Study on Preprocessing Method of TCM Prescription Data in Data Mining

S Z Zhang¹, Y R Yu² and M Z Shen²

¹ School of Software, Zhengzhou University of Light Industry, Zhengzhou, Henan, China
² School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, China
Email: zhsuzhi@zzuli.edu.cn; 758828804@qq.com

Abstract. Traditional Chinese medicine (TCM) prescriptions have been developed for thousands of years. Data forms are diverse, content is discrete and missing, and there are many uncertainties due to cultural and regional differences. Therefore, it has brought some difficulties to the mining of TCM prescriptions. Data based on the 3108 prescriptions for the treatment of typhoid fever, for example, is given priority to with data cleaning and data transformation of data preprocessing, prescriptions combined with multiple functions, expounds the unqualified prescriptions data cleansing, drug name normalization, dose for solving the problems of the unification, the data structured method, make the processed data can be effectively mining, It provides a strong support for exploring the compatibility law of prescription and the development of new drugs.

Keywords. Prescription; data preprocessing; data cleaning; data transformation.

1. Introduction

After thousands of years of inheritance and development, TCM has become one of the quintessences of Chinese culture, and the state attaches great importance to the informatization of TCM. As a new information processing technology, data mining has been widely used in TCM diagnosis, TCM property research, prescription compatibility law and other aspects, bringing vitality to the development of TCM [1]. As an important part of traditional Chinese medicine culture, prescriptions contain several the change and development of the concerted application rule and the relationship between diseases and other aspects of TCM drugs for thousands of years. And the knowledge contained in it provides a strong support for clinical medicine, concerted application rule, new drug development and so on. At present, many scholars have used data mining technology to mine and analyze TCM prescriptions. Bai et al. [2] proposed that data mining could be used in the aspects of database design, prescription compatibility, prescription efficacy and so on in prescription science, providing an effective method for the research of data information in prescription science. Li [3] mentioned that at present, data mining technology is usually used in prescription compatibility rules in three modes: classification, clustering and association rules, and a lot of achievements have been made in the research. However, due to the development of TCM up to now, it has been studied and developed by numerous schools and doctors, and its data sets are huge and lack of systematic collation. As a result, it has diversified forms, discrete data, incomplete content, and incorrect, redundant, invalid and inconsistent noise data [4], which affect the mining efficiency and accuracy.

According to relevant literature, data preprocessing takes up 60% of the whole mining work and is an important part of the data mining process [5, 6]. Data preprocessing is to process the dirty data
existing in the original prescription data set, such as incomplete, inconsistent and unable to be directly mined, and finally obtain the prescription data set of higher quality suitable for mining [7]. The preprocessing of prescription data is to convert medical information into digital information and process the redundant, blank and repeated information [8]. This paper takes the data of 3108 prescriptions for treating typhoid fever as an example. It mainly deals with the data of redundant, incomplete and special, drug standardization, dose unification and other problems, combined with a series of functions and methods of data cleaning and transformation.

2. Characteristics of TCM Prescriptions
Prescription of TCM is a combination of several drugs based on the principle of compatibility and clinical experience. It is the embodiment of treatment. It is generally composed of four parts: monarch medicine, minister medicine, adjuvant medicine and envoy medicine, which follows the four basic principles of “monarch and minister adjuvant envoy”. Prescription contains not only the law of drug compatibility for a certain disease, but also contains the diagnostic experience accumulated so far, which is of valuable research and use value. The characteristics of TCM prescriptions are mainly as follows:

2.1. Large Amount of Data and Redundancy
The development and inheritance of TCM prescription data has a history of thousands of years. Each prescription contains not only the drug information, but also the patient's self-reported condition information, the disease and treatment information determined by the doctor according to the patient's self-reported condition. Therefore, TCM prescription data has a large amount of data. And the characteristics of the drugs and their corresponding symptoms are fixed, so there are inevitably some prescriptions are the same, which leads to data redundancy.

2.2. Data Is Fuzzy and Incomplete
TCM diagnosis is to be combined with the “inspection, auscultation and olfaction, inquiry, and pulse-taking and palpation” four aspects: information, on the basis of on the basis of traditional Chinese medicine, also a lot of into the healer’s own subjective experience, therefore in the different formula, the terms of the healer with the symptoms of patients describe different, different form, cause the symptom of patients with fuzzy data. In addition, the phenomenon of the same name exists in TCM, and the names of drugs used by different doctors are also different, which leads to the vagueness of drug names. Traditional Chinese medicine prescriptions are recorded in writing, but after long-term storage, the paper will be damaged and corroded, so that the data will be lost, resulting in a certain amount of incomplete data.

2.3. Data Is Not Standardized and Unstructured
In addition to the same name of drugs, multiple names for one disease, and one name for multiple diseases, TCM prescriptions also have various forms of drug dosage, such as grams, kilograms, strains, coins, etc. And there is no unified and standard drug name, symptom description and drug dosage, so it is not standardized to a certain extent. The data of prescription usually use long text fields and other unstructured forms to describe drug composition, disease, usage and dosage, source of prescription and other information. In mining, it is necessary to separate drug name, drug dosage and disease from the long fields and standardize them, so as to obtain standardized, concise and complete data.

According to the above-mentioned data characteristics of TCM prescriptions, the difficulties brought to data mining are as follows:

1. How to clear the data that is not suitable for mining or does not need mining in the huge data set?
2. How to deal with the problem of “same and different names” of drugs?
3. How to unify drug dosage and conduct “drug name-dose” separation?
4. How to conduct structural processing of data?
3. Preprocessing of TCM Prescription Data
In order to solve the problems existing in the above-mentioned TCM prescription data and ensure the quality of the data set and the knowledge obtained from data mining, the following four processing methods are proposed for the data set used in this paper: (1) Data mining; (2) Dose normalization; (3) Data Structuring; (4) Drug name Normalization. The flow chart is shown in figure 1.

![Flow chart of TCM prescription data preprocessing.](image)

4. Examples of TCM Prescription Data Preprocessing
In this paper, the data of 3108 prescriptions for the treatment of typhoid fever collected from the Chinese Medicine Network and the Yao Zhi Network were taken as an example to conduct preprocessing for the purpose of mining their compatibility rules.

4.1. Data Cleaning
Data cleaning is one of the important parts in the process of data preprocessing. The so-called cleaning is to remove the redundant, invalid and inconsistent noise data existing in the original data set, including the data that need not be mined in the experiment.

The text cleaned the 3108 prescriptions for typhoid fever, mainly to remove the repeated, blank, special and unexcavated data in the original data set. The special data refers to some prescriptions for pregnant “women”, “children”, “babies” and other special groups, which deviated greatly from those used by ordinary people and need to be discarded. It is proposed to use Python language and PyCharm programming software to clean the data. First, import “numpy”, “pandas” and “re” modules for later use. Import “xlwt” and “xlrd” module to read the formula data stored in the Excel sheet. Next, use the drop statement and the “duplicates()” function to view and delete duplicate statements, and use the if statement, drop statement, and “split()” function to determine if there are blank lines and delete them. Search out the special prescriptions including “woman”, “pregnancy”, “child” and “baby” in the “primary treatment” attribute of the data by querying characters and delete them, and finally store the cleaned data in a new Excel table. The process is shown in figure 2.

In prescription data, the processing method of drugs is often marked in brackets, and this data does not need to be mined. In this paper, the data in brackets is cleared by Excel. After data cleaning, there were 2537 prescriptions. As the correlation between drugs was only studied in the experiment, only drug data was retained. The comparison before and after cleaning is shown in figures 3 and 4.
4.2. Dose Normalization

Traditional Chinese medicine prescriptions do not have systematic measurement data and units, so different doctors prescribe prescription dosage forms are also myriad, which brings great difficulties for data mining and analysis. Drug dose normalization means to convert dose data types and units of measurement to make them unified, and at the same time to make them numerical, which is convenient for data mining. In this paper, the pharmacometrics of 2537 typhoid prescriptions after cleaning were standardized.

4.2.1. Type Normalization. As mentioned in [9], the dosage of TCM prescriptions can be classified into the following three types: (1) Discrete. Such as 3 grams, 2 coins and so on; (2) Continuous. Such as 20.8%—30.6%, 2-3 grams and so on; (3) Fuzzy. Such as a little, amount, etc. According to research, In the prescription, the dose is mostly discrete data, so it is necessary to discretize the fuzzy data, and take the intermediate value as the final value. The calculation formula is shown below:
5

\[ Dose = \begin{cases} \text{Discrete type value} \\
\frac{\text{min} + \text{max}}{2} ; & \text{Continuous type value} 
\end{cases} \]

The drugs with fuzzy dose were mostly tonic drugs, not primary drugs, so the deletion did not affect the treatment, nor did it affect the excavation of prescription.

4.2.2. Standardization of Units. As mentioned in [10], the dosage description of drugs in prescriptions can be roughly divided into the following 5 types: (1) Unit of weight (kilogram, catty, etc.); (2) Unit of capacity (milliliter, liter, etc.); (3) Unit of length (foot, etc.); (4) Text description (amount, little, etc.); (5) Uncommonly used units (cents, coins, etc.). According to the conversion rate between units, different units of measurement are uniformly converted into metric grams(g). At present, most of the quantity values of TCM prescriptions in the database have been changed into Arabic numbers, but a few of them are still in Chinese numbers, which need to be digitized and unified into Arabic numbers.

The drug doses after normalization are all appropriate discrete values, but they are still continuous numerical variables due to the different doses among different drugs. Therefore, the variable form of “name-dose” is one-sided to some extent in prescription data mining, which is not conducive to mining results. The dose is usually not considered in the regular mining of prescriptions, but the dose should be further processed in the classification mining. Zhu et al. [9] proposed the use of K-means clustering algorithm and triangular membership function for dose fuzzy processing. The drug dose was divided into three fuzzy regions: “low (L), medium (M) and high (H)”. The central points of the three regions, C1, C2 and C3, were obtained by K-means clustering method, and the drug dose was divided into corresponding regions by triangular membership function. The function is shown below:

\[ f(x) = \begin{cases} 
\frac{1}{2} \cos \left( \frac{\pi}{14.76} (x - 5.71) \right) + \frac{1}{2} L \quad 0 < x \leq 5.71 \\
\frac{1}{2} \cos \left( \frac{\pi}{37.94} (x - 20.47) \right) + \frac{1}{2} M \quad 5.71 < x \leq 20.47 \\
\frac{1}{2} \cos \left( \frac{\pi}{13.09} (x - 39.44) \right) + \frac{1}{2} H \quad 20.47 < x \leq 58.41 \\
\frac{1}{2} \sin \left( \frac{\pi}{37.94} (x - 39.44) \right) + \frac{1}{2} M \quad 58.41 < x \leq 100 
\end{cases} \]

where C1=5.71, C2=20.47, C3=58.41, [0,100] are the relative regions where the drug dose is located.

4.3. Data Structuring

Structuration is to organize the accumulated data and make it orderly and programmatic. In this paper, the prescription data stored in the form of long field is decomposed into columns to separate various drugs. In this paper, the “Data” TAB in Excel is used to perform column decomposition of prescription drug data with the stop symbol “and” as the segmentation symbol. When mining the correlation between TCM prescriptions and drugs, the dosage of each drug in different prescriptions is different, that is, the amount of the same drug is different. Therefore, this is not a data item, and it is difficult to find the correlation in mining. For this purpose, only the drug name is retained in such experiments. The following two methods of dose deletion are described.

4.3.1. Orthogonal Decomposition Method. In this study, the orthogonal decomposition method means to separate the prescription data after cleaning by “drug name-dose”, and only the drug name data can be extracted when studying the correlation between drugs. In addition, this method can also be used to store the separated drug name and dose separately for corresponding specification before drug name and dose specification. The orthogonal decomposition operation is shown below:
5) if(pf has numeric characters) {
6) Medicine = The string before the numeric character in pf
7) Quantity = String that begins with a numeric character in pf
8) i++
9) }
10) while (++p !== null)

The orthogonal decomposition method does not need to unify the unit of measurement of drug dose to delete drug dose data, but only needs to discretize the values and unify them into Arabic numbers.

4.3.2. VBA Creates Function Deletion Method. The dose data in the prescription can be deleted by creating the DELE function through VBA. However, this method has a higher requirement for the standardization of dose data, and it needs to unify the data form and measurement unit of drug dose, and the data form is all discrete Arabic numerals. Create the DELE function is shown below [11]:

Sub dele()
Dim I As Integer
Dim j As Integer
Dim n As Integer
Dim deles As String
For i = 1 To 2537
For j = 1 To 500
For n = 1 To Len(Cells(i, j))
If (Asc(Mid(Cells(i, j), n, 1)) >= 48 And Asc(Mid(Cells(i, j), n, 1)) <= 122) Then
Else
deles = deles & Mid(Cells(i, j), n, 1)
End If
Next
Cells(i, j) = deles
Deles = “”
Next
Next
End Sub

In this paper, VBA creation function method was adopted to delete dose values in prescription drugs. Structured data are shown in figure 5.

Figure 5. Data structure.
4.4. Drug Name Normalization

There is more than 1000~1200 kinds of Chinese medicinal materials, among which about 500~600 is commonly used [9]. Due to the reasons of times and regions, the same medicine has different names, such as: “Artemisia argyi”, also known as “medical grass”, “mugwort”, “moxibustion grass”, etc. TCM names have the phenomenon of the same name and the same name of foreign bodies, the above is the phenomenon of the same name; A foreign body of the same name is a drug with a different name, such as: Cinnamon, also known as cassia, but cassia and cinnamon are not a kind of plant, cassia is warm, cinnamon is hot. In addition, there are many descriptive words starting or ending with “raw”, “cooked”, “fried” and “roast” in TCM, such as “Glycyrrhiza” and “RADIX GLYCYRRHIZAE PREPARATA”, which need to be deleted in the study.

In this study, the drug names and aliases in the Chinese Pharmacopoeia were taken as the criterion to normalize the treatment of the drugs in the prescriptions. Firstly, structured drug data are stored in set D, and TCM set M (M is empty) is established. Then, the “same and different names” table T of drugs, the list of descriptive prefix word W1 and descriptive suffixes W2 were created based on the Pharmacopoeia of Traditional Chinese Medicine. The specific process is shown below:

Start
  For each medicine in D do
    if Di. The prefix word in W1
      Di=Di. Substring(1);
    if Di. The suffix words in W2
      Di=Di. Substring(0,Di.Length-2);
    if Di in T. Another name and T.Official name in D
      M.add(T.Official name);
    Else
      M.add(Di);
  return M
End

In this algorithm, if some drugs with descriptive words can be retrieved in the table T, there is no need to do any processing; if not, the title will be deleted and the processing of the same and different names will be performed. After this process, most of the drug names are standardized, ensuring the integrity of most of the data and simplifying the data samples.

5. Conclusion

With the advent of the digital age, how to accelerate the digitalization and industrialization of TCM has become a major focus in the field of medicine. This paper describes some methods of TCM prescription data preprocessing in data mining, and verifies them through examples, which lays a foundation for the mining of drug correlation in TCM prescription in the future. However, due to the existence of a variety of properties of traditional Chinese medicine and the fact that most of the data are stored in the form of long text, it is still necessary to conduct in-depth research to obtain the ideal prescription data through data preprocessing.

References
[1] Xu J W, Xia J, Di R H, et al. 2019 Application progress of data mining technology in traditional Chinese medicine research Medical Recapitulate 25 (18) 3672-3676+3681.
[2] Bai Y, Lin J, Ren S, et al. 2019 Application and development of data mining in study of Chinese medical formulas Journal of Shandong University of Traditional Chinese Medicine 43 (01) 24-27.
[3] Li D 2021 Data preprocessing analysis of traditional Chinese medicine prescription based on data mining Integrated Circuit Applications 38 (06) 186-187.
[4] Lei W U and Shu L I 2015 An attempt on data preprocessing for text mining in TCM prescription database Chinese Journal of Library and Information Science for Traditional Chinese Medicine 39 (03) 8-11.
[5] Han J W 2000 Data Mining Concepts and Techniques (Morgan Kaufmann Publishers).
[6] Bhaya W S 2017 Review of data preprocessing techniques in data mining Journal of Engineering and Applied Sciences 12 (16) 4102-4107.
[7] Zhen C G and Zhang Y X 2019 A survey of data preprocessing in data mining International Core Journal of Engineering 5 (9).
[8] Guo R C 2018 Application of data mining in data preprocessing of traditional Chinese medicine Computer Science and Electronic Technology International Society 4.
[9] Zhu J, Ju S and Xin Y 2006 Data mining based approach to preprocessing TCM data set Computer Engineering 2006 (15) 280-282+285.
[10] Jun H J 2008 Design of preprocessing in data mining system for traditional Chinese medicine prescriptions information Computer Engineering 34 (21) 1223-1227.
[11] Tong W N, Li H-B and Wang Y L 2015 Analysis and application of TCM data preprocessing method based on Apriori algorithm Chinese Medical Equipment Journal 36 (09) 55-57+132.