Hierarchical mining algorithm for high dimensional spatiotemporal big data based on association rules

Chunlei Zhou[1], Xinwei Dong[2*], Liang Ji[1], Bijun Zhang[1], Zhongping Xu[3] and Chengping Zhang[3]

[1] Big Data Center of State Grid Corporation of China, Beijing, 100052, China
[2]AnHui Jiyuan Software Co., Ltd, Hefei, Anhui, 230088, China
[3]Beijing Sgitg Accenture Information Technology Center Co., Ltd, Beijing, 100052 China

Abstract. The traditional data mining algorithm focuses too much on a single dimension of data time or space, ignoring the association between time and space, which leads to a large amount of computation and low processing efficiency of the mining algorithm and makes it difficult to guarantee the final data mining effect. In response to the above problems, a hierarchical mining algorithm based on association rules for high-dimensional spatio-temporal big data is proposed. Based on the traditional association rules, after establishing the association rules of spatio-temporal data, the data to be mined are cleaned for redundancy. After selecting the local linear embedding algorithm to reduce the dimensionality of the data, a hierarchical mining strategy is developed to realize high-dimensional spatio-temporal big data mining by searching frequent predicates to form a spatio-temporal transaction database. The simulation experiment results verify that the algorithm has high complexity and can effectively reduce the processing volume, which can improve the processing efficiency by at least 56.26% compared with other algorithms.

1Introduction

The essential function of spatio-temporal data is to reflect the quantitative and qualitative characteristics, spatial structure and spatial relationships of the elements or phenomena in space-time and their changes over time, which is the basis for human cognition of the geographic world[1-2]. Spatio-temporal data reflects the spatio-temporal laws of human activities, and is the basic spatio-temporal framework for all big data collection and aggregation. Spatio-temporal big data is the fusion of big data and spatio-temporal data, i.e., big data with the Earth as the object, based on a unified spatio-temporal datum, and activities directly or indirectly associated with location in space-time [3]. High-dimensional spatio-temporal big data is the data with higher dimensionality and complexity based on spatio-temporal big data with the continuous development of spatio-temporal detection technology. The prevalence of the existence of high-dimensional spatio-temporal big data makes the study of high-dimensional data mining of great importance. The traditional spatio-temporal big data mining algorithms often focus on temporal granularity or spatial density, and when applied to high-dimensional spatio-temporal data mining will result in low data mining efficiency due to the high data dimensionality and large computational load of the algorithm[4-5]. With regard to the above analysis, in order to improve the efficiency of high-dimensional spatio-temporal big data mining and reduce the data mining process, this paper will study the association rule-based hierarchical mining algorithm for high-dimensional spatio-temporal big data.

2Association rule-based hierarchical mining algorithm for high-dimensional spatio-temporal big data

2.1Establishing spatio-temporal data association rules

To measure the reliability of association rules, two parameters, support and confidence, are introduced. Support refers to the percentage of data transactions that contain A and also contain B. Support indicates the prevalence of $A \Rightarrow B$ in D. Confidence refers to the percentage of the number of transactions containing A and B compared to the number of A transactions, representing the degree of certainty that $A \Rightarrow B$ holds. In data mining using association rules, it is necessary to set the minimum support and minimum confidence according to the size of the mining object and the needs of the mining purpose[6]. When the rule is greater than both the minimum support and the minimum confidence, the rule is called a strong association rule. To facilitate the calculation, a value between %100 and 0, instead of a value between 0.1 and 0, is used to represent the support and confidence. From equation (1), there is [7].

*E-mail: dongxinwei@sgitg.sgcc.com.cn

$$codfidence\left(A \Rightarrow B\right) = \frac{support\left(A \cup B\right)}{support\left(A\right)} \quad (1)$$

In equation (1), $codfidence$ is the confidence of the association rule; $support$ is the support of the association rule.

For spatio-temporal big data, the association rules between the data, i.e., spatio-temporal association rules, are established based on the traditional association rules with the introduction of spatial and temporal constraints.

The core of the spatio-temporal association rule lies in spatial association analysis and time segmentation, which is expressed as $P_1 \wedge P_2 \cdots \wedge P_m \Rightarrow Q_1 \wedge Q_2 \cdots \wedge Q_n \left[valid-time\right](s\%, c\%)$, where $P_1 \wedge P_2 \cdots \wedge P_m \Rightarrow Q_1 \wedge Q_2 \cdots \wedge Q_n$ contains at least one spatial predicate, valid-time is the effective time, $s\%$ is the support of the spatio-temporal association rule, and $c\%$ is the confidence of the spatio-temporal association rule. Spatio-temporal association rules consider the constraints of time and space, i.e., the knowledge discovered in the dataset by spatio-temporal association rules is valuable only at a certain valid time and in a certain space. After determining the above spatio-temporal big data mining association rules, data pre-processing is performed on the high-dimensional spatio-temporal big data to be mined in order to reduce the complexity of dimensionality reduction processing of the data at a later stage.

2.2High-dimensional spatio-temporal big data pre-processing

There may be duplicate data in the high-dimensional spatio-temporal big data to be processed, and in order to avoid redundant big data affecting data mining efficiency, the high-dimensional spatio-temporal big data is cleaned and processed based on the clustering principle. The detection of duplicate big data is divided into field detection and record detection, and the field detection problem is the core. If the attribute values of attributes A_i of two sets of spatio-temporal big data are x and y, respectively, their attribute similarity is calculated by the following formula [7-9].

$$S_{Ai}\left(x, y\right) = 1 - \frac{d - \min d}{\max d - \min d} \quad (2)$$

In the above equation, d is the set data attribute similarity threshold. $\min d$ is the minimum distance between the spatio-temporal big data attribute values x and y. $\max d$ is the maximum distance between data attribute values x and y. The data attribute similarity is used to calculate the matching degree of the records. That is, the weight of each attribute in the record and the calculation of attribute similarity are superimposed, and the record similarity is defined as follows.

$$Sim\left(x, y\right) = \sum_{i=1}^{n} S_{Ai}\left(x, y\right) \quad (3)$$

In equation (3), $Sim\left(x, y\right)$ is the record similarity of spatio-temporal big data. n is the total number of data attributes such as temporal and spatial dimensions of spatio-temporal big data. This formula also needs to be used in combination with the following formula for calculating attribute weights.

$$Sim2\left(x, y\right) = \sum_{i=1}^{n} w_i S_{Ai(x,y)}, \sum_{i=1}^{n} w_i = 1 \quad (4)$$

So the record similarity is calculated by first calculating the attribute similarity for attribute i. Then formula (4) is applied to calculate the similarity between records. After determining the similarity between data, different types of data attributes are matched to achieve data cleaning. The principle of data field matching is that a string is viewed as an arrangement of several different sequences of letters combined. A simple definition of this aspect of similarity can be given: the value of identical characters accounting for half of the sum of each field is called similarity. If the strings are identical, they must match, and the specific steps of field matching are as follows.

In the first step, the starting strings are sorted in the original segment. In the second step, the strings are compared against the other recorded strings in order. The number of matches is recorded. In the third step, the matching degree is calculated by the formula. For the high-dimensional spatio-temporal big data to be processed, the data with matching fields are redundant data, and the data cleaning process can be completed by deleting this part of data. After the above pre-processing steps, the high-dimensional spatio-temporal big data are dimensionally reduced to improve the efficiency of subsequent data mining.

2.3Dimensionality reduction of high-dimensional spatio-temporal big data

High-dimensional spatio-temporal big data contains a large amount of effective data information, but at the same time, the impact of dimensional disaster caused by the high dimensionality of the data also makes the mining of high-dimensional spatio-temporal big data exceptionally difficult, and it is necessary to reduce the dimensionality of the data while retaining the effective information of high-dimensional spatio-temporal big data.

In this paper, we choose local linear embedding algorithm to reduce the dimensionality of high-dimensional spatio-temporal big data. The local linear embedding algorithm converts the global nonlinear problem into a local linear problem by expressing the overall global structure information through the overlapping local neighborhoods. The global low-dimensional representation of the data points is obtained by regrouping each sample data point and its neighborhood after dimensionality reduction by some rules[10]. If for a given high-dimensional spatio-temporal large data set $X = \left\{x_1, x_2, \cdots, x_n\right\}$, the nearest neighbor points of each sample are determined by the k-

order nearest neighbor method, the Euclidean distance is calculated as follows.

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left| x_{ik} - x_{jk} \right|^2} \quad (5)$$

A local reconstruction weight matrix is calculated from the nearest neighbors of the sample points, and the following error function is defined.

$$\min \varepsilon(W) = \sum_{k=1}^{n} \left| x_i - \sum_{k=1}^{n} w_{ij} x_{ij} \right|^2 \quad (6)$$

In equation (6), x_{ij} is the k th nearest neighbor of the sample point x_i . w_{ij} is the weight between x_i and x_{ij} and the sum of all values is 1. A d-dimensional mapping is constructed from the locally reconstructed weight matrix of the sample points and their nearest neighbors, and the minimization loss function is calculated according to the following equation.

$$\min \varepsilon(Y) = \sum_{I=1}^{N} \left\| YI_i - YW_i^T \right\|^2 = tr\left(YMY^T \right) \quad (7)$$
$$M = (I - W)^T (I - W)$$

In equation (7), Y is the output vector matrix after dimensionality reduction of the data. I_i is the unit matrix. W is the weight matrix. To minimize the loss function, Y is taken as the eigenvector corresponding to the smallest d non-zero eigenvalues of M . Usually, the eigenvector corresponding to the eigenvalues between the 2nd and d+1 is taken as the output result. After the dimensionality reduction of high-dimensional spatio-temporal big data, the spatio-temporal big data hierarchical mining strategy is developed to complete the hierarchical mining of high-dimensional spatio-temporal big data.

3Simulation experimental research

Spatio-temporal data contains a large amount of data information, and the performance of data mining algorithms will directly determine the use of effective information in spatio-temporal data. Regarding the limitations of traditional data mining methods in dealing with high-dimensional data, a hierarchical mining algorithm based on association rules for high-dimensional spatio-temporal big data is proposed in the above paper. In this section, the performance of this data mining algorithm will be tested by simulation experiments.

3.1Experiment content

This simulation experiment compares the performance of the data mining algorithm proposed in this paper with the spatio-temporal transaction-based mining algorithm and the clustering mining algorithm for testing. Among them, the data mining algorithm proposed in this paper is used as the experimental group, and the spatio-temporal transaction-based mining algorithm and the clustering mining algorithm are used as comparison group 1 and comparison group 2, respectively. The performance of the data mining algorithm is compared by comparing three indexes: algorithm complexity, hardware load situation, and algorithm processing efficiency when data mining is performed by the data mining algorithm. Comprehensive analysis of experimental data

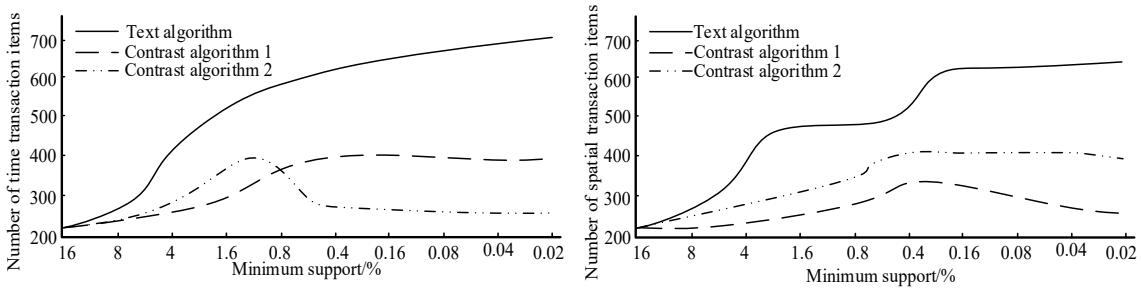
3.2Experiment Preparation

Test data for the experiment: the total number of data is 1292567, which are all high-dimensional spatio-temporal big data collected from various monitoring channels, which do not include spatial transactions that represent only binary items. The collected high-dimensional spatio-temporal big data are composed into a data set, and the total number of all data attributes in this data set is 13.

Experimental environment: Intel(R) Celeron(R) M CPU 964@3.720 GHz, 16G RAM, operating system Windows 8. All data mining algorithms in this simulation experiment are implemented on Visual C# 2010.NET development platform. The experimental data were processed by MATLAB 2012a software. A training set consisting of low-dimensional data was used to test the operation of the three algorithms on the computer simulation platform before the experiments to avoid interference during the experiments.

3.3Experimental Results

The experimental data were data mined using the three data mining algorithms, and the minimum support degree set in the data mining process was changed. The processing complexity of the algorithms is compared by comparing the change in time and space complexity of the data mining algorithms under different minimum support degrees. The algorithm complexity of the three data mining algorithms is shown in figure 2 below, where figure (a) shows the curve of time complexity of the data mining algorithm with the minimum support degree; figure (b) shows the curve of space complexity of the algorithm with the minimum support degree. In the mining process, the more the number of transaction items, the higher the complexity of the corresponding algorithm, and the less the algorithm repeats the processing computation.



(a) Comparison of time complexity (b) Comparison of spatial complexity

Figure 2 Comparison of algorithm complexity

The time complexity of the comparison algorithm increases and stabilizes as the support decreases, while the time complexity of the comparison algorithm 2 increases slowly and then decreases rapidly, indicating that the time complexity of the comparison algorithm 2 is better than that of the comparison algorithm 1 within a certain range, but overall, the time complexity of the comparison algorithm is higher. In terms of space complexity, the space complexity curve of the algorithm in the paper shows two large increases, but the overall trend is to increase as the support decreases. The spatial complexity of the two compared algorithms increases slowly and then decreases. The above analysis shows that the algorithm in the paper has a higher time complexity and space complexity than the comparison algorithm when performing data mining, i.e., the algorithm has a better complexity.

The hardware load and algorithm processing time of the data mining algorithm when mining the experimental data are shown in table 1 below. The data in the experimental dataset are divided into subsets with different data volumes and labeled in ascending order according to the data volumes. The processing efficiency of the algorithms and the load on the hardware when the algorithms run are compared for different data volumes of spatio-temporal data mining.

Table 1 Algorithm processing efficiency and hardware load comparison

Data subset number	Algorithm in the text		Comparison Algorithm 1		Comparison Algorithm 2	
	Hardware memory usage ratio /%	Data mining time /s	Hardware memory usage ratio /%	Data mining time /s	Hardware memory usage ratio /%	Data mining time /s
1	6.3	8.5	9.8	11.4	10.4	19.7
2	6.3	8.7	12.3	16.8	13.6	24.6
3	6.4	8.8	14.5	20.5	17.2	29.8
4	6.5	9.4	18.9	25.1	21.5	34.4
5	6.7	10.0	22.7	30.0	26.1	39.5

The analysis of the data in the above table shows that the processing time of the two comparison algorithms increases as the data volume of the data mining object increases, and the hardware space occupation increases, which also indicates that the load pressure on the hardware caused by the algorithms is also increasing. In comparison, the algorithm in the paper has almost no significant increase in the load pressure on the hardware when mining data of different data volumes, and the data mining time of the algorithm is significantly shorter than that of the two comparison algorithms. Further processing of the data in the above table shows that the average processing time of the algorithm in the paper is 9.08 s, compared with 20.76 s for the comparison algorithm 1 and 29.6 s for the comparison algorithm 2. From the data perspective, the processing efficiency of the algorithm in the paper is improved by at least about 56.26%. In summary, the association rule-based hierarchical mining algorithm for high-dimensional spatio-temporal big data studied in this paper has a high complexity and significantly improved processing efficiency.

4Conclusion

With the development of modern technology, people gradually realize the importance of spatio-temporal big data applications. Regarding the problems of traditional data mining algorithms, this paper proposes a hierarchical mining algorithm for high-dimensional spatio-temporal big data based on association rules, and through comparative simulation experiments, it is proved that the algorithm has a high spatio-temporal complexity and can effectively reduce the computation in the data mining process. In order to improve the complexity of the algorithm, this paper adopts the strategy of dimensionality reduction followed by hierarchical mining, while to cater for the development of spatio-temporal data, the direct processing of data should be attempted to improve the algorithm processing efficiency.

References

1. Maria Dagaeva, Alina Garaeva, Igor Anikin, et al. (2019) Big spatio-temporal data mining for

emergency management information systems[J]. IET Intelligent Transport Systems, 13(11) :1649-1657.

2. Ben Y, Ma J, Wang H, et al. (2019) A spatio-temporally weighted hybrid model to improve estimates of personal PM 2.5 exposure: Incorporating big data from multiple data sources[J]. Environmental Pollution, 253:403-411.
3. Wang X, Qin D, Zhang D, et al. (2019) Evolution Characteristics of Overburden Strata Structure for Ultra-Thick Coal Seam Multi-Layer Mining in Xinjiang East Junggar Basin[J]. Energies, 12(2):332.
4. Durga Toshniwal, Narayan Chaturvedi, Manoranjan Parida, et al. (2020) Application of clustering algorithms for spatio-temporal analysis of urban traffic data[J]. Transportation Research Procedia, 48:1046-1059.
5. Liu LY, Zhan HX, Liu JX, et al. (2019) Visual analysis of traffic data via spatio-temporal graphs and interactive topic modeling[J]. Journal of Visualization, 22(1) :141-160.
6. Jonathan R. Bradley, Scott H. Holan, Christopher K. Wikle. (2018) Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)[J]. Bayesian Analysis, 13(1) :253-310.
7. Mohammad Jafarzadegan, Faramarz Safi-Esfahani, Zahra Beheshti. (2019) Combining hierarchical clustering approaches using the PCA method[J]. Expert Systems With Applications, 137:1-10.
8. Maria Th. Kotouza, Fotis E. Psomopoulos, Pericles A. Mitkas. (2020) A dockerized framework for hierarchical frequency-based document clustering on cloud computing infrastructures[J]. Journal of Cloud Computing: Advances, Systems and Applications, 9(6) :30-38.
9. Saptarshi Das, Shamik Sural, Jaideep Vaidya, et al. (2019) Policy Adaptation in Hierarchical Attribute-based Access Control Systems[J]. ACM Transactions on Internet Technology (TOIT), 19(3) :1-24.
10. Prasad S. Nishtala, Te-yuan Chyou. (2020) Identifying drug combinations associated with acute kidney injury using association rules method[J]. Pharmacoepidemiology and Drug Safety, 29(4) :467-473.