Data-parallel clustering algorithm based on mutual information mining of joint condition

Changjiang Huang
Sontan College Guangzhou University, Guangdong Guangzhou, 511370, China
33946676@qq.com

Abstract. In order to improve the reliability service ability of the cloud storage database, the data parallel clustering process is carried out, and the data parallel clustering algorithm based on the mutual information mining of the joint condition is proposed. A large data configuration structure model of a cloud environment virtual resource is constructed, data compression and characteristic reconstruction are carried out by adopting an online dictionary learning method, a regression analysis of the cloud environment virtual resource configuration data and a point cloud structure recombination are carried out in combination with a non-linear statistical sequence analysis method, the parallel characteristic scheduling of the large data of the virtual resources of the cloud environment is realized, the mutual information feature quantity of the joint condition is mined, the characteristic quantity of the mining is subjected to the characteristic filtering and the attribute set merging processing by adopting the fuzzy C-means clustering algorithm, by using the self-adaptive optimization algorithm, the automatic retrieval of the fuzzy clustering center is carried out, and the parallel clustering optimization of the large data is realized. The simulation results show that the classification performance of the large data clustering of the cloud environment virtual resources is good, the property classification and fusion capability is high, and the error rate is lower.

1. Introduction
With the development of cloud computing and large data information processing technology, the cloud storage technology is used for virtual resource storage, the parallel storage and information management capability of the data is improved, constructing a feature extraction and parallel mining model of the cloud storage virtual resource large data, carrying out feature extraction and mining on the large data of the cloud storage virtual resource by using a joint association rule mining method, and constructing a parallel clustering model of large data, In order to improve the retrieval and information management ability of the cloud storage virtual resource big data, the relevant data clustering algorithm research has been greatly concerned by people [1].

The clustering processing of cloud storage virtual resource big data is based on the feature extraction and classification identification of data, and the statistical feature quantity of cloud storage virtual resource big data is extracted, and the relevant classification algorithm is adopted for clustering analysis. In the traditional method, the big data clustering analysis method of cloud environment virtual resources mainly includes fuzzy C mean clustering method, K mean clustering method, grid
clustering method and irregular triangular network clustering method [2], the detection statistical analysis model of cloud environment virtual resources is constructed, and the correlation rule attribute set of extraction data is realized to realize the optimization clustering of cloud environment virtual resource big data clustering. In the document [3], the parallel clustering algorithm of cloud environment virtual resource configuration data based on depth learning is proposed in the document [4], and the phase space reconstruction method is adopted to reconstructing the feature reconstruction of the cloud environment virtual resource configuration data of cloud environment virtual resource configuration data, and the method is based on cloud environment virtual resource configuration data, but the calculation cost of the calculation cost of cloud environment is not good. In the literature [4], the method of virtual resource allocation data clustering for cloud environment based on association rule mining and fuzzy C mean value is proposed. The feature extraction is carried out by using principal component analysis to optimize clustering of cloud environment virtual resource data. Aiming at the above problems, this paper presents a parallel clustering algorithm based on joint conditional information mining. Construct big data configuration structure model of cloud environment virtual resources, adopt online dictionary learning method for data compression and feature reconstruction, realize parallel feature scheduling of big data of cloud environment virtual resources, mining joint condition mutual information feature quantity, use fuzzy C mean clustering algorithm for feature selection and attribute collection processing for the feature quantity of mining, and carry out automatic retrieval of fuzzy clustering center in combination with adaptive optimization algorithm to realize data clustering optimization. Finally, simulation experiment analysis is carried out, which shows the advantages of the method in improving data clustering capability.

2. Analysis of Storage configuration structure and characteristics of data

2.1. Big data configuration structure Model

In order to realize that optimization design of the parallel clustering algorithm for the virtual resource allocation data of the cloud environment, the data storage structure analysis is carried out in combination with the distributed structure recombination method of the cloud environment virtual resource configuration data storage node, and a cloud environment virtual resource configuration data architecture model is established. The distributed mobile social network (DSNs) is used to design the cloud environment virtual resource allocation structure model, and the optimal distribution model of the cloud environment virtual resource configuration data storage node is constructed by combining the vector quantization analysis method [5]. The graph model structure of the cloud environment virtual resource configuration data is represented by a binary digraph \( G = (V, E) \), wherein, \( V \) is a vertex set deployed in the cloud environment virtual resource configuration map model distribution node, and the \( M_1, M_2 \cdots M_N \) is a set of all edges of the cloud environment virtual resource configuration data in the finite field distribution area \( G \). assuming that the \( W = \{u, w_1, w_2, \cdots, w_k\} \) is the Sink node of the cloud environment virtual resource configuration data, the Euclidean distance is used to represent the phase track pitch of the data transmission node of the virtual resource allocation data transmission node of the cloud environment, and the information coverage area of the cloud environment virtual resource allocation data, It is assumed that the transmission link layer data of the virtual resource allocation of the M cloud environments is \( x_k = [x(\eta_1), x(\eta_2), \cdots, x(\eta_k)]^T \), and the characteristic value of the initial position \( x(k-1), \cdots, x(k-M) \) of the resource allocation data is as follows:

\[
\hat{x}_s = W_s^T y
\]

Mining association rule attribute set for virtual resource allocation data for cloud environment:

\[
\max_{\hat{x}_s, s, \eta} \sum_{a, b, d, p} \sum_{x, y, \eta} \sum_{x, y, \eta, \eta} \sum_{x, y, \eta} V_p
\]
s.t. \( \sum_{x \in A} \sum_{d \in D} \sum_{p \in P} x_{x,d,p} R_{d}^{bw} \leq K_{p}^{bw} (S), b \in B \) \hspace{1cm} (3)

According to the priority attribute scheduling method, the load model for constructing the cloud environment virtual resource allocation is as follows:

\[
    r(t) = \sum_{j} \sum_{i} b_{j} \alpha_{i} p(t - iT_{j} - f_{j}T_{f} - c_{j}T_{c} - \tau_{j}) + \omega(t) = \sum_{j} \sum_{i} b_{j} \alpha_{i} p(t - iT_{j} - f_{j}T_{f} - c_{j}T_{c} - \tau_{j}) + \omega(t)
\] \hspace{1cm} (4)

Wherein

\[
p_{j}(t) = \sum_{\tau=0}^{t-1} \omega(t)
\] \hspace{1cm} (5)

In addition, the \( \omega(t) \) is the data dimension of the virtual node, and \( p_{j}(t) \) is the load between the cloud environment virtual resource configuration data Source and the Sink node. The virtual resource distribution structure model of the cloud environment is as follows:

\[
x = [s_{1}, s_{2}, \ldots, s_{K}]
\] \hspace{1cm} (6)

Where \( K = N - (m - 1) \tau \), representing the embedded dimension of the cloud environment virtual resource allocation data search feature space, \( sf \) is the delay. According to the above analysis, a big data configuration structure model of cloud environment virtual resources is constructed. Online dictionary learning is used for data compression and feature reconstruction [6].

2.2. Data feature analysis

Assuming that the statistical distribution sequence \( \{x_{1}, x_{2}, \ldots, x_{N}\} \) of virtual resource allocation data flow in cloud environment to be parallel clustering makes \( x(n) \) a set of regression analysis features, the sparse evacuation point cloud mapping of cloud environment virtual resource allocation data is carried out in \( m \) dimension reconstruction space, and the distributed reorganization structure of cloud environment virtual resource allocation data is obtained as follows:

\[
X(n) = \{x(n), x(n+\tau), \ldots, x(n+(m-1)\tau)\} \quad n = 1, 2, \ldots, N
\] \hspace{1cm} (7)

Among them, \( \tau \) represents the embedding delay of cloud environment virtual resource allocation data in high-dimensional phase space, and establishes the state transfer model [7]. The characteristic evaluation concept expression of cloud environment virtual resource allocation data is as follows:

\[
p(y | x, \theta) = \sum_{k=1}^{K} \alpha_{k} p_{k}(y | x_{k}, \sum_{k})
\] \hspace{1cm} (8)

Using the method of cloud sparse evacuation random point structure reorganization, the scattered point set of the first cloud environment virtual resource allocation data is \( P_{j} = \{p_{1j}, p_{2j}, \ldots, p_{mj}\} \), in which:

\[
f \in N_{j}(k), N_{j}(k) = \{x_{j}(k) - x(k) | \leq r_{j}(k)\}
\] \hspace{1cm} (9)
The regression analysis of the cloud environment virtual resource configuration data and the point cloud structure recombination are carried out in combination with the non-linear statistical sequence analysis method [8], and the trust relation of the cloud environment virtual resource configuration data is expressed as \( A \rightarrow B \), \( B \rightarrow C \), and the regression analysis model is expressed as follows:

\[
MSD_{ab}=1-\frac{\sum_{i=1}^{k}\left[(d_{ai,j}-\overline{d}_a)^2+(d_{aj,b}-\overline{d}_b)^2\right]}{\left|L_{ab}\right|}\times\frac{\sum_{i=1}^{k}\left[(d_{ai,j}-\overline{d}_a)^2+(d_{aj,b}-\overline{d}_b)^2\right]}{2}\]

(10)

According to the attribute set of virtual resource allocation in cloud environment, the vector quantitative decomposition is carried out, and the fuzzy domain of distributed big data retrieval of virtual resource allocation in cloud environment is recorded as \( \theta_T \), that is, the fuzzy domain of distributed retrieval of virtual resource allocation in cloud environment is as follows:

\[
a_i(\theta)=T^i_a(\theta)
\]

(11)

The frequency domain and time domain feature analysis of data in fuzzy domain is carried out, and the time-frequency distribution structure model of feature distribution in cloud environment virtual resource allocation is obtained, and the data parallel clustering analysis is carried out under the control of fusion clustering constraints [9].

3. Optimization of parallel clustering algorithm for data

3.1. Joint conditional mutual information mining

Based on the construction of big data configuration structure model of virtual resources in cloud environment, the online dictionary learning method is used for data compression and feature reconstruction and reconstruction [10]. In this paper, a data parallel clustering algorithm based on joint conditional mutual information mining is proposed. In the cloud environment virtual resource allocation big data association scheduling set, the fuzzy scheduling point set satisfies \( U_T = U^{-1} \), and \( V_T = V^{-1} \). The fusion clustering output of parallel allocation of virtual resources in cloud environment is as follows:

\[
x_{a,G} = x_{a,G} + \Delta x_i
\]

(12)

The cloud environment virtual resource allocation query model is constructed, the interference component is \( \rho(e_i | v_j) \), and the cloud environment virtual resource allocation big data optimal clustering characteristic component:

\[
\mu_b = \frac{1}{\sum_{j=1}^{n} \frac{d_{ab,2}}{d_{b,j}}} = \frac{\sum_{i=1}^{m} (\mu_a)^n x_i}{\sum_{i=1}^{m} (\mu_a)^n}
\]

(13)

(14)

The detection statistic characteristic value \( W_{BLCMV} \) of the distributed large data in the cloud environment virtual resource allocation is expressed as follows:
ACMME 2020
IOP Conf. Series: Materials Science and Engineering 914 (2020) 012030 doi:10.1088/1757-899X/914/1/012030

$w_{BCMP} = R_{\hat{\gamma}}^{-1} [a(\theta_0), C] ([a(\theta_0), C]^T R_{\hat{\gamma}}^{-1} a(\theta_0), C)]^{-1}$ \hspace{1cm} (15)

Where, $C = [c_1, c_2, \ldots, c_n]$ represents the cloud environment virtual resource configuration big data's block matching set, the description is shown below:

$U_v = \beta_{ij} \times \log(1 + \gamma_{ij} \times \sum_{j=1}^{K} \sigma_j^2 x_{ij}^T \rho_{ij} \sigma_j), v_j \in v, C_j \in C$ \hspace{1cm} (16)

Among them, $n_{ij} = \sum_{j=1}^{K} \sigma_j^2 x_{ij}^T$, $x_{ij} = (0, ..., 0, 0)$, calculates the fuzzy clustering center of virtual resource allocation in cloud environment, and the mutual information of joint conditions is extracted as follows:

$I_{SM} = A_{SM} H \left\{ \sin \left( \frac{2\pi t}{\rho_{SM}} \right) \right\} \left[ 1 - H \left\{ \sin \left( \frac{2\pi t + D_{SM}}{\rho_{SM}} \right) \right\} \right]$ \hspace{1cm} (17)

Wherein, $A_{SM}$ is the weighted output amplitude of big data parallel programming for virtual resource allocation in cloud environment, $\rho_{SM}$ is the adaptive adjustment parameter of big data parallel programming for virtual resource allocation in cloud environment, and $D_{SM}$ is an inequality constraint. According to the results of feature mining, the optimal design of data parallel clustering is carried out [11].

3.2. Implementation of clustering algorithm

The feature extraction technology is adopted to extract the average mutual information feature quantity of the cloud environment virtual resource configuration data, and the mutual information of the attribute distribution of the output cloud environment virtual resource configuration data is as follows:

$I(Q, S) = H(Q) - H(Q|S)$ \hspace{1cm} (18)

Wherein:

$H(Q|S) = -\sum_{j} \left[ p_{ij}(s_j, q_j) \right] \log_2 \left[ \frac{p_{ij}(s_j, q_j)}{p_{ij}(s_j)} \right]$ \hspace{1cm} (19)

The fuzzy C-means clustering method is adopted to realize the parallel clustering on the extracted feature quantity, and the fuzzy C-mean region distribution set of the large data of the cloud environment virtual resource allocation is obtained as follows:

$\Omega(t) = \frac{y_{ij}^2 \sigma_{ij}^2}{h(G - (N(t) - 1)\gamma_{ij})}$ \hspace{1cm} (20)

The self-adaptive information fusion method is adopted to carry out the information fusion of the data output characteristics of the cloud environment virtual resource configuration data, and the judgment criteria of the data parallel clustering are satisfied:

Norm (1):

$\sqrt{\frac{R_{(m+1)w}^2 - R_{(m)w}^2}{R_{(m)w}^2}} = \frac{|x_{y(x+y)} - x_{y+cr}|}{R_{(m+1)w}} \geq R_{ad}$ \hspace{1cm} (21)

Norm (2):
\[
R_{\ell_{n+1}} > A_{\ell_{n}}
\]

(22)

According to the fuzzy C-means judgment criterion of the cloud environment virtual resource configuration data, the main component analysis of the cloud environment virtual resource configuration data is carried out, the \( \{u_{1}, \ldots, u_{n}\} \) represents a non-trusted node set, \( \{v_{1}, \ldots, v_{m}\} \) represents a user behavior set of the cloud environment virtual resource configuration data, the self-adaptive optimization of the data clustering center is realized, and the iterative formula of the optimization of the data clustering center is as follows:

\[
x_{i}(k+1) = x_{i}(k) + \rho \frac{x_{i}(k) - x_{i}(k)}{\|x_{i}(k) - x_{i}(k)\|}
\]

(23)

Wherein, \( \|\bar{x}\| \) represents the norm of \( \bar{x} \). According to the above algorithm design, the parallel clustering processing of big data for virtual resource allocation in cloud environment is realized [12].

4. Simulation experiment and result analysis

In order to verify the application performance of this method in parallel clustering of virtual resource allocation data in cloud environment, the simulation results are carried out with Matlab and C programming software. The sampling sample database of virtual resource allocation data in cloud environment comes from cloud combination database Pearson Database, data set size 1200, training data set size 120, and the time length of collecting cloud environment virtual resource big data is 400. The dimension of feature reconstruction is 4, the sampling time delay is 0.25, and the convergence judgment threshold is 0.28. According to the above simulation environment and parameter setting, the virtual resource allocation data in cloud environment is clustering in parallel, and the time domain waveform of data distribution is shown in Fig. 1.

![Figure 1. Data distribution time-domain waveform](image1)

![Figure 2. Data feature reconstruction](image2)

Taking the data of Fig.1 as input, digging the mutual information feature quantity of the joint condition, adopting a fuzzy C-means clustering algorithm to carry out characteristic filtering and attribute set merging processing on the extracted feature quantity, and obtaining the characteristic reconstruction output of the data as shown in Fig.2.

The analysis of Fig.2 shows that the clustering processing of the cloud environment virtual resource configuration data is carried out by the method, and the feature reconstruction capability is
good and the discrete distribution is strong, so that the data optimization clustering is realized, and the clustering output is obtained as shown in Fig.3.

![Figure 3. Data cluster output](image)

Figure 3 shows that the classification performance of cloud environment virtual resource allocation data clustering is better by using this method. Test error score rate, get the contrast result as shown in Table 1.

| Iterations | Proposed method | K-means clustering | Particle cluster |
|------------|-----------------|--------------------|-----------------|
| 100        | 0.105           | 0.565              | 0.335           |
| 300        | 0.046           | 0.356              | 0.275           |
| 500        | 0.008           | 0.264              | 0.211           |
| 700        | 0               | 0.165              | 0.114           |

5. Conclusion
In the cloud storage system, the data of the cloud storage virtual resource needs to be optimized cluster analysis, and the data parallel clustering algorithm based on the mutual information mining of the joint condition is proposed in this paper. The optimal distribution model of the cloud environment virtual resource allocation data storage node is constructed, the sparse scattered point cloud mapping of the cloud environment virtual resource configuration data is carried out, and the regression analysis of the data and the reconstruction of the point cloud structure are realized. The feature extraction technology is adopted to extract the average mutual information feature quantity of the cloud environment virtual resource configuration data, and the fuzzy C-mean region distribution set of the large data of the cloud environment virtual resource allocation is obtained, so that the data optimization clustering is realized. The results show that the error rate of the parallel clustering of the data is small and the feature extraction capability is good.

References
[1] DING J, LI A X, HU Z Q, et al. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks [C] //Proceedings of the 2017 International Conference on Medical Computing and Computer-Assisted Intervention. Berlin: Springer, 2017: 559-567.

[2] KAMNITSAS, K, LEDIG C, NEWCOMBE V F, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation [J]. Medical Image Analysis, 2016, 36: 61-78.
[3] DOU Q, CHEN H, YU L Q, et al. Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks [J]. IEEE Transactions on Medical Imaging, 2016, 35 (5): 1182-1195.

[4] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation [C] //Proceedings of the 2015 International Conference on Medical Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

[5] SU Kuanhao, HU Lingzhi, STEHNING C, et al. Generation of brain pseudo-CTs using an undersampled, single-acquisition UTE-mDixon pulse sequence and unsupervised clustering [J]. Medical Physics, 2015, 42 (8): 4974-4986.

[6] JOHANSSON A, KARLSSON M, NYHOLM T. CT substitute derived from MRI sequences with ultrashort echo time [J]. Medical Physics, 2011, 38 (5): 2708-2714.

[7] JOHANSSON A, KARLSSON M, YU Jun, et al. Voxel-wise uncertainty in CT substitute derived from MRI [J]. Medical Physics, 2012, 39 (6): 3283-3290.

[8] ANDREASEN D, VAN LEEMPUT K, EDMUNDJ M. A patch-based pseudo-CT approach for MRI-only radiotherapy in the pelvis [J]. Medical Physics, 2016, 43 (8): 4742-4752.

[9] ANDREASEN D, VAN LEEMPUT K, HANSEN R H, et al. Patch-based generation of a pseudo CT from conventional MRI sequences for MRI-only radiotherapy of the brain [J]. Medical Physics, 2015, 42 (4): 1596-1605.

[10] HAN Xiao. MR-based synthetic CT generation using a deep convolutional neural network method [J]. Medical Physics, 2017, 44 (4): 1408-1419.

[11] QIAN Pengjiang, SUN Shouwei, JIANG Yizhang, et al. Cross-domain, soft-partition clustering with diversity measure and knowledge reference [J]. Pattern Recognition, 2016, 50: 155-177.

[12] TU Guangsheng, YANG Xiaoyuan, ZHOU Tanping. Efficient identity-based multi-identity fully homomorphic encryption scheme [J]. Journal of Computer Applications, 2019, 39 (3): 750-755.