Research on Association Mining Method of Frequent Itemsets in High-dimensional Multi-source Big Data

Yingshan Li

{40478507@qq.com}

Middle South Regional Air Traffic Management Bureau of CAAC , Guangzhou, Guangdong, 510422, China

Abstract. The conventional association mining method of frequent itemsets in high-dimensional multi-source big data mainly uses the framework of Parameter Server to solve the problem, which is easily influenced by the change of data attribute relationship, resulting in low accuracy of data association mining. Therefore, it is necessary to design a new association mining method of frequent itemsets in high-dimensional multi-source big data. That is, the association mining strategy of high-dimensional big data frequent itemsets is generated, and the association mining algorithm of high-dimensional multi-source big data frequent itemsets is designed, thus realizing the association mining of high-dimensional multi-source big data frequent itemsets. The experimental results show that the designed association mining method for frequent itemsets of high-dimensional multi-source big data has high accuracy, which proves that the designed association mining method has good mining effect, reliability and certain application value, and has made certain contributions to improving the processing efficiency of high-dimensional multi-source big data.

Keywords: High dimension; Multi-source; Big data; Frequent itemsets; Association; Digging; Way.

1 Introduction

With the development of information technology, the number of large databases in China is increasing day by day, and the data in large databases have certain privacy, from which we can obtain hidden knowledge and improve the effect of data decision-making. Based on this, data mining technology has emerged. Data mining, also known as data knowledge acquisition and discovery, can find a large number of hidden data knowledge information [1-3] , analyze the relevance between these information and users, and thus build a feasible scheme. Therefore, data mining is an inevitable choice to enhance the competitiveness of data decision-making.

Data analysis technology is of great significance to the development of our country. The application of data analysis technology can effectively process massive data sets and efficiently calculate data [4-7] , but it is difficult for data analysis technology to mine the correlation between data and make reference for upper-level decision-making. In fact, the mined data abandons the fuzziness of the original data and increases the pertinence and decision-making reference of the data. At present, the common data association mining technologies include data clustering, data prediction, etc., and it is necessary to select targeted methods to complete the decision according to the actual requirements of data mining. In the

PMBDA 2023, December 15-17, Nanjing, People's Republic of China

DOI 10.4108/eai.15-12-2023.2345306

process of association mining of frequent itemsets [8-10] , it is necessary to effectively process metadata, determine the relationship between data items, calculate the support of mining, and output the association rules between data items. At present, the data in most databases belong to high-dimensional and multi-source data, and its composition is relatively complex, which contains multiple frequent itemsets, so it is difficult to carry out association mining of data frequent itemsets [11-13] . If the support of association mining is not set in advance, it is easy to cause serious association mining problems if the mining threshold is determined. According to the characteristics and existing problems of frequent itemsets mining of high-dimensional multi-source big data, relevant researchers have designed several conventional association mining methods of high-dimensional multi-source big data frequent itemsets. The first is the association mining method of high-dimensional multi-source big data frequent itemsets based on Apache Spark distributed computing framework, and the second is the association mining method of high-dimensional multi-source big data frequent itemsets based on personalized tag optimization. Most association mining methods of frequent itemsets in high-dimensional multi-source big data mainly use the framework of Parameter Server [14-15] , which is easily affected by the change of data attribute relations, resulting in low accuracy of data association mining, which does not meet the current requirements of data association mining. Therefore, this paper designs a new association mining method of frequent itemsets in high-dimensional multi-source big data.

2 Design of association mining method for frequent itemsets of high-dimensional multi-source big data

2.1 Association mining strategy for generating frequent itemsets of high-dimensional big data

In the process of association mining of frequent itemsets in high-dimensional big data, it is necessary to pre-process the data to be mined, including data screening, transformation and other steps to verify whether the information has mining reference degree. Association rule mining has several basic attributes. First, there is an iterative relationship between its items and data items, second, its support represents the specified data items, and finally, frequent itemsets represent its association mining standards. Therefore, in order to improve the association mining effect of frequent itemsets, it is necessary to generate effective association mining strategies for high-dimensional large data frequent itemsets.

When a given transaction database is in the state of minimum support, its different mining paths can represent the mining requirements of frequent itemsets. At this time, the traversal operation can be completed directly without scanning the original data set. However, when faced with complex high-dimensional and multi-source frequent itemsets of big data, it is often difficult to read the ranking state of data, and it is necessary to construct an effective processing path for frequent itemsets mining. Therefore, the association mining strategy of big data frequent itemsets generated in this paper is based on the original mining database, and the association mining strategy is generated by using the greatest common divisor mode.

In the actual mining process of large data frequent itemsets, the data volume and data complexity of frequent itemsets are mainly considered, and the layout is carried out according

to the vertical layout method, and then the coding conversion is unified, thus improving the efficiency of data mining. At this time, the data conversion expression TV as shown in the following (1).

$$TV = \prod P_R \tag{1}$$

In the formula (1), P_R representing a given set of things, an effective association mining PNFP-Tree can be constructed based on the above data conversion expressions. At this time, the construction rules of PNFP-Tree are as follows:

First, if there is a node insertion relationship among nodes that does not belong to the association mining tree structure, it needs to be inserted as the only node, and its local count attribute should be adjusted to 1; Secondly, if the TVs between nodes are equal, it is necessary to adjust the local counting attribute and accumulate; Thirdly, if the global count attribute value in the node is equal to the dividend in the data set, the parent node needs to be reinserted; Fourthly, if the newly inserted node can be divisible by the original node, it is necessary to rearrange the nodes. According to the above construction rules, the number of support counts can be calculated $NS_{\sup}$, as shown in the following (2).

$$NS_{\sup} = \sum_{i=1}^{n} N_i * support \tag{2}$$

In the formula (2), N_i represents a mining node, $support$ represents the attribute value of each node. At this time, it can be assumed that different mining nodes have the same mining relationship, then mining prediction can be made, the weight balance value can be determined, and attribute inference can be completed. After repeated iterations, the resources of the candidate set can be compressed to realize efficient mining. At this time, the generated association mining strategy of frequent itemsets of high-dimensional multi-source big data is shown in Figure 1 below.

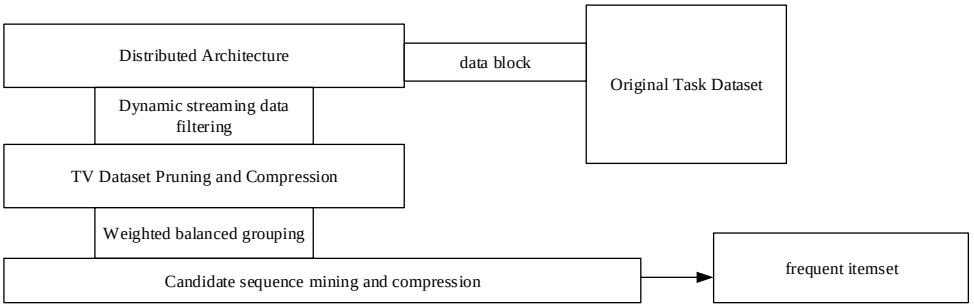


Fig. 1. Association mining strategy for frequent itemsets of high-dimensional multi-source big data.

As can be seen from Figure 1, using the above frequent itemset association mining strategy can improve the independent relationship of each node mining to the greatest extent and ensure the reliability of the final mining.

2.2 Design an association mining algorithm for frequent itemsets of high-dimensional multi-source big data

In the process of multi-source big data mining, we can reduce data based on association rules, and then update frequent itemsets according to mining requirements. Based on this, this paper designs an effective association mining algorithm for high-dimensional multi-source big data frequent itemsets. Based on the standard data, it is necessary to obtain the hidden relationship between the data and make the initial database hypothesis, which is the relationship between the set and the database D_i^* at this time is shown in the following (3).

$$D_i^* = D_i - D_i^- \tag{3}$$

In the formula (3), D_i represents the updated database, D_i^- represents deleting the database set. According to the above database relation, we can process the database support set, judge the relationship between the support numbers, and filter the global candidate item set. At this time, the principle of analysis and deduction can be used to generate a complete set of candidate items, which reduces the difficulty of association mining operation.

Under the constraint of database foundation, the specifications of frequent itemsets will change to some extent. In order to improve the constraint of data association rules, it is necessary to update the global frequent itemsets. The updated itemsets have corresponding central nodes, which meet the global support requirements. The candidate itemsets are updated iteratively, and the association mining algorithm $y(x_n)$ is obtained at this time is shown in the following (4).

$$y(x_n) = \varpi + \sum_{i=1}^{l} w(x_n, t_i) \tag{4}$$

In the formula (4), ϖ represents the training sample output values, $w(x_n, t_i)$ represents the data vector basis function, according to the above iterative processing of itemsets, the reduction calculation can be further carried out, the complexity between itemsets can be determined, and the itemsets that meet the constraints can be selected for further mining. After the data sets that meet the mining requirements are obtained, the data mining quantity specifications can be adjusted, and the real-time association mining of frequent itemsets in big data can be realized by combining rough set theory.

3 Experiment

In order to verify the mining effect of the designed association mining method of high-dimensional multi-source big data frequent itemsets, this paper sets up an effective experimental platform, and compares it with the conventional association mining method of high-dimensional multi-source big data frequent itemsets using Apache Spark distributed computing framework and the association mining method of high-dimensional multi-source big data frequent itemsets based on personalized tag optimization. The experiments are as follows.

3.1 Experimental preparation

According to the experimental requirements of frequent itemsets association mining of high-dimensional multi-source big data, this paper assumes a multi-source complex decision-making problem and generates an experimental attribute set. In this paper, sqoop platform is selected as the experimental platform, which can effectively record HDFS association mining data. The architecture of this experimental platform is shown in Figure 2 below.

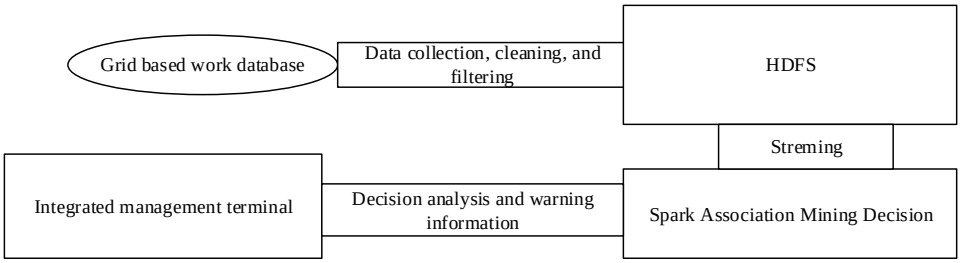


Fig. 2. Experimental Platform Architecture.

As can be seen from Figure 2, the above experimental platform can effectively make decision analysis of early warning information and report grid events. In order to improve the reliability of the experiment, this paper sets up a data cluster with the same scale and plans the Master main driver node. At this time, the CPU is Intel E5 3.10GHz/4 core, the memory is 16GB, the operating system is 64-bit CentOS 6.5, and the experimental platform is supported by Hadoop 2.7.0 and Spark 2.2.0. The experimental flow at this time is shown in the following Table 1.

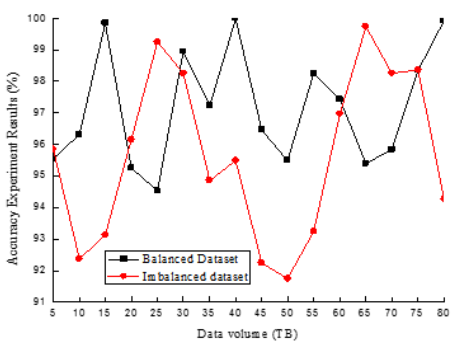
Table 1. Experimental Process.

Step	Experimental process
Step 1	Using the Sqoop tool to collect frequent itemset association data from multiple sources of big data
Step 2	Generate experimental file blocks and store them in different experimental nodes
Step 3	According to experimental requirements, record the grid time and process irrelevant storage data
Step 4	Conduct preliminary cleaning of experimental data and intercept irrelevant experimental data for filtering
Step 5	Processing raw datasets through HDFS multiple file blocks and reading dynamic data streams
Step 6	Perform RDD core calculations and use PNFPM algorithm to obtain support for association mining
Step 7	Output experimental results and conduct in-depth analysis of the experimental results

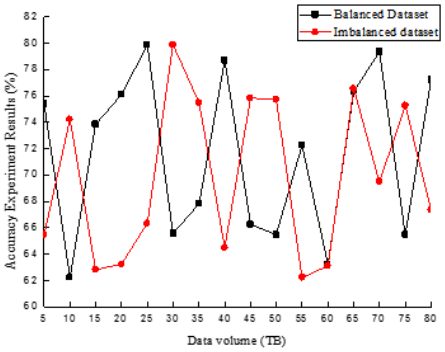
It can be seen from Table 1 that the frequent itemsets of different high-dimensional multi-source big data can be coded by combining the above experimental process, and then the T10I4100 classic dataset is used to expand, and the items that need to be enhanced are introduced into the original experimental dataset, so as to synthesize accurate experimental data.

3.2 Experimental results and discussion

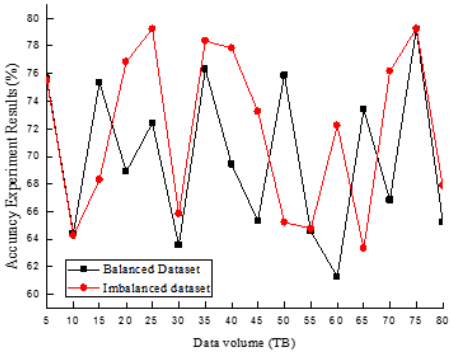
Based on the above experimental preparation, experiments can be conducted on mining frequent itemsets of high-dimensional, multi-source big data. We used the methods presented in this article, Apache Spark distributed computing framework, and personalized label optimization methods for mining, and recorded the data association mining accuracy and mining time of the three methods under different data volumes. The experimental results are shown in Figure 3 and Figure 4.



(a)Proposed method



(b)Apache Spark Distributed Computing Framework Method



(c)Personalized Label Optimization Method

Fig. 3. Accuracy Experiment Results (%).

In the experiment in Figure 3, the performance of the algorithm in handling this real-world situation can be more accurately evaluated by using an imbalanced dataset. By using a balanced dataset, the impact of differences in the number of categories on algorithm performance evaluation can be eliminated. This can ensure that each category receives sufficient attention and opportunities for correct classification, thereby objectively comparing the performance of different algorithms. Therefore, in order to comprehensively evaluate different mining algorithms, accuracy experiments were conducted on both balanced and imbalanced datasets. From Figure 3, it can be seen that the method proposed in this paper has a high accuracy in association mining for different data volumes, and is less affected by the

balance of the dataset. The overall accuracy is higher than 91%. The Apache Spark distributed computing framework method and personalized label optimization method are greatly affected by the balance of the dataset, and there are significant fluctuations in accuracy, with a numerical range of 60% to 80%. In summary, it can be seen that the method in this article uses association rules as the basis for data reduction processing, and then updates the frequent itemsets of data according to mining requirements, optimizing mining accuracy and reducing the impact of data balance.

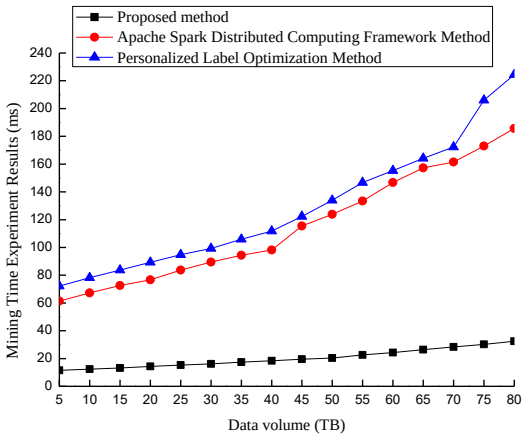


Fig. 4. Mining Time Experiment Results (ms).

From Figure 4, it can be seen that the association mining time consumed by this method is relatively short, with a time consumption of less than 33ms; The Apache Spark distributed computing framework method and personalized label optimization method consume relatively long association mining time, both exceeding 60ms. The above experimental results prove that the method proposed in this paper can mine frequent itemsets of high-dimensional multi-source big data in the shortest possible time, and has a fast mining speed.

4 Conclusion

Under the background of the parallel development of computer technology and information data technology, the number of large-scale databases is increasing, and the scale is gradually increasing. These data are highly random, including transaction data, enterprise management data, user medical data, etc. All kinds of data are complicated and extremely difficult to handle. In addition, the data in most databases are highly integrated, belonging to high-dimensional multi-source data and containing frequent itemsets. In order to solve the random changes in data analysis and make effective data decisions, it is necessary to effectively mine frequent itemsets of high-dimensional multi-source big data. Conventional frequent itemsets association mining methods lack relevant strategy support, and the overall mining performance is poor. In order to solve the above problems, this paper designs a brand-new high-dimensional multi-source big data frequent itemsets association mining method. The experimental results show that the designed frequent itemset association mining method has

good mining effect, accuracy and certain application value, and has made certain contributions to improving the processing effect of high-dimensional multi-source big data.

References

[1] X Gan,X Tang .(2021). A CNN based temporal data association rule mining model.Computer Simulation,,38(3):5.

[2] Zhang, Y., Bai, R., Han, J., Chen, Q., & Gao, X. (2021). Research on TCM Diabetes Assisted Diagnosis and Treatment Plan Integrating Association Mining and Quantitative Calculation. Procedia Computer Science, 188, 52-60.

[3] Wu, Z., & Chen, Y. (2021). Digital art feature association mining based on the machine learning algorithm. Complexity, 2021, 1-11.

[4] Zhao, Z., & Tan, Y. (2021). Research on global tourism information query method based on association mining. International Journal of Information and Communication Technology, 18(3), 288-303.

[5] Xu, R., & Luo, F. (2021). Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest. Safety science, 135, 105125.

[6] Yavari, A., Rajabzadeh, A., & Abdali-Mohammadi, F. (2021). Profile-based assessment of diseases affective factors using fuzzy association rule mining approach: A case study in heart diseases. Journal of Biomedical Informatics, 116, 103695.

[7] Saranyadevi, S., Murugeswari, R., & Bathrinath, S. (2021, February). A context-free grammar based association rule mining technique for network dataset. In Journal of Physics: Conference Series (Vol. 1767, No. 1, p. 012007). IOP Publishing.

[8] Khurana, K., & Sharma, S. (2013). A comparative analysis of association rule mining algorithms. International Journal of Scientific and Research Publications, 3(5), 0.

[9] Tiwari, M., Shanthi, V., & Mishra, A. (2021). Development of association rule mining model for gender classification. In IOP Conference Series: Materials Science and Engineering (Vol. 1022, No. 1, p. 012064). IOP Publishing.

[10] Priyanto, E., Hermawan, A., Rianto, R., & Avianto, D. Efektifitas Penggunaan Association Rules Mining dalam Personalisasi Website. JISKA (Jurnal Informatika Sunan Kalijaga), 6(1), 59-69.

[11] Zhang, X., Tang, Y., Liu, Q., Liu, G., Ning, X., & Chen, J. (2021). A fault analysis method based on association rule mining for distribution terminal unit. Applied Sciences, 11(11), 5221.

[12] Chaisoong, U., Tirakoat, S., & Jareanpon, C. (2021). Tourist information-seeking behaviours using association rule mining. ICIC Express Letters, 15(9), 915-923.

[13] Lima P L D , Fonseca R N D , Gomes R P ,et al. (2021). USE OF ASSOCIATION RULES TO PERFORM THE MINING OF A MARKETING DATABASE COMERCIAL.International Journal for Innovation Education and Research, 9(4):142-152.

[14] Nohuddin P N , Zainol Z , Hijazi M H A . (2021).Study of B40 Schoolchildren Lifestyles and Academic Performance using Association Rule Mining.Annals of Emerging Technologies in Computing, ,5(5):60-68.

[15] Orama, J. A., Borras, J., & Moreno, A. (2021). Combining cluster-based profiling based on social media features and association rule mining for personalised recommendations of touristic activities. Applied Sciences, 11(14), 6512.