INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

A data mining method for biomedical literature based on association rules algorithm

Xiaofeng Shi, Yaohong Zhao, Haijuan Du

DOI: 10.1504/IJDMB.2023.10057445

Article History:

Received:	30 December 2022
Last revised:	25 April 2023
Accepted:	04 May 2023
Published online:	22 January 2024

A data mining method for biomedical literature based on association rules algorithm

Xiaofeng Shi*

Centre of Modern Education Technology,
Changchun Institute of Technology,
Changchun, 130012, China
Email: 0218024@ccit.edu.cn
*Corresponding author

Yaohong Zhao

Faculty of Computer Science and Technology,
Changchun University,
Changchun, 130022, China
Email: 234581000@qq.com

Haijuan Du

School of Information Science and Engineering,
Jiaxing University,
Jia'xing, 314001, China
Email: hjuan_du@126.com

Abstract: There are problems in the process of biomedical literature data mining, such as high data noise, low mining accuracy, and long-time consumption. Therefore, a biomedical literature data mining method based on association rule algorithm was designed. First, set up the extraction process of biomedical literature data, introduce the factor graph decomposition global extraction function, and establish a probabilistic database to speed up the extraction. Secondly, wavelet transform is used to denoise the data, improve the effectiveness of the extracted data, and classify it based on its importance. Finally, by setting association rules for biomedical literature data mining and introducing pre pruning methods on this basis, the time cost of calculating support is reduced, mining efficiency is improved, and combining confidence and dependency, a biomedical literature data mining model based on association rules is constructed to achieve the final mining. The results show that this method improves the accuracy of literature mining, reaching 99%, and effectively reduces the mining time, with a maximum time consumption of 1.7 seconds. It has strong application performance.

Keywords: association rules; biomedical literature; data mining; wavelet transform; vector space; classification basis.

Reference to this paper should be made as follows: Shi, X., Zhao, Y. and Du, H. (2024) 'A data mining method for biomedical literature based on association rules algorithm', *Int. J. Data Mining and Bioinformatics*, Vol. 28, No. 1, pp.1–17.

Biographical notes: Xiaofeng Shi received his MS in Science and Technology of Computer from Northeast Normal University in 2005. He is currently an Associate Professor in the College of Computer Technology and Engineering of Changchun Institute of Technology. His research interests include big data technology, research on intelligent algorithm, and internet of things technology.

Yaohong Zhao received her MAEng in Computer and Applied Science from Jilin University in 2005. She is currently a Lecturer in Faculty of Computer Science and Technology of Changchun University. Her research interests include image processing, and cloud computing technology.

Haijuan Du received her Master's degree from the School of Software, Liaoning University of Engineering and Technology in 2008, and is currently a Lecturer in the School of Information Science and Engineering, Jiaxing University. Her research interests include algorithms, data processing and modern educational technology.

1 Introduction

Biomedical literature contains a lot of key information such as medical biological genes, which is important information summarised through hundreds of millions of clinical and other practical experiences (Karatzas et al., 2022). In recent years, the data volume of biomedical literature has been expanding, and the published biomedical literature data has gradually become an important medical knowledge base. The application of biomedical literature data has benefited mankind and provided strong support for the treatment of human difficult and miscellaneous diseases (Cox et al., 2020). Therefore, the effective mining of biomedical literature data has become an important research direction in the biomedical community. By using certain technologies to mine and analyse medical literature data, the new discovery of literature knowledge is finally realised (Feng and Gao, 2022). However, with the increasing amount of medical literature, the credibility of research results needs to be measured, and the effective mining of biomedical literature data is increasingly difficult (Li et al., 2021). In order to mine more effective knowledge and information, researchers in this field have studied many methods for document data mining, and made some achievements.

Momeni et al. (2020) proposed a data mining method based on feature discretisation. Use wavelet packet transform to extract data features, and preprocess the extracted features. Secondly, use supervised discretisation technology to convert continuous features into finite interval features, so as to realise data mining. However, this method did not handle too much data noise during the mining process, which has certain drawbacks. Wu et al. (2021) proposed a feature association data mining method based on machine learning algorithms, which is used for feature association and discretises continuous data feature attributes. The binary representation of data features is extended to ensure the diversity of data feature attributes. Finally, a heuristic feature mining method with minimum support is adopted to achieve data mining. This method effectively improves the accuracy of data mining, but it requires a longer sample cost and has certain limitations. Wu and Chen (2021) designed a data mining method for

structures with uncertainty of design variables. Based on the joint probability distribution of engineering design variables, a new uncertain data decision tree (DTUD) method is developed. And nine datasets are selected from the available repositories and compared with the traditional decision tree to verify its high accuracy and effectively generate the design with expected performance. The uncertain structure decision tree is used to complete the document data mining, and the research on the implementation method is based on the main components of the data. This method has good data decomposition effect in the process of document mining, but it takes a long time and has some limitations.

Therefore, in order to solve the problems of high data noise, low mining accuracy, and long-time consumption in the mining process of existing data mining methods, and improve the mining efficiency of biomedical literature data, a biomedical literature data mining method based on association rule algorithm is proposed.

The main steps of this method research are as follows:

Step 1 First, determine the distribution law of biomedical literature data by setting the extraction process of biomedical literature data, introduce the factor graph decomposition global extraction function, and establish a probabilistic database to reduce the difficulty of biomedical literature data extraction, accelerate the extraction speed, and achieve the extraction of biomedical literature data.

Step 2 Secondly, based on the above, analyse the data noise level in the extracted biomedical literature dataset, and achieve denoising of biomedical literature data through wavelet transform to improve the effectiveness of the extracted data, and place the noise reduced literature data in different dimensional feature vector spaces, complete the literature data classification, achieve the biomedical literature data preprocessing, lay the foundation for subsequent data mining, and improve the mining accuracy and mining speed.

Step 3 Finally, by setting association rules for biomedical literature data mining and introducing a pre pruning method on this basis, the number of candidate data items is reduced, and the frequency of calculating their support is avoided, thereby reducing time overhead and improving mining efficiency. And combining confidence and dependency, construct a biomedical literature data mining model based on association rules to achieve the final mining. Then, based on the support results, the confidence level of the literature data items is calculated, and by determining the dependencies between the data in the project, a biomedical literature data mining model based on association rules is established to achieve the final mining.

Step 4 By setting the experimental environment, parameters and indicators, and by means of comparative experiments, the proposed method, Wu et al. (2021) and Wu and Chen (2021) methods are compared in terms of data noise, mining accuracy and mining time to verify the feasibility of the proposed method.

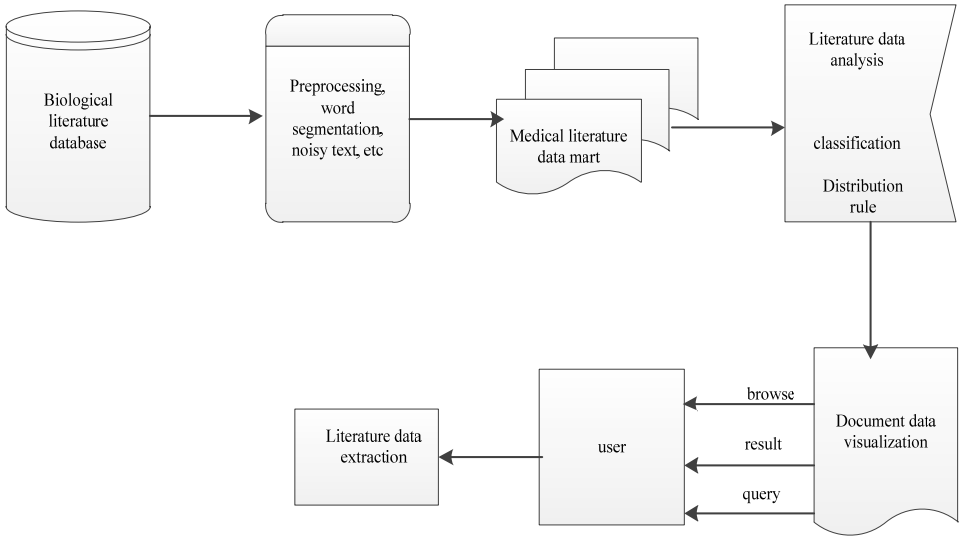
2 Research on biomedical literature data extraction and preprocessing

2.1 Biomedical literature data extraction research

In order to realise biomedical literature mining, its data needs to be preprocessed to lay the foundation for subsequent mining. Before data preprocessing, the data needs to be extracted first. In the mining process, due to the large number and variety of biomedical literature data, the process of biomedical literature data extraction is more complex (Du et al., 2021). In order to simplify the process and improve the mining effect of biomedical literature data, this paper uses a factor graph to extract relevant biomedical literature data.

In biomedical literature data, gene protein and other key data are the most critical data, and also the most in-depth research field in the medical field. Therefore, the role of relevant data extracted from biomedical literature is of great significance (Carrasco et al., 2021). Among them, biomedical literature data needs to be extracted from entity recognition, knowledge discovery, data visualisation and other methods. The general process of biomedical literature data extraction is shown in Figure 1.

Figure 1 Schematic diagram of biomedical literature data extraction process



In the biomedical literature data extraction, it is mainly aimed at the identification and extraction of key professional terms such as biological genes and proteins in the medical field. In this extraction process, there are some difficulties in the extraction rules and description of literature data. Therefore, in order to reduce the difficulty of biomedical literature data extraction and accelerate the extraction speed, this paper uses factor graph to effectively extract biomedical literature data. Factor graph (Ali et al., 2021) is a vector graph that factorises a function.

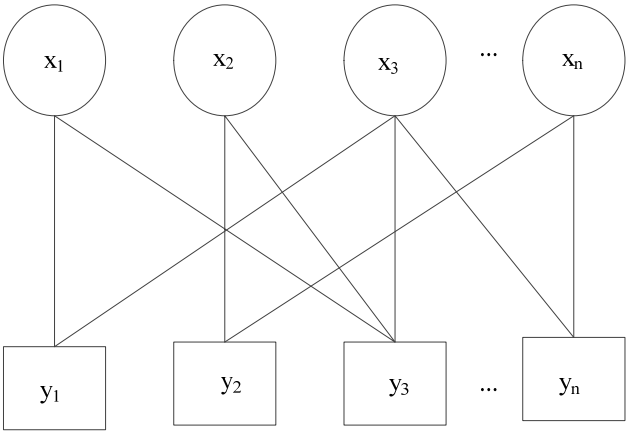
In the process of factorisation, two key vocabulary nodes, document data variable nodes and function nodes are set. It is an extraction method that decomposes a global function into the product of multiple local functions through the formula, and then reflects these factor functions and corresponding variables on the factor graph. The functional relationship between all points and edges of all medical biological data in the image is shown in Figure 2.

According to the relationship function set in Figure 2, set the factorisation equation expression of a global function in biomedical literature data extraction as:

$$g\left(x_1, x_2, ..., x_n\right) = y_1\left(x_1\right) y_2\left(x_1, x_2\right) y_3\left(x_1, x_2, x_3\right), ..., y_n\left(x_1, x_2, x_3, ..., x_n\right) \tag{1}$$

In formula (1), $g(x_1, x_2, \ldots, x_n)$ represents the global function of the biomedical literature data extraction, $y_1, y_2, y_3, \ldots, y_n$ represents each data edge of the biomedical literature, $x_1, x_2, x_3, \ldots, x_n$ represents each data point of the biomedical literature.

Figure 2 Functional relationship between medical biological literature data points and sidelines



Since the global extraction of biomedical literature data is more complex in factor decomposition (Trieu et al., 2021), in order to simplify its decomposition process, the global function is expressed as:

$$g(X) = \prod_{x \in Y} Y_n\left(X_n\right) \tag{2}$$

In formula (2), $g(X)$ represents the simplified extracted global function expression, Y_n represents the collection of edges, and X_n represents the collection of biomedical literature data points.

In the extraction of biomedical literature data, the most important thing is to separate data according to the changes of random variables in the factor graph. Generally, a probabilistic (Krawczyk et al., 2021) database is constructed based on the association between biomedical literature data. The constructed probabilistic database of biomedical literature data can be expressed as:

$$A(X) = \{D, R\} \tag{3}$$

In formula (3), $A(X)$ represents the probabilistic database of biomedical literature data, D represents the association mode of biological data literature data, and R represents the user mode of biological data literature.

During document data extraction, the data in the intersection process will be captured. Therefore, the document data captured in the process of biomedical document data interchange is defined as:

$$c_i(X) = \left\{ a \mid fid \in \left\{ X_n, y_n \right\} \right\} \tag{4}$$

In formula (4), $c_i(X)$ represents the captured literature data in the intersection, and a represents the random variable.

On this basis, the whole process of biomedical literature data extraction is realised, and the results are as follows:

$$P(x) = \exp\left\{ \sum\sum\nolimits_{i=x} v(X) \left[\frac{\left\{ X_n, y_n \right\}}{a} \right] \right\} \tag{5}$$

In formula (5), $P(x)$ represents the global data of the extracted biomedical literature, $\exp\{\}$ represents the data allocation coefficient, and $v(X)$ represents the partition function.

In the process of biomedical literature data extraction, the distribution rule of biomedical literature data is determined by setting the process of biomedical literature data extraction, the factor graph decomposition global extraction function is introduced, and a probabilistic database is constructed to finally achieve biomedical literature data extraction. Next, preprocess the extracted data to lay a foundation for subsequent mining and improve the mining precision and speed.

2.2 Biomedical literature data preprocessing research

In the process of implementing biomedical literature data mining, due to the interference of noise and other factors, the effectiveness of the above globally extracted biomedical literature data mining cannot be improved (Chen et al., 2020), resulting in low data precision. Therefore, this chapter needs to preprocess the above extracted data. In this preprocessing, the whole preprocessing process is realised mainly through document data noise reduction and classification.

The dataset of biomedical literature data extraction is determined as:

$$H_i = \left\{ h_1, h_2, ..., h_m \right\} \tag{6}$$

In formula (6), H_i represents the extracted dataset of biomedical literature data, and $\{h_1, h_2, \ldots, h_m\}$ represents the literature data composition in the set.

Analyse the noise level of all data in the biomedical literature dataset, namely:

$$z_i = \begin{cases} \sum \dfrac{u_i}{H_i} & > 1 \\ \sum \dfrac{u_i}{H_i} & < 1 \end{cases} \tag{7}$$

In formula (7), z_i represents the noise degree of the data, and u_i represents the noise coefficient in the dataset extracted from biomedical literature data.

When the data is greater than 1, it means that the data noise is high at this time, so denoising is required. When it is less than 1, it is regarded as normal medical literature data with no denoising processing.

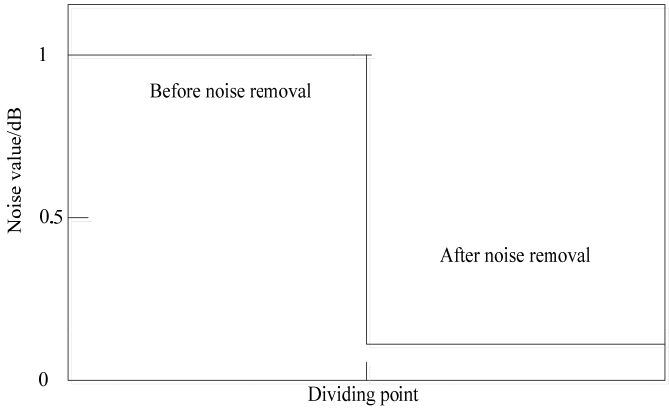
According to the determined noise level of biomedical literature data, denoise the noisy data, and denoise the noisy data by wavelet transform (Hillenmeyer et al., 2021; Sokhangoee and Rezapour, 2022) to remove the interference in the data. The noise reduction results are as follows:

$$\rho(z_i) = \frac{2}{\sqrt{3}}\pi^{0.25}\left(1-\sum\frac{u_i}{h_m}\right)^2 e \tag{8}$$

In formula (8), $\rho(z_i)$ represents the biomedical literature data after noise reduction, and e represents the wavelet base.

The change diagram of noise reduction curve of biological literature data after wavelet transform is shown in Figure 3.

Figure 3 Schematic diagram of changes in noise reduction curve of biomedical literature data after wavelet transform



Next, in order to facilitate the mining process, the biomedical literature data after the above denoising is classified and processed to speed up its mining speed. Put the literature data in the vector space, and express different objects in different dimensional feature spaces (Gu et al., 2022), namely:

$$d_i = \left(D_{1i}, D_{2i}, ..., D_{ni}\right) \tag{9}$$

In formula (9), d_i represents different dimensional feature vector space description, (D_{1i}, D_{2i}, …, D_{ni}) represents different types of biomedical literature data vector values.

In this vector space, calculate the importance of each biomedical literature data vector (Guo et al., 2020) as the basis for its literature data classification, namely:

$$L_i = \frac{\sigma}{\sum_{i\in k} d_i} \tag{10}$$

In formula (10), L_i represents the importance result of the biomedical literature data vector, σ_i represents the word frequency of the biomedical literature data vector, and k represents the number of total spatial dimensions.

On the basis of the above analysis, the classification of biomedical literature data is finally realised through reverse file frequency (Li and Li, 2021), and the classification formula is:

$$f(d_i) = \log\left|\frac{\sigma_i}{\{L_i \in \alpha\}}\right| \tag{11}$$

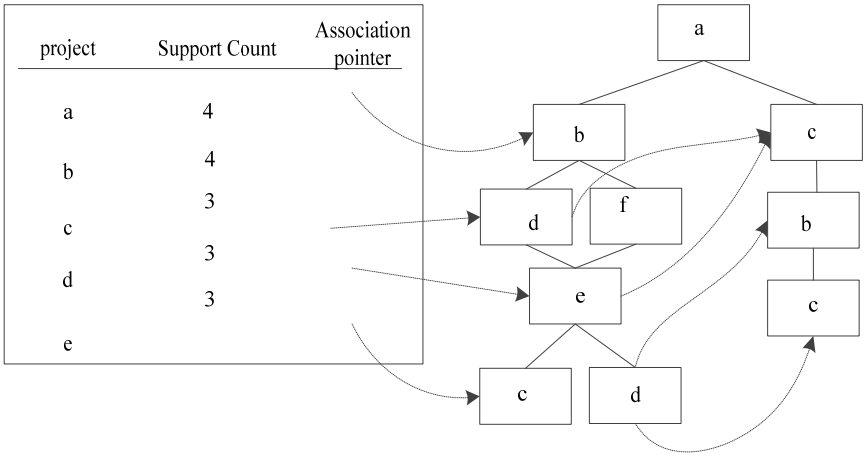
In formula (11), $f(d_i)$ represents the classification results of biomedical literature data, and α represents the data model of expanding medical literature in the vector space.

In the biomedical literature data pre-processing, analyse the data noise level in the biomedical literature dataset, realise the biomedical literature data denoising through wavelet transform, place the denoised literature data in different dimensional feature vector spaces, set the literature data classification benchmark, and achieve the biomedical literature data pre-processing.

3 Implementation of biomedical literature data mining based on association rule algorithm

Based on the biomedical literature data determined in the above chapters, in order to improve the effect of biomedical literature data mining, this paper introduces association rules algorithm to achieve data mining. Association rule algorithm is an algorithm that determines the relationship between large-scale data, finds the association between data items in different relationships, determines the closeness of different data, and then implements data mining. This algorithm has the advantage of processing large-scale data (Xu and Luo, 2021). Therefore, this method is used in this study. The schematic diagram of the relationship between the data mined by the algorithm is shown in Figure 4.

Figure 4 Schematic diagram of association rules mining data relationships



Therefore, this method is adopted for the mining of biomedical literature data. Before mining, prepare the data and use the biomedical literature data in MYSQL data as an example to represent D. The factor graph decomposition global extraction function is introduced to establish a probabilistic database and complete the extraction of biomedical literature data D. Then, wavelet transform is used to denoise the data, and it is classified based on its importance to obtain the final set of biomedical literature data items W, which completes the preparation work before mining. Next, use the proposed method to mine the prepared data. Next, use the proposed method to mine the prepared data, and the key steps for mining are as follows:

Step 1 Determine the form of association rules for biomedical literature data. Set the project set of biomedical literature data, represent all the fields in the data, T represents each item set in the literature data object set, the dataset is in W. The identification code of each medical literature data thing is single, as shown in Table 1.

Table 1 Setting details of identification codes of medical literature data

Object identification code	*Item*	*Object identification code*	*Item*
1	I_1, I_2, I_5, I_6, I_9	6	I_2, I_3, I_6, I_9
2	I_2, I_4, I_8	7	I_1, I_3, I_6, I_7
3	I_2, I_3, I_7, I_9	8	I_1, I_3, I_4, I_6
4	I_1, I_2, I_4	9	I_2, I_3, I_7
5	I_1, I_3, I_4, I_6, I_8	10	I_1, I_5, I_6, I_9

Then the expression form of association rules for biomedical literature data is:

$$K \in T \to Q \in T \tag{12}$$

In formula (12), K / Q represents different arbitrary medical literature data things included in T.

Step 2 Determine the support of biomedical literature data items (Zhang et al., 2020). This value is mainly used to determine the frequency of occurrence of biomedical literature data objects based on the rules set in step 1, that is, its activity range is used to measure the importance or scope of application of association rules, reflecting the universality of rules. The calculation formula is:

$$S(X) = \frac{\left|T(X \cup Y)\right|}{|T|} \tag{13}$$

In formula (13), $S(X)$ represents the support results of the biomedical literature data and things project, and X and Y respectively represent items in biomedical literature data.

However, due to the large number of candidate itemsets generated by the connection of various data item items, the time cost of calculating the support of candidate data item items is high, which reduces the efficiency of mining. Therefore, the method of pre pruning is introduced to calculate the number of occurrences of each item in the set of item items T in the literature data. Before generating the candidate itemset, it is determined that some itemsets are not

frequent. This part of the itemset is pre pruned to reduce the number of candidate data item items and avoid calculating their support frequency, thereby reducing time overhead and improving mining efficiency.

Step 3 Calculate the confidence level of the association rules for biomedical literature data. The reliability of association rules of this data mining is calculated by the determined support results, which is used to measure the accuracy of the association rules. It reflects the probability of the results also being true when the premise of the association rules is true. The calculation formula is:

$$C(X) = \frac{|S(X)|}{|T(X)|} \tag{14}$$

In formula (14), $C(X)$ represents the confidence results of the item correlation rules for biomedical literature data.

Step 4 Based on the above, determine the degree of dependence between the data in the biomedical literature data item. Through the determination of this dependency, the correlation between biomedical literature data is clarified (Zhang et al., 2020), in order to improve the effectiveness of subsequent data mining models and further improve the accuracy of data mining results, and the determination formula is expressed as:

$$U(X, Y) = \frac{u(X \cup Y)}{u(X)u(B)} \tag{15}$$

In formula (15), $U(X, Y)$ represents the degree of dependence between the data in the biomedical literature data project, and u represents the degree of dependence coefficient.

Step 5 Based on the above obtained credibility and the degree of dependence between data in the biomedical literature data item, build a biomedical literature data mining model based on association rules. The final biomedical literature data mining is realised through this model, and the results are as follows:

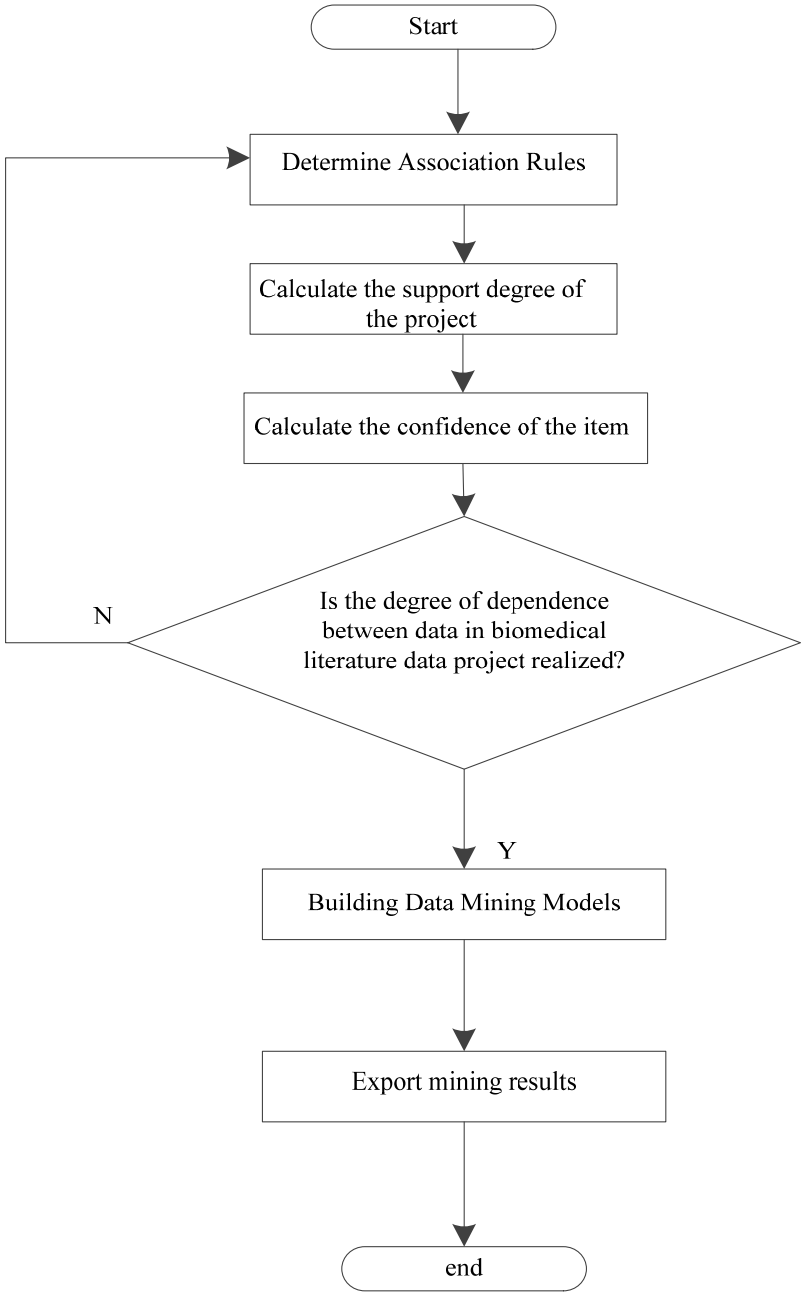
$$\tau^2(X, Y) = \frac{[u(X)u(B) - u(X \cup Y)^2]C(X)}{u(X) \cap u(B)} U(X, Y) \tag{16}$$

In formula (16), $\tau^2(X, Y)$ represents the biomedical literature data mining results.

The implementation process of biomedical literature data association rule mining is shown in Figure 5.

In the biomedical literature data association rule mining, by setting the association rules of biomedical literature data mining, calculate the support and confidence of the literature data item, and by determining the dependency between the data in the item, finally build a biomedical literature data mining model based on the association rules to achieve data mining.

Figure 5 Implementation process of biomedical literature data association rule mining



4 Experimental analysis

4.1 Experimental scheme design

Biomedical literature data in MYSQL data were selected as the experimental object in the experimental test. In the test, 2 GB medical literature data was selected for data mining, and Windows 10 system was used for data mining. The operating memory of this system is 16 GB, which can support the whole process of data mining experiment. The parameters of the specific experimental settings are shown in Table 2.

Table 2 Test parameter setting

Parameter	*Content*
Types of medical literature data mining/species	5
Sample medical literature data/type/article	1,000
Mining result processing software	Spss9.0
Medical literature data noise range/dB	[0, 1]
Collection of items in mining/piece	5
Mining iterations/time	100
Mining data support/%	>90
Data confidence/%	>90

According to the set experimental scheme, the sample medical literature is mined, and Wu et al. (2021) and Wu and Chen (2021) methods are used as comparison methods to compare with the proposed methods. In order to ensure the effectiveness of experimental mining, the experimental environment set by the three methods in the experiment is relatively consistent, and the data obtained have been processed for many times. The experimental indicators tested in the test include the following three types:

1 Biomedical literature data noise: this indicator mainly reflects the data processing before data mining by different methods to ensure the cleanliness of data

2 Precision analysis of biomedical literature data mining. This index reflects the overall effectiveness of different mining algorithms. The closer the value is to 100%, the better the mining effect;

3 Time consuming analysis of biomedical literature data mining: this indicator reflects the speed of data mining and is one of the key indicators.

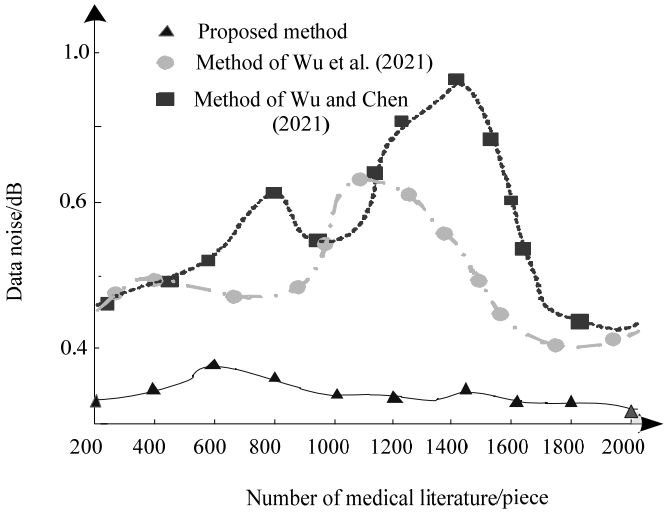
4.2 Analysis of experimental results

4.2.1 Comparison of biomedical literature data noise results

In the experimental test, the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method first tested and analysed the noise reduction results of the sample literature data before conducting the sample biomedical literature data. In the experiment, 2000 pieces of sample medical literature data of different types were selected, and the noise reduction research was carried out for these 2000 pieces of data. The results are shown in Figure 6.

The analysis of six experimental results shows that there are some differences in the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method for noise reduction of sample biomedical literature data. It can be seen from the curve trend in the figure that the noise of the three methods changes to a certain extent with the change of sample literature data. Among them, the noise change curve of Wu et al. (2021) method and Wu and Chen (2021) method is the most prominent, and fluctuates greatly. In contrast, the fluctuation of the proposed method is small, and the noise is controlled below 0.4 dB. It can be seen that the noise reduction effect of the proposed method is better. This is because the proposed method realises biomedical literature data denoising through wavelet transform in data processing, and places the denoised literature data in different dimensional feature vector spaces, thus effectively improving the effect of noise reduction.

Figure 6 Comparison of biomedical literature data noise results



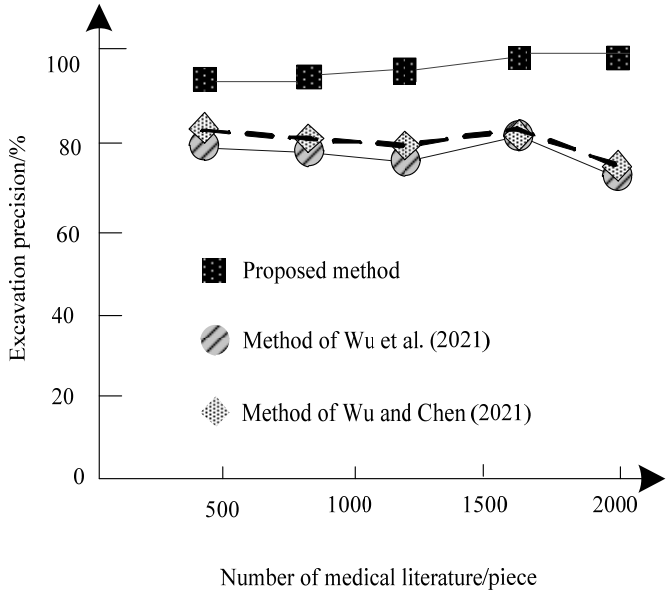
4.2.2 Comparison of precision results of biomedical literature data mining

Next, the experiment tests the mining accuracy of the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method in the sample biomedical literature data mining, and analyses the results in detail. The precision mining of 2,000 selected medical literature sample data is shown in Figure 7:.

By analysing the experimental results in Figure 7, we can see that there are some differences in the accuracy of the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method in data mining of sample biomedical literature. Among them, the proposed method has the highest accuracy of about 99% for sample biomedical literature data mining, while the mining accuracy of the methods in Wu et al. (2021) and [Wu and Chen (2021) is below 82%, which is lower than the accuracy of the proposed method's data mining. This is because the proposed method uses wavelet transform for noise reduction, which effectively improves the effectiveness of the extracted data and improves the accuracy of subsequent data mining. Based on the calculation of confidence, the degree of dependence between data in biomedical literature data items is

determined to improve the effectiveness of subsequent data mining models and further improve the accuracy of data mining results. It can be seen that the mining effect of the proposed method is more significant.

Figure 7 Comparison of precision results of biomedical literature data mining



4.2.3 Comparison of time-consuming results of biomedical literature data mining

At the end of the experiment, the time consuming of the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method in sample biomedical literature data mining is tested. 2000 selected medical literature sample data, and the mining time is shown in Table 3:

Table 3 Comparison of time-consuming results of biomedical literature data mining (s)

Document data volume/piece	*The proposed method*	*Wu et al. (2021) methods*	*Wu and Chen (2021) methods*
500	1.2	1.6	1.9
750	1.3	1.8	2.1
1,000	1.5	1.9	2.3
1,250	1.5	2.3	2.4
1,500	1.6	2.6	2.6
1,750	1.7	2.9	2.9
2,000	1.7	3.0	3.2

Analysing the experimental results in Table 3, it can be seen that the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method have a large difference in the time consumption of sample biomedical literature data mining. When the sample

biomedical literature data volume is 1,000, the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method take 1.5 s, 1.9 s and 2.3 s to mine the sample biomedical literature data respectively. When the sample biomedical literature data volume is 1,500, the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method take 1.6 s, 2.6 s and 2.6 s to mine the sample biomedical literature data respectively. When the sample biomedical literature data volume is 2,000, the proposed method, Wu et al. (2021) method and Wu and Chen (2021) method take 1.7 s, 3.0 s and 3.2 s to mine the sample biomedical literature data, respectively. It can be seen from the comparison of data results that the mining time of the proposed method is shorter, which is due to the classification preprocessing of the extracted biomedical literature data by the proposed method, and before constructing the biomedical literature data mining model based on association rules, the method of pre pruning is introduced to calculate the number of occurrences of each item in the collection T of literature data items. Before generating the candidate itemset, it is determined that certain itemsets are not frequent. This part of the itemset is pruned in advance to reduce the number of candidate data item items, avoid calculating their support frequency, reduce time overhead, improve mining efficiency, and achieve fast mining research.

5 Conclusions

1 In order to solve the problems of low precision and long time consuming in biomedical literature data mining, a biomedical literature data mining method based on association rule algorithm is designed. This method has the characteristics of high mining precision and fast speed, and can effectively implement data mining.

2 The proposed method realises data extraction by setting the process of biomedical literature data extraction and introducing the factor graph decomposition global extraction function; Then, wavelet transform is used to reduce the noise of biomedical literature data, complete the classification of literature data, and achieve data preprocessing; Finally, by setting association rules for biomedical literature data mining and introducing pre pruning methods, the time cost of calculating support is reduced, and mining efficiency is improved. Combining confidence and dependency, a biomedical literature data mining model based on association rules is constructed to achieve mining.

3 The proposed method is verified from three aspects of noise control, mining accuracy and mining time consumption in biomedical literature data. Compared with the methods in Wu et al. (2021) and Wu and Chen (2021), the proposed method has significant data mining effects, which can effectively control noise and make the noise in data fluctuate less. It is controlled below 0.4 dB, with high mining accuracy of about 99%, and short mining time, with the maximum time consumption of 1.7 s. It has high excavation speed.

4 Biomedical literature data mining method based on association rule algorithm can effectively realise efficient and accurate mining of biomedical literature data, and has broad application prospects.

References

Ali, I., Dreij, K., Baker, S. et al. (2021) 'Application of text mining in risk assessment of chemical mixtures: a case study of polycyclic aromatic hydrocarbons (PAHs)', *Environmental Health Perspectives*, Vol. 129, No. 6, pp.6700–6711.

Carrasco, A., Volberg, C., Pedrosa, D.J. et al. (2021) 'Patient safety in palliative and end-of-life care: a text mining approach and systematic review of definitions', *American Journal of Hospice and Palliative Medicine*, Vol. 38, No. 8, pp.1004–1012.

Chen, C.W., Tsai, C.F., Tsai, Y.H. et al. (2020) 'Association rule mining for the ordered placement of traditional Chinese medicine containers: an experimental study', *Medicine*, Vol. 99, No. 18, p.20090.

Cox, J., Mcbeath, D., Harper, C. et al. (2020) 'Co-occurrence of cell lines, basal media and supplementation in the biomedical research literature', *Journal of Data and Information Science*, Vol. 3, No. 24, pp.4781–4792.

Du, X., Xu, H. and Zhu, F. (2021) 'A data mining method for structure design with uncertainty in design variables', *Computers & Structures*, Vol. 244, No. 2, pp.106457.1–106457.13.

Feng, B. and Gao, J. (2022) 'AnthraxKP: a knowledge graph-based, anthrax knowledge portal mined from biomedical literature', *Database*, Vol. 11, No. 3, pp.63–71.

Gu, L., Fei, Z. and Xu, X. (2022) 'Enhancement method of weak Lidar signal based on adaptive variational modal decomposition and wavelet threshold denoising', *Infrared Physics & Technology*, Vol. 120, No. 7, pp.1–7.

Guo, C., Wang, B., Wu, Z. et al. (2020) 'Transformer failure diagnosis using fuzzy association rule mining combined with case-based reasoning', *IET Generation Transmission & Distribution*, Vol. 14, No. 11, pp.14–18.

Hillenmeyer, S., Davis, L.K., Gamazon, E.R. et al. (2021) 'Data and text mining stams: string-assisted module search for genome wide association studies and application to autism', *Bioinformatics*, Vol. 32, No. 14, pp.15–3822, Oxford, England.

Karatzas, E., Baltoumas, F.A., Kasionis, I. et al. (2022) 'Darling: a web application for detecting disease-related biomedical entity associations with literature mining', *Biomolecules*, Vol. 12, No. 4, pp.520–530.

Krawczyk, K., Chelkowski, T., Laydon, D.J. et al. (2021) 'Correction: quantifying online news media coverage of the COVID-19 pandemic: text mining study and resource', *Journal of Medical Internet Research*, Vol. 23, No. 7, p.31544.

Li, C. and Li, W. (2021) 'Automatic classification algorithm for multisearch data association rules in wireless networks', *Wireless Communications and Mobile Computing*, Vol. 12, No. 12, pp.1123–1128.

Li, T., Zhou, Z., Zhang, K. et al. (2021) 'Direct infusion-tandem mass spectrometry combining with data mining strategies enables rapid chemome characterization of medicinal plants: a case study of polygala tenuifolia', *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 20, No. 4, p.114281.

Momeni, Z., Hassanzadeh, E., Abadeh, M.S. et al. (2020) 'A survey on single and multi omics data mining methods in cancer data classification', *Journal of Biomedical Informatics*, Vol. 107, No. 41, pp.103466–103466.

Sokhangoee, Z.F. and Rezapour, A. (2022) 'A novel approach for spam detection based on association rule mining and genetic algorithm', *Computers and Electrical Engineering*, Vol. 7, No. 9, pp.97–97.

Trieu, H.L., Miwa, M. and Ananiadou, S. (2021) 'BioVAE: a pre-trained latent variable language model for biomedical text mining', *Bioinformatics*, Vol. 38, No. 3, pp.872–874.

Wang, T., Xiao, B. and Ma, W. (2022) 'Student behavior data analysis based on association rule mining', *International Journal of Computational Intelligence Systems*, Vol. 15, No. 1, pp.1–9.

Wu, J.M., Srivastava, G., Yun, U. et al. (2021) 'An evolutionary computation-based privacy-preserving data mining model under a multithreshold constraint', *Transactions on Emerging Telecommunications Technologies*, Vol. 32, No. 3, pp.1–19.

Wu, Z. and Chen, Y. (2021) 'Digital art feature association mining based on the machine learning algorithm', *Complexity*, Vol. 2021, No. 1, pp.1–11.

Xu, R. and Luo, F. (2021) 'Risk prediction and early warning for air traffic controllers' unsafe acts using association rule mining and random forest', *Safety Science*, Vol. 135, No. 24, p.105125.

Zhang, H., Chen, J., Qiang, Y. et al. (2020) 'DART: a visual analytics system for understanding dynamic association rule mining', *The Visual Computer*, Vol. 17, No. 1, pp.524–561.