International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 5.22 (SJIF-2017),e-ISSN:2455-2585
*International Conference on*
Recent Explorations in Science, Engineering And Technology (ICRESET'19)
Volume-5, Special Issue-March, 2019.

# Data Mining of Medical records by the Apriori Algorithm using WEKA tool

B.Dhanalakshmi[1]

[1]Assistant Professor ,Department of Information Technology

KGiSL Institute of Technology-Coimbatore,

***Abstract— Apriori algorithm is an association rule mining algorithm, with lot of applications, as it is easy to use. Apriori algorithm is best to find frequently occurring patterns. This paper uses Apriori algorithm to build medical record analysis of patients with different disease. Medical records analysis plays important role diagnosing a patient. Medical record analysis is the process of organizing, summarizing, and analyzing medical records by engaging people with the help of software which takes medical data and gives an output representing all possibilities of cause and effects of the disease. The medical record analysis process, when performed effectively, helps improving trial outcomes of attorneys and settlement results. By data mining based on Apriori algorithm, we can analyze the characteristic of a specific syndrome. Thus paving way for prevention and precaution methods. Weka tool provides effective algorithms and accurate results of analysis of large amount of data. In this paper we are using weka tool to analyze medical records of patients and produce a chart structure depicting the analysis results.***

***Keywords- Data Mining; Apriori Algorithm; WEKA tool***

## I. INTRODUCTION

Apriori algorithm, a classic algorithm, is useful in mining frequent item sets and relevant association rules. Usually, you operate this algorithm on a database containing a large number of transactions. One such example is the items customers buy at a supermarket. It helps the customers buy their items with ease, and enhances the sales performance of the departmental store. This algorithm has utility in the field of healthcare as it can help in detecting adverse drug reactions (ADR) by producing association rules to indicate the combination of medications and patient characteristics that could lead to ADRs.

### Apriori Algorithm – An Odd Name

It has got this odd name because it uses 'prior' knowledge of frequent itemset properties. The credit for introducing this algorithm goes to Rakesh Agrawal and Ramakrishnan Srikant in 1994. We shall now explore the apriori algorithm implementation in detail.

### Apriori algorithm – The Theory

Three significant components comprise the apriori algorithm. They are as follows.

- Support
- Confidence
- Lift

This example will make things easy to understand.As mentioned earlier, you need a big database. Let us suppose you have 2000 customer transactions in a supermarket. You have to find the Support, Confidence, and Lift for two items, say bread and jam. It is because people frequently bundle these two items together.Out of the 2000 transactions, 200 contain jam whereas 300 contain bread. These 300 transactions include a 100 that includes bread as well as jam. Using this data, we shall find out the support, confidence, and lift.

**Support**

Support is the default popularity of any item. You calculate the Support as a quotient of the division of the number of transactions containing that item by the total number of transactions. Hence, in our example,

Support (Jam) = (Transactions involving jam) / (Total Transactions)

= 200/2000 = 10%

**Confidence**

In our example, Confidence is the likelihood that customer bought both bread and jam. Dividing the number of transactions that include both bread and jam by the total number of transactions will give the Confidence figure.

Confidence = (Transactions involving both bread and jam) / (Total Transactions involving jam)

= 100/200 = 50%

It implies that 50% of customers who bought jam bought bread as well.

**Lift**

According to our example, Lift is the increase in the ratio of the sale of bread when you sell jam. The mathematical formula of Lift is as follows.

Lift = (Confidence (Jam – Bread)) / (Support (Jam))

= 50 / 10 = 5

It says that the likelihood of a customer buying both jam and bread together is 5 times more than the chance of purchasing jam alone. If the Lift value is less than 1, it entails that the customers are unlikely to buy both the items together. Greater the value, the better is the combination

## II APRIORI ALGORITHM

### *A. Theory Foundation*

R_ apriori Algorithm is mainly based on rough set theory and Apriori association rules algorithm, the basic knowledge as follows:

#### *1) Rough set theory*

In 1982, Poland mathematician Pawlak proposed rough set theory. Any information system (or information sheet) in rough set theory called I can be describe with an ordered 4-tuple $<U, A, V, f>$, where: $U = \{x_1, x_2 \ldots x_n\}$ are all limited collection of samples, $A=C\ D$ is a set of all finite number of attributes, $C$ is condition attributes set that is the characteristics of the object, $D$ is the decision attribute set that is the type of study object and $C \cdot D = ij$. Suppose $a$ is either an attribute, $x_i$ is either an object, then $f(x_i, a)$ is value of $x_i$ in $a$ attribute, while $V$ is the value range of attribute $A$. [8]

Pawlak attribute importance reduction algorithm confirms mainly the importance of the attribute based on the information system changes in size classification ability while it removed. Attribute importance is defined as follows:

A given set of information systems

$IS$ □ $(U, C, V, f)$, $\forall B \subseteq C$ and $\forall a \in C - B$, define

$$sig(a, B; C) = card(U / ind(BU\{a\})) - card(U / ind(B))card(U)$$

as the attribute importance of $a$ to the attribute set $B$

Where $U$ is the domain of information systems, $C$ is the attribute set, $V$ is the value for the attribute, $f$ is the information function, *ind* is undifferentiated relationship attribute set of the information system between, *card* is the base for the attribute set. General attribute reduction is processed according to the importance of attribute. [10-11]

#### *2) Association rules and apriori algorithm*

First introduce the data mining association rules, association rules discovery data mainly to an interesting correlation between item sets like $A$ Ÿ $B$ such a law can establishment of rules.

Concrete realization depends on the degree of interest rules, including those of support and confidence of two terms.

Degree of support (support) *S* refers to the transaction in the rules the frequency. *A* ☐ *B* degree of support for *S* to

$S(A \square B) = T(A * B) \quad T$

Where, $T(A * B)$ refers to the number of transactions of the data set contains $A * B$ , *T* refers to the total number of that matters. Degree of support is too low, said the rules are not general.Degree of confidence (confidence) *C*, said association rule *A* Ÿ *B* intensity, defined as:

$C(A \text{ Ÿ } B) = T(A * B) \quad T(A)$

Where, $T(A * B)$ refers to the data set contains transactions number of $A * B$ , $T(A)$ refers to data set that contains transactions number of *A*. The lower confidence means less credibility of the rules.

Association rule *A* ☐ *B* means that the confidence level of *A* given *B*, that is the conditional probability, that is

$C(A \square B) = P(B / A)$

Data mining association rules is found to have a user-specified minimum support degree $S_{min}$ and minimum confidence degree $C_{min}$ of association rules. Namely:

*A* ☐ *B* is equivalent:

$(S(A \square B) \geq S_{\min}) \wedge (C(A \square B) \geq C_{\min})$ [9]

### *B. R_Apriori Algorithm Description*

To the problem with clear decision-making field by mining association rules, improved R_Apriori algorithm can be form by integrating rough set theory with the Apriori algorithm to solve the problem raised in the preamble as follows:

About the problem of the efficiency of Apriori algorithm and the validity of the mining rules on account of the large amount of attributes set, we can first get the nuclear of attribute set by rough set attribute reduction, then the association rule mining to the nuclear data. In certain extent, it can improve the efficiency and effectiveness of mining;

For inefficient Apriori algorithm raised from the needs of scanning all attribute sets to obtain each frequent attribute set, we can solve as follow:

 Frequency set can be obtained by scanning the set of attributes, which assumed to be $L_1=\{X_1^1, X_1^2, ...,X_1^M\}$. Denote the set of samples of $X_1^P$ as $S(X_1^P)=\{t_1^P , t_2^P ,..., t_m^P\}$, $1<P<M$. Obviously the number of the element of $S(X_1^P)$ is transactions number, namely $|S(X_1^P)|$ ,and $|S(X_1^P)| / |T| > S_{min}$ ($S_{min}$ is minimum support degree);

Frequent set and above 2-frequency set can be obtained just by set operations of frequent set. According to the character of frequent sets: the attribute which is not attributing of low frequency set must not be attributor of high frequency. So the attributes of 2-frequent sets must include the frequency set $L_1$. For any two item sets $X_1^i$ and $X_1^j$

in $L_1$ (where $1<i,j<M$), $S(X_1^i)$ $S(X_1^j)$ includes all transaction in $X_1^i$ and $X_1^j$. Then its number of elements namely $|S(X_1^i)$ $S(X_1^j)|$ is the number of its transactions. So if $|S(X_1^i)$ $S(X_1^j)| / |T| > S_{min}$, then $\{X_1^i, X_1^j\}$ $L_2$, otherwise $\{X_1^i, X_1^j\}$ $L_2$;

then intersection operation to all 1-frequency set and to get their support, you can get $L_2$; if $\{X_1^i, X_1^j\}$ $L_2$, you can also calculate $|S(X_1^i)$ $S(X_1^j)| / |S(X_1^i)|$ and $|S(X_1^i)$ $S(X_1^j)| / |S(X_1^j)|$ which is the confidence of $X_1^j => X_1^i$ and $X_1^i => X_1^j$. And

comparison with the minimum confidence degree $C_{min}$, association rules is determined.

- Thus intersection operation to the frequent item sets $L_{K-1}$ and $L_1$, and then determines the number of elements in the intersection set, all frequent sets $L_K$ can be obtained. For the problem of the sort of strength of the rule model, it could be sorted by support degree, confidence degree and order of mining:

Given two rules $r_i$ and $r_j$, $r_i > r_j$ (i.e., $r_i$ precedes $r_j$ or $r_i$ has higher precedence over $r_j$) if one of the following holds good:

- The confidence degree of $r_i$ is greater than that of $r_j$
- Their confidences degree are the same but support of $r_i$ is greater than that of $r_j$
- Both the confidences and supports of $r_i$ and $r_j$ are the same, but $r_i$ is generated before $r_j$

The algorithm described as follows:

*U = RS(U,C,D,f)*

//attribute reduction according to decision set

*L(1) = Apriori(U,1)= {$X_1^1$, $X_1^2$, ... ,$X_1^M$}*

//Calculate 1-frequency set

*$S(X_1^P)$={$t_1^P$ , $t_2^P$ ,... $t_m^P$}  1<P<M*

//record of 1-frequency set transaction sets

*k=2*

*Do while L(k)=null{*

*L(k)=null*

*For i=1 to | L(k-1)|*

*For P=1 to | L(1)|*

*If $X_1^P$ E(L(k-1)) then*

// E(L(k-1)) is( k-1)- frequency set

*If | $S(X_1^P)$∩ S(E(L(k-1)))| /| T |> $S_{min}$ then L(k) = L(k) U{ $S(X_1^P)$∩ S(E(L(k-1)))}*

//Is the elements of k-frequency set

*If | $S(X_1^P)$∩ S(E(L(k-1)))| /| $S(X_1^P)$| > $C_{min}$*

*then*

//To determine whether S(E(L(k-1))) => $S(X_1^P)$ is the rule

*R = R U{ S(E(L(k-1))) => $S(X_1^P)$} Sort(R, S(E(L(k-1))) => $S(X_1^P)$)// Sort*

*if | $S(X_1^P)$∩ S(E(L(k-1)))| /| S(E(L(k -1)))| > $C_{min}$ then* //To determine whether $S(X_1^P)$ => S(E(L(k-1))) is the rule

*R  = R U{ $S(X_1^P)$ => S(E(L(k-1)))} Sort(R, $S(X_1^P)$ => S(E(L(k-1))))//sorted*

*Next*

*Next*

*}*

Output rules R

## III RESULTS

Purpose of this study in patients with bacterial pneumonia in the elderly as a target for observation. Clinical experience with a rich selection of TCM clinical experts, the research objects dialectic treatment, to observe and collect large number index information of different time points. Of these clinical data, to analyze the process of TCM syndrome differentiation changes of various indicators and their mutual relations, screening older persons with bacterial pneumonia certificate of TCM clinical efficacy evaluation index, we need to adopt data mining processing techniques (data preprocessing, modeling, methods selection, etc.) . Data mining is found in large amounts of data from the hidden knowledge and laws, both as a knowledge access technology, but also a data processing. The technology evolved from artificial intelligence, so many of the technical achievements of artificial intelligence can be transplanted into the data mining systems, such as the traditional statistical, clustering, decision tree, set theory, correlative rules, rough set theory, artificial neural network, genetic algorithms and evolutionary computing. For exploring the intrinsic link between multiple variables, artificial intelligence, data mining techniques, correlative rules is used mainly to seek contact, while the R_Apriori correlative rules algorithm is commonly used. Therefore, based on R_ Apriori algorithm for correlative rules is used to explore the course of treatment, changes in the relationship between the various indicators.

Acquisition from Henan Medical College Hospital, Beijing Medical University Affiliated Hospital, Shandong Medical University Affiliated Hospital, Nanjing Medical University Affiliated Hospital and Affiliated Hospital of Changchun University of Chinese Medicine and other five hospitals, 450 cases of elderly patients with bacterial pneumonia admitted with including basic information, medication day, the 4th day, the 7th day and the 14th day. It also includes syndrome indicators, the efficacy of Western diseases, indicators of quality of life indicators, a total of 285 indicators for the information. The above information as a basic sample set of data mining, is completed data management by MICROSOFT SQL SERVER database.

### *A. Data Preprocessing*

#### *1) To deal with the null values data in the database*

Using PowerBuilder the database is processed. The tables are incomplete in some indicators, data collection items phenomenon (target item data to a null value), for the difference of the data to carry out (null value data are not subtracted), null value data items in the database will be assignment for the arithmetic mean of the indicator data.

#### *2) To reduce data dimensionality*

Processed by statistical analysis of indicator data, it is found that many indicators data, the probability of a margin of 0 for more than 95%, meaning that at some stage in the treatment of the index did not change significantly. Such as indicators of FT (expressed abdominal pain), WFY (verbal response, said no), GZ (said tongue dry), H5 (indicated pulse Hung), HB (said tongue flower stripping), KT (indicated expectoration), is as Fig. 1 which the change in statistical law.

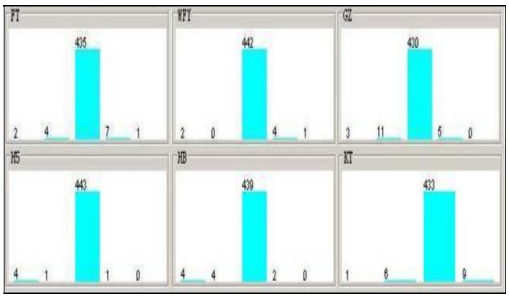


Figure 1.Change in indicator data

Because of these indicators change was not obvious, they can be understood not to be the main indicator of treatment of pneumonia from the medical diagnosis. They will be to discard in the data analysis, resulting in data analysis to reduce the process indicators and to improve the algorithm efficiency. To some extent to reduce mining results invalid.

*3) To change numeric variables into multi-value variable*

As the R_Apriori algorithm is not sensitive enough to deal with numerical variables, five numeric variables indicator data such as body temperature, heart rate, respiration, systolic blood pressure and diastolic blood pressure are converted to multi-valued variables. Based on the above five indicators for the analysis of the data distribution, the conversion is processed. For example,body temperature, in general, lower than 36 called low temperature, 36-38 called normal temperature, 38-39 called low-grade fever, more than 39 called high fever. So it can keep the body temperature indicator in accordance with the above categories and is divided into four categories ABCD .

*B. Data Mining Analysis of R_Apriori Algorithm*

Using R_Apriori algorithm, we have a data mining process after the initial various index data. Seven types of syndrome diagnostic criteria are obtained through the analysis and of in the orderly's pneumonia as follows:

- Wind-cold Syndrome: Fever, Cough, Stethoca-tharsis, thin tongue fur, White tongue fur, Floating pulse
- Wind-heat Syndrome: Fever, Aversion to wind, Headache, Cough, Stethocatharsis, Red tongue, Yellowish fur
- Phlegmatic Hygrosis Syndrome: Cough, Excessive phlegm, Tur chest and diaphragm, Anorexia, Dyspnea and tachypnea, abdominal distension
- Phlegm-heat Syndrome: Fever, Cough, Stethoca-tharsis, Chest pain, Yellowish fur, Greasy fur, Dry stool, Slippery pulse
- Fire-heat Syndrome: Fever, Cough, Stethoca-tharsis, Red tongue, Yellowish fur, Slippery pulse
- Qi Deficiency Syndrome: Cough, Stethoca-tharsis, Fatigue and lack of strength, White fur, Spiritlessness
- Yin Deficiency Syndrome: Cough, Night sweat, Stethoca-tharsis, Sputumless, Dry stool, Thready rapid Pulse

After comparison with TCM clinical medicine, these TCM syndrome diagnosis standard of the orderly's pneumonia is matched up with the actual situation on the whole.[12-13] The method of building syndrome diagnosis standard of Traditional Chinese Medical by R_Apriori algorithm has much practical significance and worthy of promotion.

## ACKNOWLEDGMENT

R.B.G. thanks National Nature Science Foundation (973 Program) under Grant number: 2006CB504605. "WEKA" stands for the Waikato Environment for Knowledge Analysis, which was developed at the University of Waikato in New Zealand. WEKA is extensible and has become a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost every platform. WEKA is easy to use and to be applied at several different levels. You can access the WEKA class library from your own Java program, and implement new machine learning algorithms. There are three major implemented schemes in WEKA. Implemented schemes for classification. Implemented schemes for numeric prediction. Implemented "metaschemes". Besides actual learning schemes, WEKA also contains a large variety of tools that can be used for pre-processing datasets, so that you can focus on your algorithm without considering too much details as reading the data from files, implementing filtering algorithm and providing code to evaluate the results. This tool is very useful for data mining and analysis medical related data can be analysed very effectively.

## REFERENCES

[1] Cai Yue-jun, Data Mining Technology and Its Application in Traditional Chinese Medicine[D], Hangzhou: Zhejiang University, 2003

[2] Cheng-hua, Application Research of Data Mining in Diabetes Data[D], Beijing: Institute of Software Chinese Academy of Sciences, 2003

[3] Zhou Jie, The Research of Data Mining in Several Ways and its Application in the Databases Chinese Medicine [D], Southwest Jiaotong University, 2003

[4] Liu Qiang, Study on Data Mining and Network Structure about the Heart-Qi Deficiency Syndrome [D], Changsha: Hunan University of Traditional Chinese Medicine, 2003

[5] Yu Hui, Study on Medical Knowledge Acquirement and Discovery [D], Tianjin: Tianjin University, 2003

[6] Tang Zhong-liu and Wang Zhe-dong, Genetic algorithm in the application of medical discriminant analysis[J], Acta Academiae Medicinae Suzhou, 2000,683-684

[7] ZOU Zong-feng, Data Mining in the Application of Case-mix Control Program in the Hospital Charges Development [D], Guangzhou: Jinan University, 2003

[8] Luo Laipeng. Algorithm for mining frequent itemsets based on rough set [J]. Computer and Modernization. No. 10 (2005)

[9] Jiawei Han,Micheline Kamber. Data Mining Concepts and Techniques[M].Fan Ming, Men Xiaofeng, translation. Beijing: China Machine Press,(2001)

[10] Jue Wang, Duo-Qian Miao, Yu-jian Zhou, "Rough Set Theory and its Application: A Survey", Pattern Recognition and Artificial Intelligence, (1996),9 (4): 337 ~ 344

[11] Z.Pawlak. Rough Sets, International J. of Computer and Sciences [J]. (1982),11(5)341 ~ 356

[12] Du Wenbin, The Research of Symptomatic Coronary Artery Disease Diagnostic Standard and Pharmacodynamic Evaluation Model Based on Neural Network [D], Shenyang: Liaoning University of Traditional Chinese Medicine, 2004

[13] Yu Xue-qing, LI Jian-sheng, etc.. Literature Analysis on Pattern Types and Symptom Features of Pneumonia[J]. Acta Universitatis Traditionis Medicalis Sinensis Pharmacologiaeque Shanghai, Vol 22, No. 2, (2008): 26 ~ 29