modify the sample data. Firstly, the associated semantics of the sample data is standardized, and the frequency of different semantic query words is counted, from which the core query words are determined, and the four attributes of the core query words are determined, and then the similarity between different sample data and the words is calculated. If the similarity distance formula is adopted, the calculation formula of similarity Q is as follows:

$$Q = \left(\sum_{i=1}^{4} \left(a_k - a_j \right)^k \right)^{\frac{1}{k}} \tag{10}$$

Among them, i is the four attributes of the core query word, a is the associated data of the database sample data, and k, j is the visual spatial dimension and the quantitative value of similarity distance of the data, where j is 2 or 3, when k is 1, it represents the real distance between the core query word and the spatial latitude, and when $k \neq 1$, it is the exact distance, representing the sum of absolute wheelbase on the spatial dimension [9]. By transforming formula (10), we get the optimal path S of transformation infimum distance in visual space dimension:

$$S = \frac{1}{c^Q} \tag{11}$$

where, c is the frequency of database sample data and core query words. When $0 < c < 1$, the optimal path S value is between (0, 1). The smaller c, the closer S value is to 0. When f $c > 1$, the smaller c, the closer S value is to 1. According to the frequency of different sample data, the values of k, j and K are determined to make S reach the limit value, and the final path of data transformation is obtained. By using this path, the associated data of sample data is transformed, and the optimal mode of resource integration is obtained. Remove the query words that are not related to the associated semantics, improve the service function of the associated data, and according to the unified specification of the small sample database, map the associated data and resource ontology to form the integrated data of the small sample database [10, 11]. Calculate the sample data sharing path, let the other database download ciphertext, decrypt to obtain the complete sample data. Multiple wireless links are set up in the transmission range. Minimizing the average energy consumption of data transmission is taken as the target of channel selection. The time is divided into fixed slot periods, so that the database of the opposite party can generate sample data at a given rate, fix the transmission power of the provider, and detect the channel every other period. The channel detection process is shown in Fig. 4.

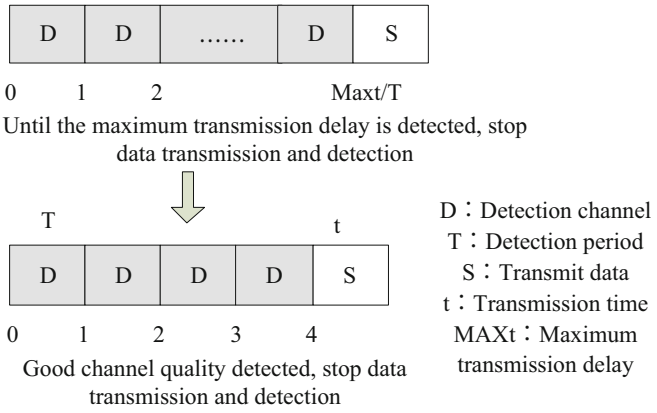

Fig. 4. One round channel detection process

The detection duration of the control channel is far less than the detection period, so that the transmission time of the database sample data can meet the requirements of the transmission delay. All the detected good channels are filtered and the channel transmission rate R is calculated:

$$R = H \log_2 \left(1 + \frac{\phi P}{\xi H}\right) \tag{12}$$

Among them, ϕ is channel gain, P is power of energy data transmission, H is channel bandwidth, and ξ is channel power spectral density. According to formula (12), transmission rate is proportional to channel gain, so real-time detection of transmission rate of each channel, capture the time with large channel gain value, obtain the channel of optimal transmission time, thus improving the transmission amount of energy data and reducing the average energy consumption of unit data. The channel is used as a sample data sharing link, and the ciphertext data is sent to the shared link, and the block chain is broadcast throughout the network, so that it is transmitted from the block head of the blockchain to the public key address of the other database. When the user receives the ciphertext, the private key (B, c) matching the public key (B, a) is selected. The formula of G of matching algorithm is:

$$G = \frac{\beta \prod\limits_{\mathrm{i} \in N} \left(\mathrm{e}(B, \mathrm{a}) \mathrm{e}(B, \mathrm{c})\right)^{\varpi_{\mathrm{i}}}}{\mathrm{e}(B, \mathrm{a})} \tag{13}$$

Among them, β is the mapping order of the keyword index tree, N is the set of ciphertext index structures, i is the smallest subset of index structures, ϖ_{i} is the subset hidden matrix, $e(B, \mathrm{a})$ and $e(B, \mathrm{c})$ are the attribute value sets of public key and private key respectively. After the information data is downloaded, the energy data encrypted by the public key is decrypted by using the matched private key, so that the other

database can interpret the shared sample data. So far, the optimal integration and sharing path of sample data is selected.

2.4 Sample Data Transmitted by Clustering Database

The cluster label of the small sample database is consistent, the cluster center of the database is output, and the transmitted sample data is clustered. Firstly, the eigenvalue of database is calculated, and each data object is regarded as a cluster class by using the classification hierarchy algorithm. Through the iterative cycle, the adjacent cluster classes are merged until all the cluster classes are merged into one cluster class. According to the specific similarity function and neighborhood threshold, the distribution of the sample data in the whole space is calculated, the known data sets are decomposed hierarchically, the high-dimensional observation sample data sets are found, the maximum eigenvalue and minimum eigenvalue of the big data are calculated, and the intrinsic geometric structure of the high-dimensional sample data is obtained. For a given small sample database, the support degree L_i of itemset i is defined:

$$L_i = \frac{|k_i|}{l} \tag{14}$$

Among them, k_i is the number of occurrence of item set i in small sample database and l is the number of local databases in small sample database. L_i is the characteristic information of small sample database. If the frequent item sets of two small sample databases have more common elements, the similarity between small sample databases is relatively large. Select the database with close support degree S_i of item set, calculate the similarity of feature value, use iterative relocation technology, redistribute the category of data objects, and determine the probability of classification of small sample database, The concept of link in clustering process of small sample database is obtained. Using ADMM algorithm, the distinguishing features are selected, the local geometry is linked to the database, the cross-correlation matrix is used to update the sampling probability of each small sample database, the cluster quality value of data samples is transmitted, the frequent sequence pattern in the small sample database is mined, the sequence prefix in the small sample database is analyzed, and the frequent term set is projected one by one. The variable data of feature values are standardized by using the same dimension minner distance, and the data close to the distance are classified into the same category. The sample data in local database is processed in a standard way. The two-dimensional logical table of database is used to record all fields and analyze the hidden structure of sample data, so that all hidden variables can correspond with data clustering, and ensure the integrity and consistency of the label of sample data cluster.

Input the small sample database for clustering analysis, use genetic algorithm to get the sample data of each database, get the sample data and the cluster to which the sample belongs, get the sum of the dimensions of the sample data of different clusters, get the center value of genetic operation, get the distance between the sample data and K clusters, and collect the small sample database with high similarity. The distance

between clusters is calculated. The data set with multiple sample data is divided into k classes of K-means clustering algorithm. The distance from the center of each cluster to the center of the whole domain is taken as the distance between clusters to calculate the distance within clusters. If the distance within the cluster is smaller, the closer the distance of the same cluster data is, the better the clustering effect is; If the distance within the cluster is larger, the distance between the data in the same cluster is larger, and the clustering cohesion is smaller. According to the classification utility of the clustering results, a new node is created as the clustering center of the small sample database, which divides the clustering groups, processes the low dimensional sample data by clustering, constructs a multi-dimensional low dimensional space, locally embeds the dimension of the sample data, reduces the dimension of the original sample data, realizes the dimension reduction of the sample data, and obtains the low dimensional representation of the small sample database.

On this basis, the similarity between data samples is taken as the objective function of the original data samples, and the sample data is embedded into the high-dimensional space to reveal the potential structure of the sample data. The small sample database is decomposed by matrix decomposition and sparse coding, and the matrix factors are non negative constrained to ensure the neighborhood of each sample data, which conforms to the non negative performance of visual data [12]. In each iteration of the small sample database, the weight of clustering partition is used to measure the quality of clustering partition, the sampling probability distribution of data samples is calculated, new sample data is randomly sampled, new K-means clustering partition is generated, and multiple clustering results are gathered, The small sample database after clustering is obtained. So far, the clustering of sample data in database transmission is completed.

2.5 Select Cluster Head and Relay Node for Sample Data Clustering

For the clustered database sample data, the cluster head and relay node are selected to realize the data communication between small sample databases.

The energy consumption of cluster head node includes collecting, fusing and forwarding the data of the node members in the cluster, and the energy consumption is large. The selection of cluster head is an important factor to ensure the elimination of differences between databases. Select a relay node in each cluster member, use the cluster head to collect the data from the members, analyze and process the sample data, and send the processed data to the relay node. The relay node is mainly responsible for data transmission, sharing the communication work between the cluster head and the base station. Cluster heads consume more energy when they receive their members' energy. When the cluster head communicates with the base station, the energy consumption is greatly reduced, and the relay node can be used as the relay data transmission. When communicating with the remote base station, the energy consumption of other parts is reduced, which plays a role in balancing the network energy load [13].

When the cluster head is selected, the topology of the cluster can be determined, and the relay node can be selected according to the application requirements in the

cluster. The distance between the relay node and the cluster head and the distance between the relay node and the base station can be calculated respectively. According to the minimum value of the sum of distances, the relay node is selected. After selecting the cluster head and relay node, the energy consumption of cluster members transmitting data to the cluster head and the energy consumption of the cluster head sending data to the relay node and the relay node sending data to the base station are calculated respectively to provide energy for data communication between small sample databases. So far, the selection of cluster head and relay node is completed.

2.6 Eliminating Differences Between Small Sample Databases Based on Feature Extraction

In the communication process of small sample database, the feature extraction model is used to fuse sample data, and then the differences between small sample databases are eliminated.

Firstly, backup the differences between small sample databases. The process is: to block data files, generate verification sum data files, search for matching blocks in small sample database according to the length of data files and data file blocks, obtain differential data files according to search results, and transfer unmatched data blocks to backup center. Assume that data blocks need to increase the number of shared blocks in the backup center, Then, you need to assign a unique block number to get the instruction file. The backup center data file difference, data file differential backup process is: according to the instruction file, get the modified data file, the backup center obtains the data file sent by the client, reconstruct a data block, and construct the final instruction file according to the check and the length of the modified data block, Complete the differential backup of the data blocks in the small sample mode.

After the backup of the differences between small sample databases, the feature extraction model is used to fuse the node data in the database, and the training of the feature extraction model is completed. The difference elimination between small sample databases is realized by using data fusion.

The feature extraction model uses machine learning convolution neural network to make the output number of features equal to the number of convolution cores. The convolution kernel of fixed number is used to convolute the sample data to ensure convolution extraction features. Only the cognitive area of convolution core is included. In the convolution process, the edge features of the bottom layer are extracted first, and then convolution operation is used, Further, the feature extraction of the underlying features is carried out. After repeated operation, the high-level features with abstract significance are obtained.

In order to control the sample data output accurately, the latter layer of neurons can be connected with the local features of the neural network front layer, ensure that all connection information is summarized in the high-level network, strengthen the relevance of the database sample data, reduce the network parameters of local connection, and obtain the global information of the sample data. During the convolution core movement, the data can be obtained, The inner parameters of convolution core are fixed, and the convolution core of convolution layer is input by weight sharing method,

so as to reduce the parameter of neural network. The convolution expression formula of feature extraction model is as follows:

$$A_{n,m} = \sum_{v=1}^{V} \sum_{u=1}^{U} w_{v,u} \times a_{i+v,j+u} + b \tag{15}$$

Among them: $A_{n,m}$ is the characteristic element of the convolution output of the cognitive region of the sample data in row n and column m; $w_{v,u}$ is the convolution kernel weight parameter of the machine learning neural network in row v and column u; V is the width of the convolution kernel of the neural network; U is the height of the convolution kernel of the neural network; $a_{i+v,j+u}$ is the correlation semantics of the $i+v$ and $j+u$ columns of the sample data; i and j are the i and j columns of the convolution kernel sliding in the cognitive region.

After the convolution layer is completed, the pooling layer is used to retain the main features of the output feature elements, enhance the anti-interference of the sample data features, and retain more feature data. Through the activation layer, the feature data learned by the machine is mapped to the label space, and the feature data of the small sample database is received comprehensively. The cluster head sends the trained feature data to the terminal node, fuses the data in the database, and sends the fused data to the cluster head node, and classifies the data. Data fusion is used to eliminate the differences between databases, which can reduce the amount of data and reduce the energy consumption in the process of database difference elimination, So as to realize the data collection, fusion and transmission between small sample databases.

So far, we have completed the design of the difference elimination method between small sample databases based on feature extraction.

3 Experiment and Analysis

The design method is compared with the two groups of traditional methods to eliminate the differences between small sample databases, and the efficiency and accuracy of eliminating the differences between small sample databases are compared.

3.1 Experimental Preparation

The test environment of the experiment is: 2G memory client, OS is Windows 2008 R2 500G disk, CPU is IntelCoreTM 2 Duo17500, Microsoft SQL 2008 R2 is installed in the client, and database is created in the client, which is developed by object-oriented Java language. The created small sample database includes Oracle, SOL Server2000, SQLServer2005, MYSOL, Access, DB2, XML, COIL20 UMIST, FERET, PIE, BA, ISOLET1. Small sample databases are of the same type and are distinguished by different table spaces. The specific distribution is shown in Table 1.

Table 1. Deployment of small sample database

Database type	Database distribution	Database name
Oracle	16.200.11	TIAGT
SOLServer2000	16.200.11	HAL
sQLServer2005	16.200.11	LE
MySOL	16.200.11	UAIDLL
Access	16.111.25	UIPJ
DB2	16.111.25	UWTAHJ
XML	16.111.46	nlabdcl
COIL20	16.110.36	nlabdc2
UMIST	16.111.18	simu2
FERET	16.200.34	VMcine
PIE	16.110.25	CZGT
BA	16.110.25	VRDisp
ISOLET1	16.110.17	GIS

A distributed parallel computing environment is built. Five computers are set up in the LAN to form a Ha-doop cluster. One computer is defined as a Master node, and the other computer is defined as a Slave node. All computers have 2 GB memory and 3.20 GHz. Thirteen types of databases and one FTP file server are distributed in the LAN. The sample data of the database is shown in Table 2.

Table 2. Database summary sample data

Database type	Number of samples	Characteristic number	Number of categories
Oracle	1448	1022	22
SOLServer2000	571	645	291
sQLServer2005	1470	1290	281
MySOL	1422	1021	278
Access	1407	325	86
DB2	152	602	20
XML	1491	1217	291
COIL20	114	1008	129
UMIST	1091	437	198
FERET	902	572	116
PIE	1210	209	402
BA	1002	319	273
ISOLET1	1472	1934	73

3.2 Experimental Result

Experimental Results on the Efficiency of Eliminating Differences Between Databases

The design method and two traditional methods are used to record the time cost of difference elimination between small sample databases. Set the number of iterations to 20, change the network bandwidth of the LAN, and the comparison results of elimination efficiency are shown in Fig. 5.

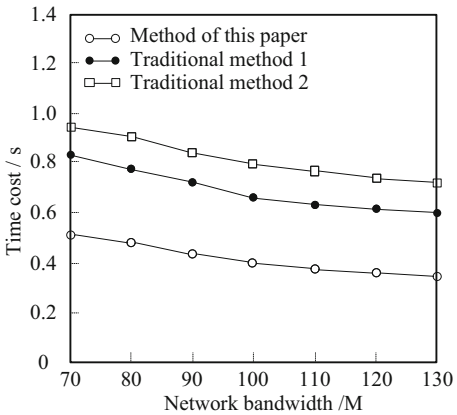

Fig. 5. Elimination time under different network bandwidth

Set the network bandwidth of the local area network to 100 m, change the number of iterations, and the comparison results of the elimination efficiency of the three groups of methods are shown in Fig. 6.

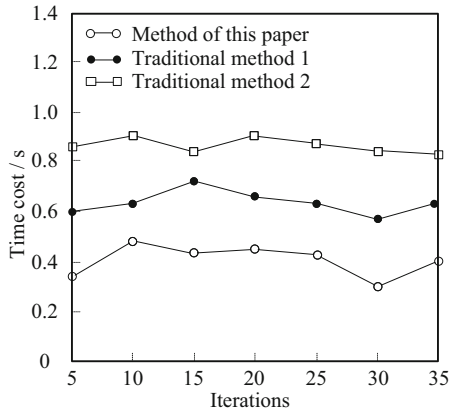

Fig. 6. Elimination time under different iterations

As can be seen from Fig. 6, the elimination time is always less than 0.5 s after the application of the method in this paper. After the application of traditional method 1, the elimination time was between 0.6 s and 0.75 s. After the application of traditional method 2, the elimination time was always above 0.8 s. In summary, under different network bandwidths and different iteration times, it takes time to eliminate the differences among the small sample databases of the method in this paper. It is obviously lower than the two sets of traditional methods, which proves that it can improve the efficiency of eliminating the differences between databases.

Experimental Results of the Accuracy of Eliminating the Difference Between Databases

The design method and two traditional methods are used to test the correctness of difference elimination between small sample databases. Set the number of iterations to 20, change the network bandwidth of the LAN, and eliminate the error rate. The comparison result is shown in Fig. 7.

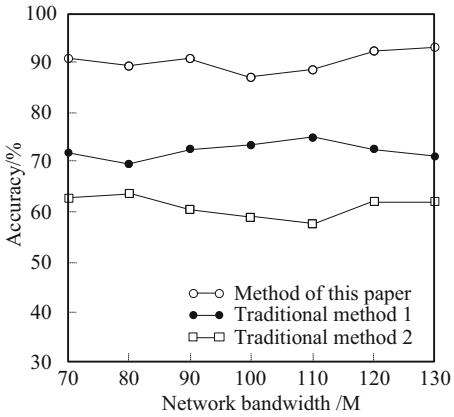

Fig. 7. Elimination accuracy under different network bandwidth

Set the network bandwidth of the local area network to 100 m, change the number of iterations, and the comparison results of the elimination accuracy of the three methods are shown in Fig. 8.

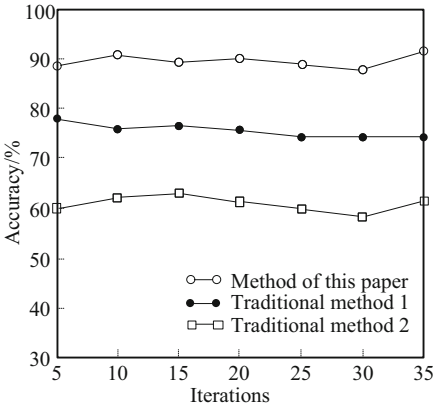

Fig. 8. Accuracy of elimination under different iterations

As can be seen from Fig. 8, under different network bandwidths and iterations, the accuracy of difference elimination among small sample databases is around 90% after the application of the method in this paper. The accuracy of traditional method 1 is less than 80%, and that of traditional method 2 is less than 65%. Therefore, the accuracy of the proposed method in eliminating the differences between small sample databases is significantly higher than that of the two groups of traditional methods, which proves that the proposed method can correctly eliminate the differences between databases.

4 Conclusion

This paper designs a method to eliminate the differences between small sample databases based on feature extraction. After securely storing the sample data, the data attributes are discretized to sort the primary and secondary relationships of the sample data. Then select the optimal integration and sharing path of sample data. On the basis of clustering sample data, select the cluster head and the relay node to realize the data communication among small sample databases. Then feature extraction model training data is used to eliminate the differences among small sample databases by fusion of communication sample data. This method gives full play to the technical advantages of feature extraction, shortens the time consuming of eliminating differences among small sample databases, and fully guarantees the correctness of eliminating differences among small sample databases. However, there are still some shortcomings in this study. In future studies, subject analysis and data mining functions will be added to the database to provide favorable conditions for data integration of small sample databases.

References

1. Guo, Y., Zuo, J.: DAO pattern database elimination elimination simulation based on big data analysis. Comput. Simul. **36**(12), 336–340 (2019)
2. Zhang, J., Wang, R., Jiang, X., et al.: Research and implementation of synchronization technology of heterogeneous database. Software Eng. **24**(1), 6–9+5 (2021)
3. She, J., Guo, Y.: Design of virtual database resource reorganization system based on multimedia technology. Modern Electron. Technique **44**(2), 86–90 (2021)
4. Licheng, L.I.U., Yifan, X.U., Guicai, X.I.E., et al.: Outlier detection and semantic disambiguation of JSON document for NoSQL database. Comput. Sci. **48**(2), 93–99 (2021)
5. Shen, D., Yang, G.: Heterogeneous database integration middleware system for privacy protection. Comput. Technol. Dev. **30**(1), 99–105 (2020)
6. Liu, S., Li, Z., Zhang, Y., Cheng, X.: Introduction of key problems in long-distance learning and training. Mob. Networks Appl. **24**(1), 1–4 (2018). https://doi.org/10.1007/s11036-018-1136-6
7. Liu, S., Sun, G., Fu, W.: e-Learning, e-Education, and Online Training, pp. 1–386. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63955-6
8. Xiong, X., Peng, X., Cao, X.: Research on Oracle database heterogeneous resource integration method based on improved ORM. Electron. Des. Eng. **28**(21), 38–41+46 (2020)
9. Zhi, H., Tao, L., Yao, S., et al.: Research on synchronization strategy of small heterogeneous database based on Json. J. Meteorol. Res. Appl. **41**(1), 48–53 (2020)
10. Xiao, G.: Research and implementation of heterogeneous database update synchronization. Software Guide **18**(10), 182–185 (2019)
11. Liu, S., Liu, X., Wang, S., Muhammad, K.: Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. Neural Comput. Appl. **33**(4), 1055–1065 (2021). https://doi.org/10.1007/s00521-020-05021-3
12. Xiong, H., Xu, D.: JSON based heterogeneous database integration model. Digital Technol. Appl. **38**(10), 33–35 (2020)
13. Li, R., Ren, Z., Huang, G., et al.: Design and implementation of heterogeneous architecture for database query acceleration. Comput. Eng. Sci. **42**(12), 2169–2178 (2020)