ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

Data Mining of Traditional Chinese Medicine Ingredients

Xiaofei Xu[1], Keping Mao[1], Gang Wang[2]

[1]School of Automation, Beijing Information Science and Technology University, Beijing, 100192, China

[2]School of Life Sciences, Beijing University of Chinese Medicine, Beijing, 100192, China

Abstract

With the development of traditional Chinese medicine, based on data mining, it was discussed the compatibility laws and application characteristics of prescriptions containing different principal components. The data mining results would be applied to provide references for clinical science and standardized use of prescriptions. Taking the prescriptions included in the national population and health science data sharing platform as the research object, through data set construction and data analysis, the digital visualization of prescription drug association analysis is realized, and the classical data mining algorithm is optimized to improve the accuracy and recall rate. By comparison of using IBM SPSS Modeler, IBM SPSS Statistics and other software for analysis and drawing the Association rules, the Apriori algorithm has been researched advantage for prescriptions data mining in weka, and the construction of the man-machine interface design.

Keywords

Association rule analysis; Prescriptions data; Apriori algorithm; Cluster analysis.

1. THE FUNDAMENTAL PRINCIPLE

1.1. Introduction

As the development of electronic digitalization of medical prescriptions data, the research of the inland algorithms of mining association rules are almost achieved by the single minimum support; According to the value of data analysis between traditional Chinese medicine prescriptions, there would be to summarize the rules and make intelligent decisions, and to help improve the diagnostic efficiency[1-10].

The paper studied the classic mining data Apriori algorithm, which overcomes the shortcoming that it needs to scan frequently the database itemset and leads to low mining efficiency. So according to the specific data, through creating data dimension, storing multidimensional cube, processing data explosion and analyze multidimensional cube, which the association rules of prescriptions have been realized the increase of efficiency and simplicity by the improved Apriori algorithm. There are many analysis methods to build datasets in the "statistical report" and "data analysis" modules, frequency analysis, association rules, and complex Systematic entropy clustering[10-16]. The paper has been integrated some methods to form an high-frequency number, classification, nature, taste and meridian analysis, association rule analysis and cluster analysis, and display the results through visualization and human-computer interaction.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

1.2. The mathematical principle

The data mining of prescriptions is to vectorize the association rules between the data sets of the main drug components in several prescriptions and other drug components by analyzing the prescription component database of prescriptions, and to conduct association analysis, cluster analysis[1], classification analysis, anomaly analysis, special group analysis， confidence degree and evolution analysis; All the express from the quantitative calculation results, which the rules can be found in the user understandable way (such as visualization).

For example, in the analysis of prescription data, it can reflect the probability that the patient would purchase drug A or B at the same time. According to the analysis results, it can effectively reflect the patient's medication history or the patient's resistant drugs, so as to help doctors compare when prescribing prescriptions, and improve the efficiency and fault tolerance of doctors[2].

The classical Apriori algorithm for data mining was first proposed by R. Agrawal et al. In 1993, it is used to mine association rules between frequent item sets in customer transaction data. Its core is a recursive algorithm based on the idea of two-stage frequent sets. It is one of the most famous association rule mining algorithms in the world. And the association rules: an implication expression in the form of a=>b, in which a and B are not empty sets, and a or b is empty [3-10]; Among which the several important concepts are: confidence, promotion and support;

(1) support

The paper proposes a new mathematical model for defining the matching support, which can remarkably improve the stability of matching, and that can be understood by shopping in the nearby supermarket. For example, if 100 users buy drugs in the pharmacy, and 80 users have purchased drug a, the support is 80%, that is, the association rules are strong. Here, and again, that's going to be important to point out, drug A, not only a certain drug, but also a set or any specified cluster, and those which would be eliminated under the minimum support level. In the paper, which is designed how to find the most frequently used medicinal material in the compound prescription for a certain kind of disease, that is, the one with the highest support among the medicinal materials, and successively find out several medicinal materials with strong association rules, so as to help doctors quickly lock in the main medicinal materials when prescribing prescriptions for this kind of disease.

(2) the degree of confidence

The degree of confidence refers to the frequency of item set B in transactions containing[11-14]. The expression is: The target adjustment is simple and the degree of confidence and precision are high.

$$C(A \Rightarrow B) = \frac{S(A \Rightarrow B)}{S(A)} \tag{1}$$

The probability theory can be understood as the probability that the person who buys drug a will buy drug B, which can be simply expressed as:

$$P(B \mid A) = \frac{P(AB)}{P(A)} \tag{2}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

The probability of the occurrence of drug A in the traditional Chinese medicine prescription was designed in this paper also has the occurrence of drug B, which can be better help the doctors to select auxiliary drugs and allergic drugs when issuing prescriptions [15-16].

the degree of lifting

The lifting degree refers to the ratio of the frequency of item set B in transactions containing item set A to the frequency of item set B in all transactions. Its expression is:

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{S(B)} = \frac{S(A \Rightarrow B)}{S(A) \times S(B)} \tag{3}$$

The degree to which the occurrence of drug a increases the occurrence probability of drug B, expressed by probability theory, that is:

$$L = \frac{P(AB)/P(A)}{P(B)} \tag{4}$$

If the lifting degree is greater than 1, it is recommended; if it is less than 1, it is not recommended. It can intuitively help doctors to choose various auxiliary drugs, extremely effectively reduce the thinking time when prescribing prescriptions, and also avoid patients misleading doctors to prescribe wrong prescriptions because they are not familiar with their own conditions.

2. DATA SET CONSTRUCTION AND ANALYSIS OF THE SYSTEM

Data source: a total of 988 data sets consisting of 38 incomplete prescriptions and 26 medicinal materials were collected for diseases such as low back pain (national population and health science data sharing platform, http://dbcenter.cintcm.com). The data is sorted as shown in Table.1.

Table 1. Example of some prescription data in the data set used in this paper

Achyranthesbidentata	Kawakami	...	Chain Fern	Epimedium
1	1	...	0	0
1	1	...	0	0
0	1	...	0	0
1	1	...	0	0
1	0	...	1	0
...	...	...	...	...
1	0	...	0	1
1	1	...	0	0

As shown in Table.1, each row in the table represents a group of prescriptions, and each column indicates whether the drug appears in the prescription. Which means that the drug is contained in the prescription, and 0 means that the drug is not contained in the prescription.

Association analysis is one of the most commonly used analysis methods in data mining. Through association mining, hidden connections between item sets can be found from massive

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

data [10]. Association rule analysis obtained 67 combinations of traditional Chinese medicine. In the association rules of drug pairs, "Peony-liujinu", "Peony-Xianlingpi" and "Peony-Patriniavillosa" have high support.

Here, the Apriori algorithm was used to generate association rules, which is divided into two steps: Step 1: generate frequent item sets; Step2: generate association rules; Rules: association rules mined by Apriori algorithm, the output show in Fig.1.

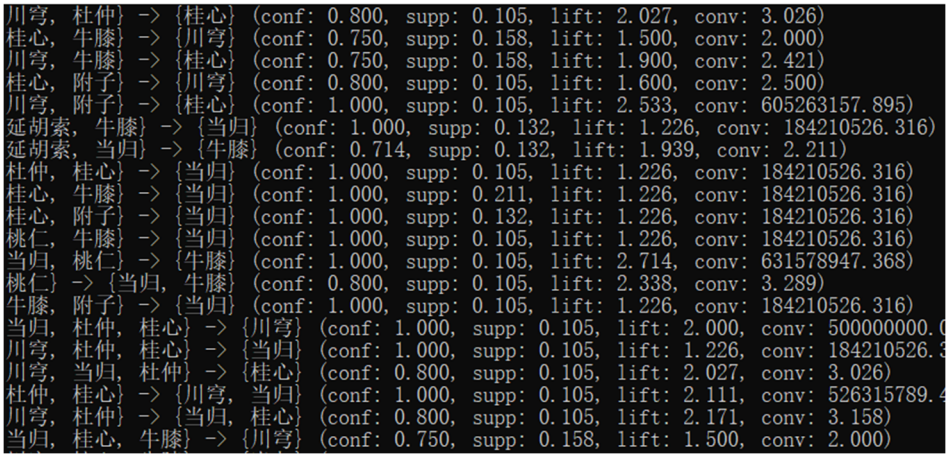


Figure 1. The output of the data Association analysis

Here, in Fig.1 which "conf" stand for the degree of confidence "soup" stand for support degree, "lift" stand for lift degree，"conv" stand for vector convolution in algorithm. By converting the results generated in Fig.1 into a data table, you can clearly see the association rules between drugs and the quantitative representation of strong association rules, as shown in Table.2

Table 2. Association rules

Association rules	souport degree	Confidence degree	Lift degree
Cnidiumofficinale Makino→ Chinese angelica	0.474	0.947	1.161
Eucommiaulmoides→ Chinese angelica	0.158	0.857	1.051
Corydalis yanhusuo→ Chinese angelica	0.184	0.875	1.073
Guixin→ Chinese angelica	0.368	0.933	1.144
Peach kernel → Chinese angelica	0.105	0.800	0.981
no →	0.105	0.800	0.981
Rehmanniaglutinosa→ Chinese angelica	0.158	0.857	1.051
Achyranthesbidentata→ Chinese angelica	0.342	0.929	1.138
monkshood→ Chinese angelica	0.211	0.889	1.090

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

3. THE IMPROVEMENT OF THE APRIORI ALGORITHM

The design of traditional Chinese medicine data mining system is shown in Figure 4.1. This module is designed for traditional Chinese medicine association rule mining. Therefore, this paper focuses on association rule mining. In this study, frequency statistical analysis, association rule analysis, cluster analysis and other methods were used to explore the clinical application and compatibility of Gastrodiaelata containing prescriptions. Weka was used to call Apriori algorithm (weka.association.apriori) to overcome the low efficiency of multiple scanning of data sets.

The experimental comparison results between Apriori and other benchmark models are shown in Table. 3: the accuracy and recall reach 91% and 93% respectively.

Table 3. Comparison of experimental results between Apriori and other methods

	Amazon-Book		Last-FM		Yelp2018	
	recall@20	ndcg@20	recall@20	ndcg@20	recall@20	ndcg@20
GAT	0.1453	0.0997	0.0912	0.1325	0.0702	0.0826*
GATsamle	0.1434	0.0942	0.0920	0.1334	0.0705	0.0825
GCN	0.1423	0.0924	0.0865	0.1296	0.0694	0.0787
RGCN	0.1487*	0.0984	0.0887	0.1345	0.0710	0.0821
FMRGCN	0.1475	0.0981	0.0896	0.1344	0.0704	0.0822
HAN	-	-	0.0914	0.1357	-	-
Priori+B	0.1541	0.1073	0.0936	0.1423	0.0734	0.0901
Apriori+C	0.1546	0.1089	0.0930	0.1404	0.0742	0.0896
Improve	+3.97%	+9.23%	+2.41%	+4.86%	+4.51%	+9.08%

It can be seen that Apriori outperforms other benchmark models in three data sets. Apriori+B represents the training results on Graph B data. It can be seen that the impact of Graph B and Graph C on Apriori, which is different on different data sets. This may be determined by the different emphasis of data on CF side and kg side; It can be found from Table.3 that the data of last FM on the CF side is much more than that on the kg side, and the edge types of the data on the kg side are also relatively few, indicating that the unidirectional structure is more suitable for this kind of chart structure data. The Han model can only be run on the second dataset because it is hierarchical by the type of edge, which will be run on other datasets.

The comparison between Apriori algorithm and GAT model is shown in Fig.2.

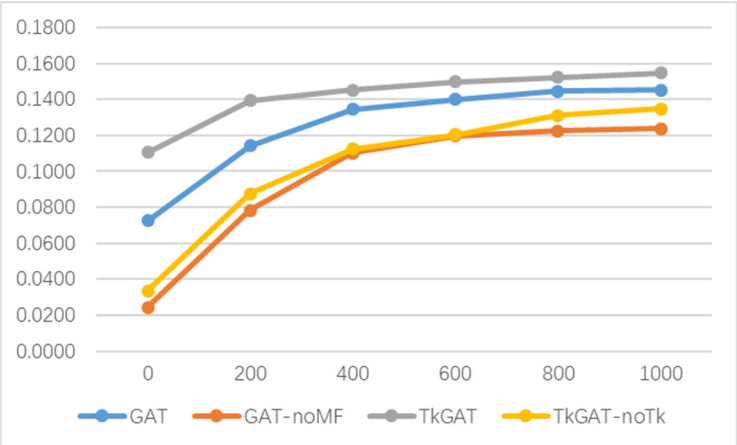


Figure 2. The comparison between Apriori algorithm and GAT model

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

It can be seen that whether Apriori or GAT models, initialization using matrix decomposition can greatly shorten the training time and improve the final effect, which fully illustrates the effectiveness and efficiency of initialization. An appropriate initialization result can greatly improve the model and save a lot of training time. At the same time, it can be noted that the supplement of KG side data is also very helpful for the model. However, it should be explained here that Apriori uses the results of Tucker decomposition during initialization. Therefore, even Apriori-noKG actually uses additional kg information, but it does not consider the entity nodes on the kg side when aggregating node information in the network, so it is not surprising that its performance is better than the other two models.

Next, when comparing the performance of Apriori with other benchmark models, as shown in Table.4: part of the data in the above table comes from the paper KGAT, where KGAT-3 represents the KGAT model using three-layer network structure, and KGAT-4 represents the KGAT model using four-layer network structure. We can see that Apriori has achieved the best results on the three data sets, as shown in Table.4.

Table 4. Comparison of experimental results between Apriori and other methods

	Amazon-Book		Last-FM		Yelp2018	
	recall@20	ndcg@20	recall@20	ndcg@20	recall@20	ndcg@20
FM	0.1345	0.0886	0.0778	0.1181	0.0627	0.0768
NFM	0.1366	0.0913	0.0829	0.1214	0.0660	0.0810
GCMC	0.1316	0.0874	0.0818	0.1245	0.0659	0.0790
RippleNet	0.1336	0.0910	0.0791	0.1238	0.0664	0.0822
KGAT-3	0.1489	0.1006	0.0870	0.1325	0.0712	0.0867
KGAT-4	0.1503*	0.1015*	0.0871*	0.1329*	0.0722*	0.0871*
Apriori+C	0.1546	0.1089	0.0930	0.1404	0.0742	0.0896
Improve	+2.86%	+7.29%	+6.77%	+5.64%	+2.77%	+2.79%

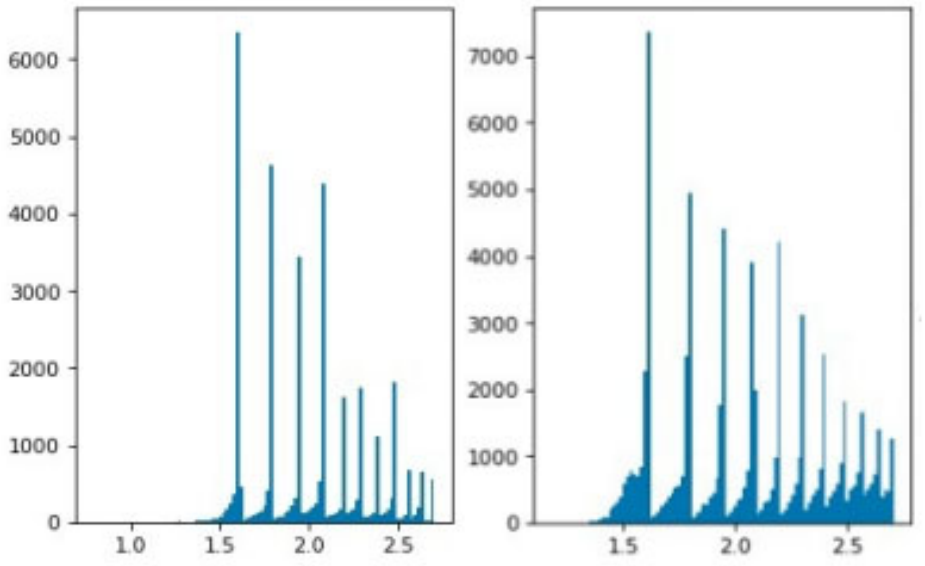
It can be seen different with which the experimental results that the effect of sampling on different data sets. By analyzing the differences of the three data sets, it can be found that since this paper samples the data on the CF side, the data set with more data on the CF side can obviously obtain better performance in sampling; For data sets with much less data on the CF side than on the KG side, sampling may damage performance to some extent. Of course, since the focus of this paper is on the structure of the model and the use of Tucker's results to enhance the performance of the model, the proposed sampling strategy is too simple, and subsequent research can continue to dig around this point.

Finally, to analyze the effectiveness of Apriori, we can choose to use the weight entropy around the computing nodes to measure an attention model. For any node i, the entropy of its attention distribution can be calculated by the following formula:

$$H(\alpha_i) = -\sum_{j \in N_i} \alpha_{ij} \log \alpha_{ij} \tag{5}$$

For the entropy of attention distribution of the trained Apriori and GAT models, select the node degree 5∼15. As shown in Fig.3, the left is GAT and the right is Apriori:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002



(a)GAT (b) Apriori

Figure 3. Comparison of attention entropy

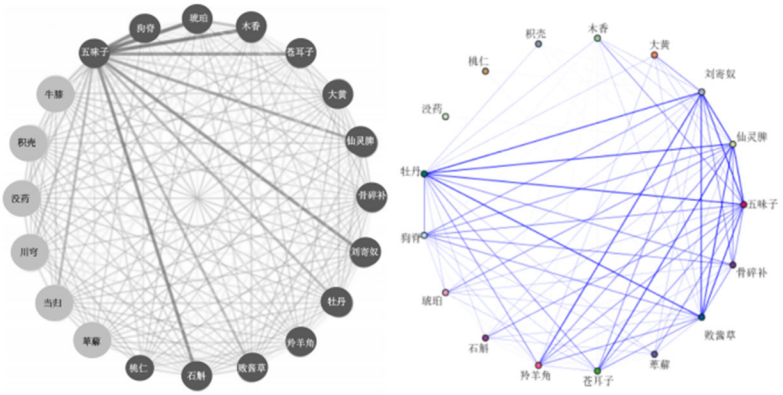
In Fig.3, the abscissa represents entropy, and the ordinate represents the number of entropy nodes. It can be seen that the entropy distribution of Apriori is more uniform, while the entropy presented by GAT model is more sharp, indicating that Apriori's attention mechanism is more effective than the ordinary GAT model.

For the integrated output of the module, the interaction is achieved between the page presentation layer and the module application layer. However, the module design needs to be jointly debugged with the software used in the module. It needs to negotiate with the software development company to obtain the application source code or data interface. In the next step, the data analysis will be realized with the help of computing tools. In the future, computing tools will be used to help intelligent decision-making, to implement a tool that can track patient information and provide feedback quickly. Apriori algorithm needs to scan the database many times and link frequent item sets many times to generate a large number of frequent item sets, which leads to the disadvantage of low mining efficiency. For the study of medical drug data, Apriori algorithm finds an appropriate number of datasets, researches and researches, summarizes the attributes of datasets, and discusses data processing methods, Setting of minimum support and minimum confidence.

4. THE VISUAL GRAPHICS OF THE DATA MINING

The association rules of data of Table.2 had been analyzed in the Apriori algorithm of IBM SPSS modeler 18.0, Set the front item support threshold ≥ 0.1, confidence threshold ≥ 0.8, and the maximum number of front items was 5. Based on the drug relationship co-occurrence diagram output by Cytoscape, 41 groups of confidence data between drugs are selected as the corresponding relationship to draw the drug relationship co-occurrence diagram. As shown in the Fig.4, The comparison output of visual graphics of Cytoscape and SPSS Modeler:

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

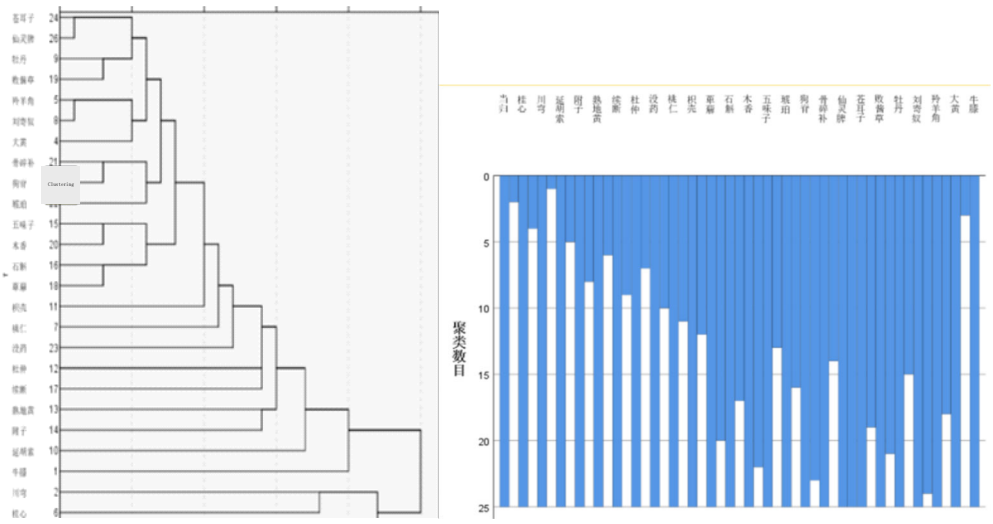


(a) byCytoscape (b) by SPSS Modeler

Figure 4. The comparison of visual graphics of Cytoscape and SPSS Modeler

As shown in Fig.4, the knotted dots represent the herbs, the thickness of the line indicates the strength of the association rules, the circular node represents the Chinese medicine, and the thickness of the connection represents the strength of the correlation. In the correlation diagram generated by SPSS Modeler, the correlation line between the non-eliminated medicinal materials with weak correlation is almost invisible, and it is precisely for this reason that the insufficient thinking has been considered, and Cytoscape can draw the co-occurrence relationship of all drugs according to specific data, but because the platform needs to manually change the attribute arrangement of the picture, it is not particularly in line with the goal of this article, the time relationship.

The proposed design of this paper has used SPSS Modeler and Cytoscape to show the co-occurrence relationship between drugs from the local and the whole, respectively, but it is difficult to reflect the contrast or similarity between drugs; Using IBM SPSS Statistics 23.0 software for systematic cluster analysis, the similarity between variables is measured by pearson correlation coefficient, the genealogical map can intuitively represent the similarity between drugs, and the icicle chart can make the drug have a sharper contrast. After adding variables and inserting the corresponding data, select System Cluster analysis and adjust the parameters as required to get the results, where the spectral plot and vertical icicle plot are shown in Fig.5.



(a) Pedigree (b) Icicle

Figure 5. Pedigree & Icicle

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

The icicle chart is the default visual graph output of the system, and the genealogical graph is the visual output of artificial selection.

(a) The word cloud by PyQt5 (b) human-computer interaction interface

Figure 6. The simplified Visual output of interacting with a computer

The paper uses multiple platforms to realize the purpose of association rule mining and data analysis, and uses Python language to implement Apriori algorithm to process fusion intelligent data; In order to be better analyzed and designed the system, the main modules of the system, including the cluster analysis module of IBM SPSS statistics platform, The Cytoscape platform draws the drug co-occurrence diagram, the python language implements the Apriori algorithm, and the python language is used for the visual interface design. For example, the word cloud and human-computer interaction interface are designed for the generation of simple formula data, as shown in Fig.6. The data mining analysis results and visual presentation of formulas are realized, and the data and pages are processed interactively.

5. CONCLUSION

To sum up, this study systematically analyzed and compared the cluster analysis module of IBM SPSS statistics platform, and the drug co-occurrence diagram was drawn on the Cytoscape platform; Finally, we choose to use Weka to call the optimized Apriori algorithm, which can overcome the shortcomings of low efficiency of multiple scans of data sets. The experimental comparison results between Apriori and other benchmark models show that the accuracy and recall are 91% and 93% respectively. The Apriori algorithm implemented by Python language can be used for data fusion processing and data mining of prescription, including the clinical application, drug composition, sexual and flavor tropism, high-frequency traditional Chinese medicine and its association rules of the prescription containing the main components. The compatibility characteristics and application rules of the prescription containing the specific main components are summarized, and the main diseases containing the Gastrodiaelata prescription are discussed, and the clinical application direction ofGastrodiaelata is defined, which provides scientific and the standard use of Gastrodiaelata reference.

REFERENCES

[1] Ye Xiaobin. Based on data mining and analysis, the drug composition rules of spleen and stomach disease in "Tai Hospital Secret Cream Dan Pill Loose Formula"[J].Jiangxi Journal of Traditional Chinese Medicine,2022,53(06):32-37.Doi:10.3969/j.issn.0411-9584.2022.6.jxzyy202206013

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

[2] XieYuling, LiDan, LiuYufeng, WenJianying, ZhangYuqin, XuWei, TangJiangshan. Analysis of medication rules of Director Tang Jiangshan in the treatment of chronic atrophic gastritis based on data mining[J].Straits Pharmacy,2022,34(04):136-139.DOI:10.3969/j.issn.1006-3765.2022.04.039

[3] Shi Xiaoqian,Niu Yang. Research on the Application Law of Aromatic Medicine in ming and Qing Dynasty Temperature Disease Prevention and Atypical Transmission Syndrome Based on Data Mining[J].Chinese Journal of Basic Medicine of Traditional Chinese Medicine,2022,28(03):451-457.DOI:10.19945/j.cnki.issn.1006-3250.2022.03.003.

[4] WeiZhili, Zhang Xiaobo, GaoRui, JianShengnan, Ma Xiaoju, Wen Yueqiang, Zhou Xin, Shen Tao. Data Mining of Ume-containing Prescriptions and Network Pharmacological Analysis of Ume-Ginseng for the Treatment of Diabetes Mellitus[J].Clinical Journal of Traditional Chinese Medicine,2022,34(03):505-511.DOI:10.16448/j.cjtcm.2022.0322.

[5] JiHongyun,YuZhengke. Research on the medication rules for the treatment of palpitations in the Dictionary of Traditional Chinese Medicine Prescriptions based on data mining[J].Hunan Journal of Traditional Chinese Medicine,2022,38(02):19-22+54.DOI:10.16808/j.cnki.issn1003-7705.2022.02.005.

[6] Yang Luyan, NieFumin, Jin Weijie, Yang Xiaodong, Li Benfa, Huang Donghan, Huang Min. Discussion on the rules of medication for the treatment of cradle carbuncle in the "Dictionary of Traditional Chinese Medicine Prescriptions"[J]. Yunnan Journal of Traditional Chinese Medicine, 2022, 43 (02): 24-28.DOI:10.16254/j.cnki.53-1120/r.2022.02.023.

[7] Zhang Suzhi, YuYiran, ShenMinzhe, LiHongwei. Application of Association Rule Algorithm in Traditional Chinese Medicine Prescription[J]. Electronic Technology, 2022, 51(02): 74-78.

[8] LanShaohang, LiNana, ShiChaojia, HuangBo, ChenJianping, Fang Gang. Analysis of formula rules of adversum-astragalus containing chicken blood vine-astragalus drug based on data mining[J].Herald of Traditional Chinese Medicine,2022,28(01):152-155.DOI:10.13862/j.cnki.cn43-1446/r.2022.01.022.

[9] Song Zhe,HuangZhiyan,FengYu,ZhangBaoqing,LuoGuangzhi. Based on data mining, the formula rules of the prescription of calamus-yuanzhi medicines containing calamus-yuanzhi medicine[J]. Chinese Journal of Chinese Materia Medica, 2022, 47(06): 1687-1693.DOI:10.19540/j.cnki.cjcmm.20211203.502.

[10] Zhang Chengdan, LiMengqi, DengYang, YangJin, OuyangXiaoyong. Based on data mining, the pharmacological rules of traditional Chinese medicine prescription dictionary for the treatment of sores and ulcers [J]. Chinese Ethnic Folk Medicine,2021,30(22):16-20+46.Doi:10.3969/j.issn.1007-8517.2021.22.zgmzmjyyzz202122006.

[11] Wu Chunxing, PeiZhifei, BaiQingyun, WangBolong. Analysis of compatibility rules and application characteristics of tianma-containing prescriptions based on data mining[J].New Drugs and Clinical Pharmacology of Traditional Chinese Medicines, 2021, 32 (10): 1562-1567.DOI:10.19378/j.issn.1003-9783.2021.10.022.

[12] Wang Jiajun, ChenQingyao, WangJian, YuanJianmei. Based on data mining and network pharmacology, the compatibility law and mechanism of action of huanglian-containing formula in the treatment of ulcerative colitis [J]. Chinese Herbal Medicine, 2021, 52 (19): 5984-5995.Doi:10.7501/j.issn.0253-2670.2021.19.021.

[13] WengJiajun, XieYilin, Zhang Xuanshuo, Gao Cui, Cui Cang, Zhao Jiaxiong, BaiXufeng, Zhu Yanchen, Hu Huiming, LvGuiyuan. Analysis of forsythia-containing formulas and their anti-inflammatory mechanisms in the Dictionary of Traditional Chinese Medicine Prescriptions based on data mining

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202302_9(2).0002

and network pharmacology[J].Chinese Journal of Experimental Formulary, 2021, 27 (22):181-193.DOI:10.13422/j.cnki.syfjx.20211319.

[14] Zhang Yingxin,GuiMeng,SuWenyu,DouMingjie,Lu Dan. Based on data mining, a study on the composition law of formulas containing Bayantian formula[J].Special Products Research,2021,43(04):60-64+69.DOI:10.16720/j.cnki.tcyj.2021.094.

[15] Zhang Yun. Based on literature data mining, the dialectical classification of primary osteoporosis and the rules of drug use[D].Shandong University of Traditional Chinese Medicine, 2021.DOI:10.27282/d.cnki.gsdzu.2021.000022.

[16] Villavicencio Charlyn Nayve; Macrohon Julio Jerison Escudero; Inbaraj Xavier Alphonse; Jeng JyhHorng; Hsieh JerGuang.COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA [J]. Algorithms, 2021, 14 (7): 201.Doi:10.3390/A14070201.