ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240129001302
Original scientific paper

# Enhancing Medical Big Data Analytics: A Hadoop and FP-Growth Algorithm Approach for Cloud Computing

Rong HU*, Xueling YANG

**Abstract:** Effective mining of relationships within massive medical datasets can profoundly enhance clinical decision-making and healthcare outcomes. However, traditional data mining techniques falter in extracting actionable associations from large-scale medical data. This research optimizes the Frequent Pattern Growth algorithm and incorporates it into a Hadoop framework for scalable medical data analytics. Empirical evaluations on real-world patient diagnosis records demonstrate the proposed approach's computational and learning efficiency. For instance, with the Break-Cancer database, the optimized algorithm requires just 0.04 seconds at 0.22 minimum support, significantly faster than existing methods. Experiments on diagnostics data generate 267 informative association rules at 0.31 support - markedly higher than 71, 126 and 233 rules produced by other comparative techniques. By enabling rapid discovery of data-driven health insights, the enhanced medical data mining framework provides a valuable decision-support system for better clinical practice. Ongoing explorations focus on further optimizations for automated disease prediction and treatment recommendations to continuously augment data-to-diagnosis applicability.

**Keywords:** cloud computing; frequent pattern growth; Hadoop; MapReduce; medical big data

## 1 INTRODUCTION

The rapid development of information technology and computer technology has created conditions for the progress of various fields such as engineering design, commercial activities, and science and technology. With the accumulation of data, various fields have stored rich historical data. Medical big data is one of them, and various cases, patients, and other data are constantly growing [1]. At present, a variety of complex diseases such as hypertension, coronary heart disease pose a serious threat to human health, causing physical and mental pain to patients and increasing the social burden [2, 3]. Therefore, the prevention and treatment of diseases are of great significance. Mining association rules between data from massive medical databases is beneficial for forming a comprehensive and scientific medical data information management system for disease prevention, drug efficacy evaluation, and disease monitoring. The formation and development of diseases have certain patterns. By analyzing the patient's disease development status, reasonable and effective intervention measures can be formulated. Faced with the growth of massive medical data, how to obtain effective information from it has become an urgent problem to be solved [4, 5]. Therefore, data mining techniques have played an important role in data processing in preventive medicine. However, medical data itself has characteristics such as incompleteness, complexity, and redundancy, making data mining difficult. The existing data mining algorithms are effective for targeting network user data. However, there is unsuitability in feature extraction and classification for specific medical data. The performance of mining association features between data is weak and cannot better extract detailed data features. Therefore, the Frequent Pattern Growth (FP-Growth) is introduced to construct medical data feature mining algorithms. In response to the shortcomings of this algorithm in operation, a pruning strategy is introduced to optimize it. Then, it is combined with Hadoop to construct corresponding medical big data analysis algorithms. It is expected that it can more effectively extract the correlation features between medical data, providing effective support for the treatment and prevention of diseases. The purpose of the research is to leverage the advantages of cloud computing, explore the association rules between medical data, and optimize the feature extraction methods of medical data. Then, cloud computing service methods are integrated with medical big data to optimize the informationization level of medical data systems and achieve more effective disease diagnosis and resource allocation layout. The contributions of the research are as follows. The study first utilized the FP Growth algorithm to construct a feature mining algorithm for medical data to extract detailed features among medical data. Then, based on the FP Growth feature mining algorithm combined with Hadoop, a medical data analysis method was constructed, achieving feature analysis of massive medical data. The study consists of four parts. The first part is the domestic and foreign research results on frequent pattern growth algorithms and medical data. The second part constructs a medical data analysis algorithm based on Hadoop and frequent pattern growth tree. The third part conducts experimental verification on the proposed method in the study. The fourth part summarizes the research results and proposes future research directions.

## 2 RELATED WORKS

Association rules are extensively applied in many fields. With the explosive growth of information data, the efficiency of mining association rules has become a very serious problem. Numerous scholars have conducted in-depth research on it. In business competition, the relationship between organizations and customers is crucial for attracting customer interest, which has a direct impact on improving corporate profits. Therefore, Ugwu N. V. et al. used frequent pattern growth algorithms to identify products of interest to customers. Based on the identification results, the customer's purchasing habits were analyzed. The results showed that this method could accurately analyze customers' interests and habits [6]. Satria C. et al. used FP-Growth and Apriori data mining methods to analyze food stall data. Those best-selling

foods/beverages were recommended. The total sales revenue based on the improved method increased significantly by 2.37 times compared to the previous sales revenue [7]. Ye Z et al. combined particle swarm optimization with FP-Growth to improve the efficiency of association rule mining. A parallel conditional frequent pattern tree (FP-Tree) was proposed to address the memory overflow caused by a large amount of data. By grouping data, the excessive data volume preventing the construction of an FP-Tree is solved. The proposed algorithm generated some data redundancy. However, the efficiency was significantly higher than traditional parallel frequent pattern algorithms [8]. Rachid K. M. et al. proposed an algorithm called TrajGrowth to directly process raw data. This method did not require any preprocessing steps or laborious parameter setting to execute. The proposed method was more accurate. It had better processing time, and avoided redundant patterns compared to discretization method [9]. FP-Growth used FP-Tree to store database information in compressed form. Jamsheela O. et al. improved the data mining process by modifying existing and new data structures. It was implemented using the basic FP-Growth. The experimental results indicated that this method improved mining performance [10]. Transforming growing medical big data into user-friendly medical knowledge is a global issue. Numerous scholars have conducted in-depth research on it from different perspectives. Wu X. et al. conducted research on medical data from data collection, data transmission, and data sharing to address privacy protection issues during the transformation of medical big data. A medical big data privacy protection platform based on the Internet of Things (IoT) was proposed. The experimental results indicated that the proposed platform had advantages and practicality [11]. By analyzing the commonalities between medical contexts, patients could be provided with unified phenotypic characteristics. Ahmad J. H. F. proposed a mechanism for generating a unified Named Entity Recognition labelled medical corpus. The corpus provided a data set of 14407 endocrine patients diagnosed with diabetes and complications in comma separated value format. Different experiments were conducted using common Natural Language Programming techniques and frameworks such as TensorFlow, Keras, Long Short Term Memory (LSTM), and Bi-LSTM. Experiments showed that the maximum accuracy of this method was 0.8846 [12]. Big data analysis can provide personalized drug regimens, clinical risk interventions and predictive analysis, standardize medical terminology, and improve healthcare. Therefore, Gou X. et al. summarized different types of medical big data, including electronic health records, medical image data, medical informatics, remote medical monitoring, biomedical data, and other data sources. In addition, potential challenges and future research directions related to big data healthcare were discussed [13]. Manikandan P. et al. proposed a merged feature selection classification strategy to reduce medical big data. The proposed system was jointly executed by an ant optimizer and an ensemble classifier. Compared with random forest (RF), SVM, and Bayesian classifiers, the proposed method accurately and effectively reduced large medical data [14]. Hurley D. et al. studied the role of medical big data and wearable IoT medical system in remote monitoring and nursing of diagnosed or suspected patients with COVID-19. Artificial intelligence driven biosensors during the COVID-19 pandemic were analyzed and estimated in terms of diagnosis, monitoring and prevention. Deep machine learning and cloud computing were key to healthcare based on the IoT. Medical IoT systems could remotely monitor patients with chronic diseases [15]. In summary, the frequent pattern growth algorithm has achieved significant research results in fields such as data mining and feature extraction. The continuous growth of big data resources in the medical field has provided effective support for the further development and transformation of medicine. However, in existing research, most of the research results have been transformed from the data resources themselves into the required resources. Facing the abundant digital resources, how to quickly and effectively obtain the necessary key information from the massive data resources has become an urgent practical problem to be solved. Therefore, the improved frequent pattern growth algorithm is used for extracting medical big data resources. Then it is combined with Hadoop to construct a medical big data analysis model based on an improved frequent pattern growth algorithm. It is expected that this method can effectively achieve the extraction and analysis of medical resources, enhance the collective experience of medical treatment, and promote the reform and innovation of medical technology.

## 3 CONSTRUCTION OF MEDICAL DATA ANALYSIS MODEL BASED ON HADOOP AND FREQUENT PATTERN GROWTH

Analyzing the association rules between medical data has important reference value for the diagnosis and treatment of various diseases, especially in health examinations, public health defence, and epidemic disease control. Based on the correlation between patient illness, diagnosis, and medication, clinical diagnosis and treatment work can be continuously optimized. This chapter will analyze the frequent pattern growth algorithm. Then it is combined with Hadoop to construct a medical data association rule analysis algorithm.

### 3.1 Medical Data Feature Mining Based on FP-Growth

The association rule mining algorithm can be used for data processing after multiple iterations. Data mining based on association rules is to find associations between data from a database. The connections between the data obtained from this can be represented by association rules or frequent itemsets [16]. $X \rightarrow Y$ is the representation of association rules. There is no mutual relationship between $X$ and $Y$. The size of association rules is represented by support and confidence, as shown in Eq. (1).

$$\begin{cases} S = \dfrac{\sigma(X \cup Y)}{N} \\ C = \dfrac{\sigma(X \cup Y)}{\sigma(X)} \end{cases} \tag{1}$$

In Eq. (1), $\sigma(X)$ represents the support of $X$. Generally speaking, the association rule mining is divided into two steps, as shown in Fig. 1.

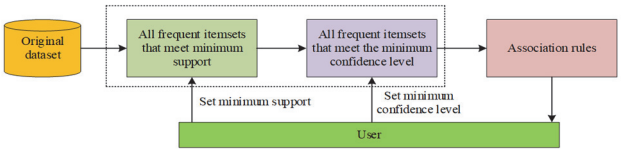



**Figure 1** Mining steps for association rules

In the association rule mining process shown in Fig. 1, all frequent itemsets contained in the transaction database are first identified as candidate rules. By setting the minimum support, the entire transaction database is traversed to find all frequent itemsets that meet the minimum support requirements. Therefore, strong association rules can be determined. In all candidate rules, according to the confidence calculation method shown in Eq. (1), the confidence of each frequent itemset is calculated to obtain association rules that meet the minimum confidence requirement. In association rules, commonly used algorithms include Apriori algorithm and FP-Growth algorithm. In the FP-Growth algorithm, a data structure of Frequent Pattern Tree (FP-tree) is used. FP-tree is a special prefix tree composed of a frequent item header table and an item prefix tree. The FP-Growth is a mining algorithm for association rules based on FP-Tree. In the FP-Growth, the principle is to store the dataset in a specific frequent pattern tree structure to discover frequent itemsets or pairs of items. Conditional patterns are created to mine frequent patterns in FP-Tree to obtain the final association rules [17]. The Apriori algorithm requires frequent scanning of the database for data calculation. On the contrary, the FP-Growth only needs to scan the data twice to obtain frequent patterns, which can effectively improve the algorithm's operational efficiency. The FP-Growth has two stages in the specific operation process. They are building FP-Tree and mining frequent patterns from the constructed frequent pattern trees. In FP-Tree construction, the data needs to be scanned twice. The first scan calculates the frequent itemset to obtain the occurrences of all item items in the dataset. Then, items below the minimum support threshold are removed and the remaining dataset items are arranged in descending order according to the decreasing support, forming a frequent itemlist FList. The second scan uses frequent item lists to construct FP-Tree, which consists of an item header table and a prefix tree. The construction process is shown in Fig. 2.

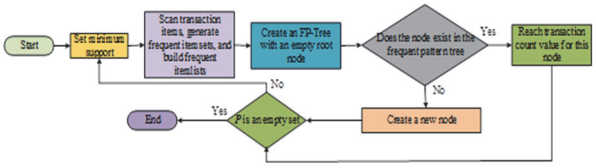



**Figure 2** Construction process of FP-Tree

In Fig. 2, firstly, dataset $D$ and minimum support min-sup are input. The root node of FP-Tree is created. Each piece of data in the dataset is arranged in the order of frequent items in FList. The arranged data is in the form of $[p|P]$. $p$ represents the element in the first position of a piece of data. $P$ represents the complement of $p$. Then, a function is introduced to calculate the data, as shown in Eq. (2).

$$f(x) = insert - Tree([p|P], T) \tag{2}$$

In Eq. (2), $T$ represents the conditional pattern tree. If the inserted data record has the same prefix node as the branch in FP-Tree, then both share these nodes. On the contrary, a new node is createed for $f(x)$ to represent the inserted new data. The FP-Tree function is iterated until the $P$ value is empty. Then the second stage of frequent pattern mining is executed. This process is implemented using the $g(x)$ function, as shown in Eq. (3).

$$g(x) = Growth(T, \beta) \tag{3}$$

Firstly, whether there is a single path in the frequent pattern tree $g(x)$ is determined. The path is represented as $L$. If the path $L$ exists, the pattern $\beta \cup \alpha$ for all subsets $\beta$ in the path is generated. The support of this generation mode is the count value of the smallest node in $\beta$. If the path does not exist in the frequent pattern tree, the pattern $B$ for each term $p_i$ in FP-Tree is generated, as shown in Eq. (4).

$$B = p_i \cup \alpha \tag{4}$$

The count value of this mode is the count value of $p_i$. Then the corresponding conditional pattern base and conditional pattern tree $T_B$ for $B$ are constructed. If $T_B$ is not an empty set, the $Growth(T_B, B)$ function is iterated until $T_B$ is an empty set. Based on the above analysis process of FP-Tree, FP-Growth needs to continuously add relevant medical data when constructing FP-Tree. During this process, some transactions share a prefix transaction, but still generate a new branch. In response to this issue, the research adopts a parallelization approach to solve it. The pruning strategy is introduced to effectively prune FP-Tree to further simplify the structure [18]. In the process of object data mining, some shared or duplicate object items do not need to be searched again. Therefore, the subtrees in this node should not be searched again. For such subtrees, pruning strategy can be used to remove them from the search space. In a frequent pattern tree, the union of the head and tail (HUT) of a node is checked. If the HUT of that node is a frequent itemset, no subset of its HUT needs to be checked again. Then all subtrees with that node as the root node are pruned. The implementation process is as follows. The frequent itemsets in an item are sorted to obtain a list. The obtained table is traversed in ascending order to obtain all prefix paths containing that node, and all paths on FP Tree including that point are obtained by ending at that point. If only a unique path is obtained through the previous step, build a frequent itemset. If not, the items within the path need to be merged to build a completely new process, repeating this process until there is only one path for all itemsets. Specifically, if transaction item $I$ is not frequent on a certain path, then $I$ has prefix paths $L_1$, $L_2$ in FP-Tree. $L_1 \in L_2$. The $I$ in path $L_2$ can be fused with path $L_1$. This operation can effectively reduce

the iterations. The support of frequent term $I$ in set $L_1$ and set $L_2$ is $m$ and $n$, respectively. The conditional pattern basis $W$ when pruning strategy is not used is shown in Eq. (5).

$$W = \{(L_1: m), (L_2: n)\} \tag{5}$$

The frequent pattern generated at this time is shown in formula (6).

$$M = L_1: m + n \tag{6}$$

The conditional pattern base of item $I$ after adopting pruning strategy is shown in Eq. (7).

$$W' = \{(L_1: m + n)\} \tag{7}$$

The frequent pattern generated at this time is shown in Eq. (8).

$$M' = L_1: m + n \tag{8}$$

The specific implementation process of FP-Tree after adopting pruning strategy is shown in Fig. 3.

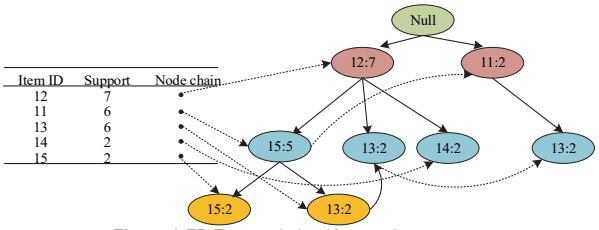



**Figure 3** FP-Tree optimized by pruning strategy

In Fig. 3, node $I$ has two paths, $I_1$ and $I_2$. The two belong to an inclusive relationship. Based on the principle of pruning strategy, $I_2$ in the second path can be merged with the previous path and the node can be deleted. After pruning each node, the FP-Tree shown in Fig. 2 can be obtained.

### 3.2 Construction of Medical Data Model Based on IFP-Growth

With the continuous progress of internet technology, the demand for data processing is constantly strengthening. In this situation, Hadoop technology has emerged. Hadoop is a computing method based on cloud computing as the basic framework. It is characterized by scalability, low cost, efficiency, and strong reliability. Hadoop is developed on the basis of Google's cluster system, mainly including MapReduce and HDFS. MapReduce can more easily achieve the expected goals when dealing with items with a large amount of medical data [19]. Meanwhile, the MapReduce also has unique advantages in universality. The MapReduce model mainly consists of two modules, namely Map and Reduce. The specific operating mode is shown in Fig. 4.

During the operation of MapReduce, the input and output of the raw data are presented in the form of (*key*, *value*). In the functional module of Map, the raw data is processed through corresponding instructions. Based on the Map function, the initial key-value pairs are converted into temporary key-value pairs.

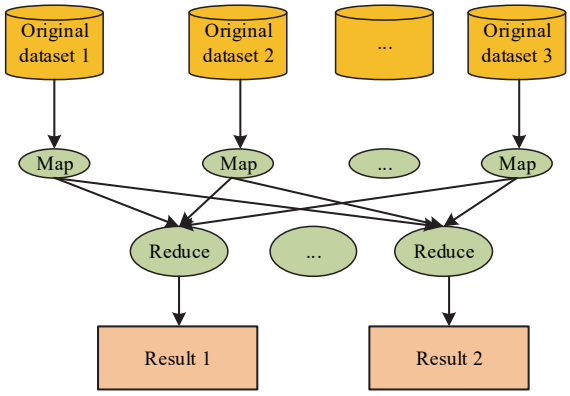



**Figure 4** Running mode of MapReduce

Under this operation, the original data remains unchanged. Different Maps belong to a membership relationship. In the Reduce operation, the same *key* value in the temporary key value pair is integrated. In the specific operation process, the Map function and Reduce function are determined based on actual needs. If the medical event related data pairs input in the established MapReduce model are (1, 3) and (2, 5) respectively, the Value values in the model need to be squared and then mapped, resulting in temporary key value pairs (3, 49) and (2, 25). Finally, the object data is summarized and calculated. The execution process of the MapReduce model is shown in Fig. 5.

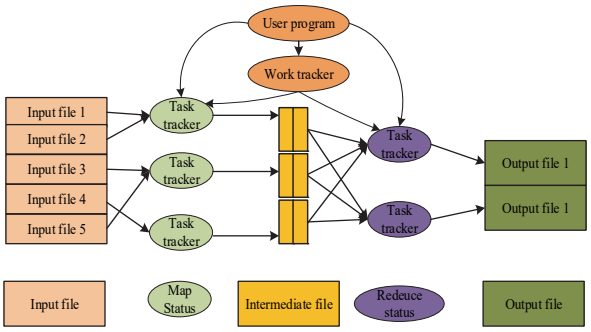



**Figure 5** Execution process of MapReduce model

The framework structure of the MapReduce model consists of a main node and multiple sub nodes. The main node is mainly responsible for assigning corresponding tasks to the currently idle nodes to execute the task. Segmentation points are used to complete the work tasks assigned by the main node. In the MapReduce model, a task may be divided into multiple different subtasks. At the same time, the Redcut node also needs to analyze and integrate the actual task processing results of each sub node to optimize the model and overall work efficiency. Therefore, the MapReduce model in Hadoop is used as a data processing framework to parallelize the FP-Growth algorithm. Based on Hadoop and improved FP-Growth medical data analysis model (IFP-Growth) is constructed. In the execution of the MapReduce model, the file data is first input and divided to form multiple different blocks. Different work tasks are executed separately. The same intermediate files are merged. The final processing result is output[20, 21]. The cloud computing medical big data

algorithm based on Hadoop's and FP-Growth includes the following basic processes. The first step is to use raw data to calculate frequent itemsets and encode them. Then, the original data is grouped based on the encoded frequent itemset. The frequent pattern trees are constructed for grouped data. Finally, each constructed frequent pattern tree is used to mine frequent itemsets, integrate the mined frequent itemsets, and obtain the final result. Each grouped data can use a node in the cluster to run FP Growth while performing tree building and mining operations, achieving parallelism. Finally, the frequent itemsets mined by each child node are integrated to obtain the final result.The parallelization processing of FP-Growth based on the MapReduce model is shown in Fig. 6.

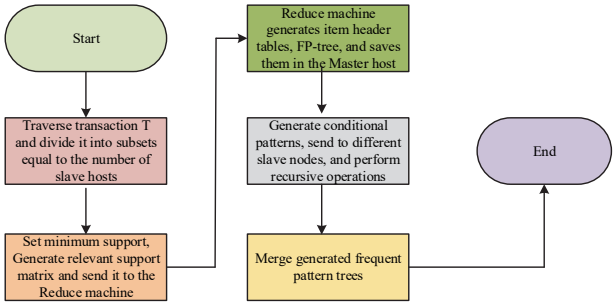



**Figure 6** Parallelization of FP growth based on MapReduce model

The set of things *key* is sequentially traversed to obtain a subset equal to the sub nodes. Each slave host traverses a subset of things in medical data through minimum support, generating frequent itemsets and support matrices. The obtained results are sent to the Reduce host in the form of key value pairs. Next, the Reduce host generates an FP-tree and the data support matrix based on the received data. Then, the item header table is generated in the form of frequent itemsets. Finally, FP-tree is recursively mined. The conditional patterns are generated based on each item in the item header table. The recursive operations are performed on them. Slave nodes perform recursive operations on the basis of conditional patterns and merge the generated frequent patterns to obtain all association rules between relevant medical data.

## 4 EXPERIMENTAL ANALYSIS OF MEDICAL DATA MODEL BASED ON IFP-GROWTH

To verify the performance of the proposed IFP-Growth in medical big data processing, a corresponding experimental environment was constructed to verify the performance of this method. In addition, real medical data was collected to analyze the application effectiveness of this method.

### 4.1 Performance Verification of Medical Data Analysis Models

To verify the performance of the proposed algorithm in large-scale medical data processing, a corresponding experimental environment is established. The experimental environment used for the study is as follows. The operating system is a Linux system, which includes a Hadoop cluster with 4 nodes (1 master node and 3 slave nodes). The hardware configuration is also the same. The experimental data used in the study came from several medical datasets in the UCI dataset, including the Diabetes, Heartstatlog, and Blood datasets, all collected from relevant patient data records from 130 hospitals in the United States from 1999 to 2008. To ensure the effectiveness of the rules and the quality of the experiment, the experimental environment design is shown in Tab. 1.

**Table 1** Experimental parameter settings

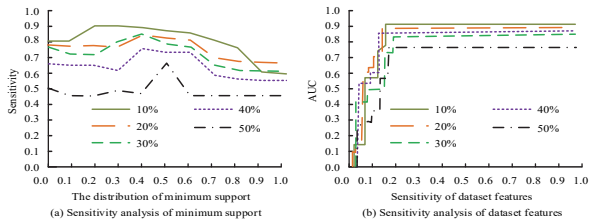
| Projects | Parameters |
|---|---|
| CPU | Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10 GHz |
| Hard disk | 1 TB |
| Network | 100 mbps |
| Memory | 32 GB |
| Operating system | CentOS7.4 |
| JDK version | 1.8 |
| Hadoop | 2.7.6 |
| Spark | 2.4.0 |
| HBase version | 1.1.9 |

To verify the performance of the IFP-Growth in different environments, the clustered environment was used as the comparitive condition. This algorithm was run in both clustered and distributed environments. In the Diabetes medical disease data sample, each experimental environment was run 5 times. The time consumption of this method in different operating environments was shown in Tab. 2. From Tab. 2, in traditional environments, the Hadoop-FP-Growth method proposed in the study had significantly lower time consumption when processing different numbers of data samples compared to the pre-improved method. In a distributed environment, the performance was also better than the unimproved method.

**Table 2** Time consumption in different environments

| Number of sets of things / pieces | Clustered environment / ms | | Distributed environment / ms | |
|---|---|---|---|---|
| | FP-Growth | IFP-Growth | FP-Growth | IFP-Growth |
| 5000 | 357 | 154 | 45 | 29 |
| 10000 | 405 | 186 | 72 | 54 |
| 15000 | 462 | 257 | 109 | 96 |
| 20000 | 547 | 299 | 128 | 114 |
| 25000 | 611 | 348 | 146 | 127 |
| 30000 | 656 | 371 | 185 | 138 |
| 35000 | 689 | 395 | 223 | 151 |

The sensitivity of minimum support and dataset feature parameters is analyzed. Fig. 7a shows the sensitivity test for minimum support, and Fig. 7b shows the sensitivity test for dataset features. Extract experimental samples from the dataset at a ratio of 10% - 50% to evaluate the performance of the model. In Fig. 7a, when the support level is below 0.5, the performance of the model continuously optimizes. When the support level is higher than 0.5, the performance of the model decreases. In Fig. 7b, the AUC values of the model remain stable above 0.65 in different sample ratios, indicating a low error rate of the model.




(a) Sensitivity analysis of minimum support (b) Sensitivity analysis of dataset features

**Figure 7** Parameter sensitivity test

Subsequently, the operational efficiency of the proposed method under different support conditions was analyzed. The Apriori algorithm and FreeSpan algorithm were used as comparative methods to analyze the proposed method. The results were shown in Fig. 8. All three methods were tested in the Diabetes, Heartstatlog, and Blood datasets. As the support level increased, the running time of all three methods showed a significant decrease. In the Blood dataset, when the support level increased to 0.3, the running time of the IFP-Growth was 4.82 s. The running time of Apriori and FreeSpan algorithms was 7.56 ds and 9.49 s respectively. The efficiency of the method proposed in the study is significantly better than the other two methods during operation. This method consumes less time, which is more efficient in the experiment.

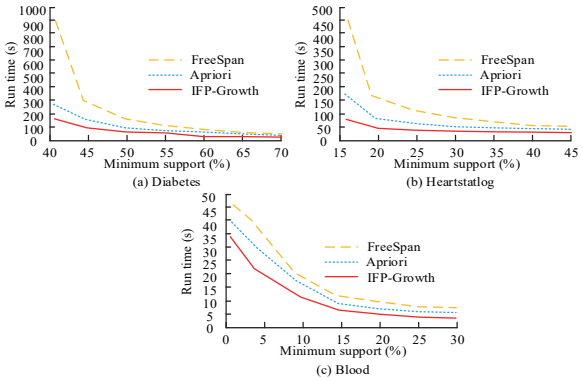



**Figure 8** Comparison of runtime on different datasets

The memory usage of the proposed method when running on different datasets was shown in Fig. 9. In Fig. 9, the memory consumption of the proposed method was always the smallest in the three datasets used for experiments. As support increased, the memory usage of the three medical tree association rule mining algorithms gradually decreased. In the Diabetes dataset, when the support level reached 0.7, the memory usage of the proposed method in the study was 5100 MB. The memory usage of FreeSpan and Apriori algorithms was 16 × 100 MB and 12 × 100 MB, respectively. By comparison, the method proposed in the study consumes less memory during operation. The model can achieve better performance by consuming fewer resources during runtime.

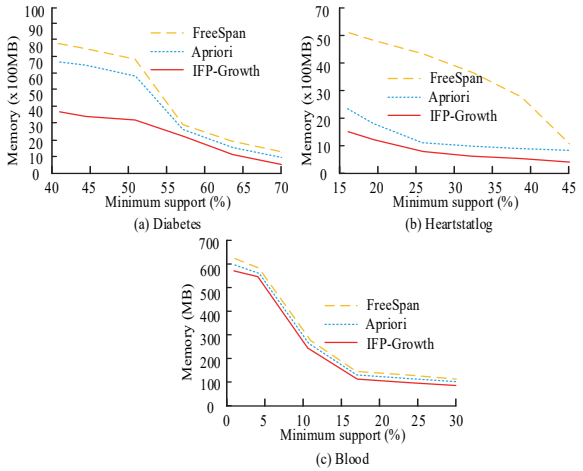



**Figure 9** Comparison of memory usage on different datasets

## 4.2 The Application Effect of Medical Data Analysis Models

To verify the application effect of the IFP-Growth in actual medical data, the Diagnosis, Agaricus-leptota, Break cancel, and Clinical-cancer datasets were used as examples for analysis. The average values of 5 experiments on the Apriori, FreeSpan, and the IFP-Growth algorithm were statistically analyzed. The results were shown in Fig. 10. In Fig. 10, as support increased in all datasets, the runtime of the proposed method significantly decreased. Taking the Break-Cancer dataset as an example, when the support level was 0.22, the time consumption was 0.04 s. The time consumption of the other three methods was 0.19 s, 0.05 s, and 0.05 s, respectively. Comparison shows that the proposed method has significantly lower time consumption during operation compared to other methods. The operational efficiency of this method is significantly better than the other three methods. It can more quickly mine the association rules and connections between data samples in medical data sample processing.

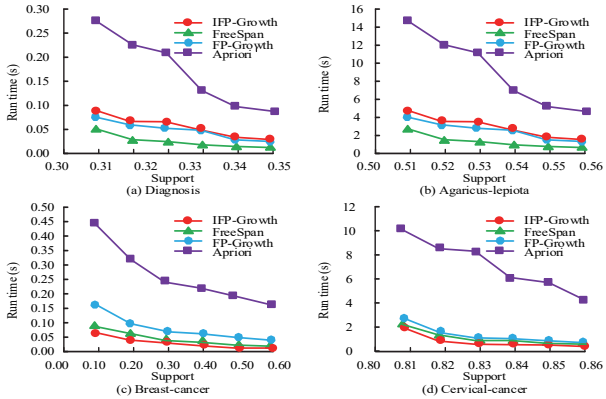



**Figure 10** Comparison of operational efficiency in different datasets

Diagnosis, Agaricus-leptota, Break cancel, and Clinical-cancer were used as datasets. The number of rules generated in each dataset was compared. The results were shown in Fig. 11. In Fig. 11, under different dataset conditions, the IFP-Growth method proposed in the study generated the most abundant number of rules between the data. In the Diagnosis dataset, when the support level was 0.31, the proposed method generated 267 rules. The rules generated by the other three methods were 71, 126, and 233, respectively. The method proposed in the study can more effectively mine the association rules between medical data, analyze the association rules between medical data, and achieve effective data feature analysis.

The predictive performance and feature selection ability of different methods are shown in Fig. 12. Fig. 12a shows the predictive performance of different methods. Fig. 12b shows the feature selection ability of different methods. In Fig. 12a, the predictive performance of IPF Growth, FreeSpan, PF Growth, and Apriori are 93.05%, 85.26%, 81.07%, and 73.64%, respectively. The above data indicates that the designed method has more significant advantages in predictive performance under different support levels. It means that the proposed method can more effectively present the intrinsic connections in medical datasets. In Fig. 12b, the feature selection abilities of IPF Growth in the Diagnosis, Agaricus leptota, Breast cancer, and Clinical cancer datasets are 82.35%, 76.41%,

67.52%, and 78.92%, respectively. In different datasets, this method exhibits more obvious feature selection advantages, indicating that it can more effectively identify the features and connections between data, thereby achieving association rule mining between medical data.

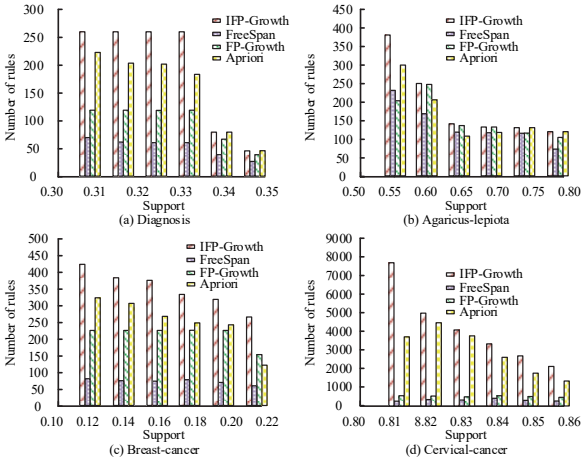



**Figure 11** Number of rules generated in different datasets

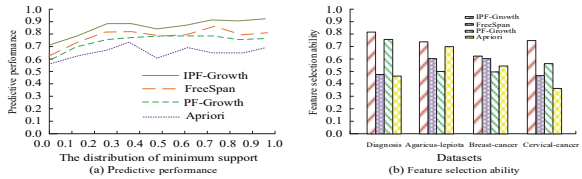



**Figure 12** Prediction performance and feature selection ability of different methods

The study further validates the model performance by collecting detection record data of hypertensive patients in a hospital in a certain region. A total of 5563 patient data are collected. The collected data is preprocessed using methods such as noise reduction or deletion to clean up the default and noise values. The remaining 5000 data are used for model application testing. Then the Python is used to extract the data and obtain the required dataset for the experiment, ensuring data quality. To ensure the effectiveness of the rules and the quality of the experiments, stability and scalability are used to evaluate the research method. Fig. 13a shows the stability testing of different methods on real datasets. Fig. 13b shows the scalability testing of different methods on real datasets. In Fig. 13a, the stability test results of IPF Growth, FreeSpan, PF Growth, and Apriori are 91.23%, 86.10%, 78.46%, and 73.82%, respectively, indicating that the research method has better stability. Meanwhile, the research method has relatively ideal stability performance for different datasets. In Fig. 13b, the scalability test results of IPF Growth, FreeSpan, PF Growth, and Apriori are 88.76%, 82.51%, 76.32%, and 68.95%, respectively. This indicates that the designed method has better load capacity, indicating that the model has good problem-solving performance.

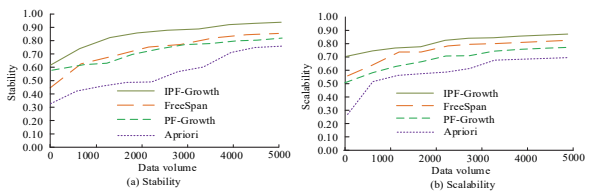



**Figure 13** Comparison of model performance on real datasets

## 5 CONCLUSION

Analyzing the association rules can effectively analyze the features and explore the implicit relationships between data. To better analyze the association rules between medical data, the frequent pattern growth algorithm is improved. Then it is combined with Hadoop. An association rule mining algorithm is constructed for medical data. Taking the Break-Cancer dataset as an example, when the support level was 0.22, the time consumption was 0.04 s. The time consumption of the other three methods was 0.19 s, 0.05 s, and 0.05 s, respectively. In the Diabetes dataset, when the support level reached 0.7, the memory usage of the proposed method was 5100 MB. The memory usage of FreeSpan and Apriori algorithms was significantly higher at 16 × 100 MB and 12 × 100 BMB, respectively, compared to the methods proposed in the study. In the Diagnosis dataset, when the support level was 0.31, the proposed method generated 267 rules. The rules generated by the other three methods were 71, 126, and 233, respectively. This indicates that the proposed method could more effectively mine association rules between medical data, achieving effective data feature analysis. At present, the proposed method has been tested in both distributed and clustered environments. The medical data mining method proposed in the study can effectively analyze the characteristics of patient data, optimize the work efficiency of doctors, improve patient recovery, and provide reference for the prevention and treatment of related diseases.At present, the proposed method has been experimentally tested in both distributed and cluster environments. However, the operational efficiency in cluster environments still needs to be improved, which is a future research direction. In addition, for the real patient data, the study only used the relevant data of diabetes patients in a certain area for the test. In future research, more types of disease sample data will be collected to verify their performance. The association rules between data under different disease conditions will be analyzed.

## 6 REFERENCES

[1] Happawana, K. A. & Diamond, B. J. (2022). Association rule learning in neuropsychological data analysis for Alzheimer's disease. *Journal of Neuropsychology, 16*(1), 116-130. https://doi.org/10.1111/jnp.12252

[2] Ebrahimnejad, A., Ghiyasi, M., Nikhbakht M., & Najjarian H. (2023). Fuzzy Goal Programming Approach for Identifying Target Unit in Combined-Oriented DEA Models: Application in Bank Industry. *Economic Computation and Economic Cybernetics Studies and Research, 57*(1), 5-22. https://doi.org/10.24818/18423264/57.1.23.01

[3] Banihashemi, S. A. & Khalilzadeh, M. (2023). Performance Evaluation Optimization Model with a Hybrid Approach of NDEA-BSC and Stackelberg Game Theory in the Presence of Bad Data. *Economic Computation and Economic Cybernetics Studies and Research, 57*(2), 293-312. https://doi.org/10.24818/18423264/57.2.23.18

[4] Chanpa R., Jamali M. A. J., Hatamlou A., & Anari B. (2023). Cluster Head Selection Algorithm on the Basis of Mass Defense of Bees in IoT. *Economic Computation and Economic Cybernetics Studies and Research, 57*(3), 187-202. https://doi.org/10.24818/18423264/57.3.23.11

[5] GOYAL, P., Verma, D. K., & Kumar, S. (2023). Diagnosis of Plant Leaf Diseases Using Image Based Detection and Prediction Using Machine Learning Approach. *Economic Computation and Economic Cybernetics Studies and Research*, *57*(4), 293-312. https://doi.org/10.24818/18423264/57.4.23.18

[6] Ugwu, N. V. &Udanor, C. N. (2021). Achieving Effective Customer Relationship using Frequent Pattern-Growth Algorithm Association Rule Learning Technique. *Nigerian Journal of Technology, 40*(2), 329-339. https://doi.org/10.4314/njt.v40i2.19

[7] Satria C., Anggrawan A., & Mayadi. (2023). Recommendation System of Food Package Using Apriori and FP-Growth Data Mining Methods. *Journal of Advances in Information Technology*, *14*(3), 454-462. https://doi.org/10.12720/jait.14.3.454-462

[8] Wang, C., Bian, W., Wang, R., Chen, H., Ye, Z., & Yan, L. (2020). Association rules mining in parallel conditional tree based on grid computing inspired partition algorithm. *International Journal of Web and Grid Services, 16*(3), 321-339. https://doi.org/10.1504/IJWGS.2020.109475

[9] Khatir, M. R., Lebbah, Y., & Nourine, R. (2020). A pattern-growth approach for mining trajectories. *Multiagent and Grid Systems, 16*(2), 117-133. https://doi.org/10.3233/MGS-200324

[10] Jamsheela, O. & Raju, G. (2022). SR-mine: Adaptive transaction compression method for frequent itemsets mining. *Arabian Journal for Science and Engineering, 47*(8), 9641-9657. https://doi.org/10.1007/s13369-021-06298-9

[11] Wu, X., Zhang, Y., Wang, A., Shi, M., Wang, H., & Liu, L. (2022). MNSSp3: Medical big data privacy protection platform based on Internet of things. *Neural Computing and Applications, 34*(14), 11491-11505. https://doi.org/10.1007/s00521-020-04873-z

[12] Shafqat, S., Majeed, H., Javaid, Q., & Ahmad, H. F. (2022). Standard ner tagging scheme for big data healthcare analytics built on unified medical corpora. *Journal of Artificial Intelligence and Technology, 2*(4), 152-157. https://doi.org/10.37965/jait.2022.0127

[13] Gou, X. & Xu, Z. (2021). An overview of Big Data in Healthcare: multiple angle analyses. *Journal of Smart Environments and Green Computing, 1*(3), 131-145. https://doi.org/10.20517/jsegc.2021.07

[14] Ramachandran, S. K. & Manikandan, P. (2021). An efficient ALO-based ensemble classification algorithm for medical big data processing. *International Journal of Medical Engineering and Informatics, 13*(1), 54-63. https://doi.org/10.1504/IJMEI.2021.111864

[15] Hurley, D. & Popescu, G. H. (2021). Medical big data and wearable internet of things healthcare systems in remotely monitoring and caring for confirmed or suspected COVID-19 patients. *American Journal of Medical Research, 8*(2), 78-90. https://doi.org/10.22381/ajmr8220216

[16] Lefa, M., Hatem, A. B. D., & Salem, R. (2022). Enhancement of Very Fast Decision Tree for Data Stream Mining. *Studies in Informatics and Control, 31*(2), 49-60. https://doi.org/10.24846/v31i2y202205

[17] Carter, D., Kolencik, J., & Cug, J. (2021). Smart internet of things-enabled mobile-based health monitoring systems and medical big data in COVID-19 telemedicine. *American Journal of Medical Research, 8*(1), 20-29. https://doi.org/10.22381/AJMR7220205

[18] Fang, Y., Luo, B., Zhao, T., He, D., Jiang, B., & Liu, Q. (2022). ST-SIGMA: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting. *CAAI Transactions on Intelligence Technology, 7*(4), 744-757. https://doi.org/10.1049/cit2.12145

[19] Maier, M. I., Czibula, G., & Delean, L. R. (2023). Using Unsupervised Learning for Mining Behavioural Patterns from Data. A Case Study for the Baccalaureate Exam in Romania. *Studies in Informatics and Control, 32*(2), 73-84. https://doi.org/10.24846/v32i2y202307

[20] Danjuma, M. U., Yusuf, B., & Yusuf, I. (2022). Reliability, availability, maintainability, and dependability analysis of cold standby series-parallel system. *Journal of Computational and Cognitive Engineering, 1*(4), 193-200. https://doi.org/10.47852/bonviewJCCE2202144

[21] Zamfiroiu, A., Sharma, R. C., Constantinescu, D., PANĂ, M., & Toma, C. (2022). Using Learning Analytics for Analyzing Students' Behavior in Online Learning. *Studies in Informatics and Control, 31*(3), 63-74. https://doi.org/10.24846/v31i3y202206

**Contact information:**

**Rong HU**, Associate Professor
(Corresponding author)
School of Intelligence Technology, Geely University of China, Chengdu, Sichuan, 641423, P. R. China;
No. 123, SEC.2, Chengjian Avenue, Eastern New District, Chengdu City, Sichuan Province
E-mail: 727749104@qq.com

**Xueling YANG**, Master lecturer
School of Intelligence Technology, Geely University of China, Chengdu, Sichuan, 641423, P. R. China;
No. 123, SEC.2, Chengjian Avenue, Eastern New District, Chengdu City, Sichuan Province
E-mail: 498452989@qq.com