Medical Diagnosis Data Mining Based on Improved Apriori Algorithm

Wenjing Zhang
College of Information Science & Technology, Agricultural University of Hebei, Baoding, China

Donglai Ma*
Hebei Software Institute, Baoding, China

Wei Yao
College of Information Science & Technology, Agricultural University of Hebei, Baoding, China

Abstract—With the wide application of computer science and technology, the amount of data generated by various disciplines increased rapidly. In order to discover valuable knowledge in these databases, people use data mining methods to solve this problem. The application of association rule mining is an important research topic in data mining. As the association rule technology becomes more mature, it is a new research that how to use this method to find out the intrinsic association rules from a large number of medical data, providing an effective basis for clinical disease surveillance, evaluation of drug treatment and disease prevention. This paper uses Apriori, the classic algorithm of association rule, for data mining analysis of medical data. According to the characteristics of medical data, it improved the Apriori algorithm. Using the improved Apriori algorithm, it finds frequent item sets in a database of medical diagnosis, and generates strong association rules, in order to find out the useful association relationship or pattern between the large data item sets. The results show that, the improved Apriori algorithm can dig out association rule models about the properties and nature of the disease from a medical database, which can assist doctors in medical diagnosis. Therefore, it is a worthy research direction that using data mining method to process and analyze the data of disease prevention and drug treatment in the field of medicine.

Index Terms—Data Mining; Association Rule; Apriori; Expert System; Knowledge Base

I. INTRODUCTION

In recent years, the applications of data mining technology in medical field are more and more widely. It obtained encouraging results in disease diagnosis, the treatment, organ transplantation, genetic research, image analysis, rehabilitation, drug development, scientific research, etc. Spine Hospital in University of Southern California used Information Discovery for data mining, and this technology has been applied to oncology, liver pathology, survival probability prediction of hepatitis, predict, urology, diagnosed cases of thyroid, rheumatology, dermatology diagnosis, cardiology, neuropsychology, gynecology, obstetrics and other medical fields [1]. Jiawei Han and Micheline Ka-mher elaborated the application problems of data mining in the DNA data analysis field, such as the semantic integration of heterogeneous and distributed genetic data, search and compare the DNA sequence similarity, gene sequences identification, disease-causing genes in different stages of the disease. Muggleton proposed using inductive logic to program, and predict a second structure of protein according amino acid sequence information. From the perspective of medical diagnosis, Igor Kononko etc. elaborated the applications in the medical field of statistical or pattern recognition method, symbol rule inductive learning, and artificial neural network these three types of machine learning algorithms. According to the analysis of ECG, EEG and other medical speculation signals, Miroslav Kubat proposed to use a decision tree to initialize the neural network, which can greatly improve the classification accuracy of test samples. Vysis used neural network technology for the analysis of protein drug development. Robeit Groth studied on the application of clustering technology in the patients rehabilitation after operation. In China, Chen Aibin, Xia Limin of Central South University studied face detection using boosting machine learning methods. Chen Xuefeng [2] of the Fourth Military Medical used the database and data mining technology to build a database of hematologic malignancies analysis system, which could not only help doctors make a preliminary diagnosis but also has powerful mining and analysis capabilities on data. Fu Chunfeng [3, 4] studied the application of LogitBoost, one machine learning method, in discriminant analysis and its application in the medical field prospect. Data mining applications in medicine has its own advantages, because the data collected from medicine is general true and reliable, not affected by other factors, and the data set have strong stability. All of these are very favorable conditions for the mining results maintenance and continuously improving the quality of mining models [5]. With the introduction of electronic medical records, medical records with computer storage have been more common in the hospital. If the collected data will be further aggregated from each hospital, the amount of data is quite large, and the data are real patient
data. From this data set, using a variety of data mining techniques to understand the mutual relations and development law of various diseases, summarize the therapeutic effect of various treatment options, as well as disease diagnosis, treatment and medical research are very valuable. Medical information mainly includes pure data such as sign parameters, laboratory results, images such as EEG, ECG signals, B ultrasound, CT; texts such as patient’s identity recording, symptoms description, as well as animations, voice, video information for science education. It has characteristics of large information capacity, complex data types, multiple redundant vacancy value and cumbersome content association [6].

For these features of medical information, there is a great difference between medical data mining and ordinary data mining, determining the specificity of medical data mining. Medical database is a huge data resource, and a lot of the same or part same information stored in it each day. For example, to some diseases, the patient presenting symptoms, laboratory results, the treatment may be exactly the same. Before data mining, these large number of vague, incomplete, noise original information, must be cleaned and filtered to ensure data consistency, and turn into a suitable form for mining.

Data mining technology is good at finding implicit and meaningful knowledge from massive data which is lack of priori information, predicting future trends and behavior, and making proactive decisions based on knowledge. It is this advantage makes data mining techniques are widely used in medical research and has made many valuable results.

The medical diagnosis expert system turn the diagnose experience of experts into the rules. As long as we input the patient’s symptoms to the system, it will be able to make judgments quickly, thus reducing errors of the doctor’s subjective judgment. However, the diagnostic criteria are drawn up based on the experience of one or a few experts, which is lack of objectivity and universality. In addition, the inference rules and conclusions of experts system are pre-designed, and the clinical manifestation of some patients may not within this range. Therefore, it has some limitations [7]. Data mining can dig out valuable diagnostic rules, by processing a large historical data of patient data base. Thus, doctors can make conclusions according to the patient’s age, gender, auxiliary examination results, the physiological and biochemical indexes etc., eliminating the interference of human factors and enhancing the objectivity. In addition, due to the large amount of data processing, the diagnostic rules have better application universality. Currently, there are many foreign success cases in this regard. Such as using Bayesian learning classification method to diagnosis CT image automatically [8], using machine learning methods to analysis breathing pressure-volume curve of patients in intensive care [9], using association rules to find CT examination indications of head trauma patients [10], using data mining for the automatic detection of liver cancer genetic syndrome and the research on racial differences of non-malignant respiratory system disease epidemic of uranium miners [11, 12]. All of these have achieved the desired results, and show the broad application prospect of data mining techniques for the disease diagnosis [13].

With the continuous improvement of hospital informatization level in recent years, especially the establishments of integrated information database of the hospital, doctors have collected a large amount of patient’s information data. These data indicate the information between the diseases classification and patient’s clinical manifestation, i.e., the disease diagnosis knowledge. How to quickly, effectively establish a relatively complete knowledge base is the recognized bottleneck problem for expert system. The quality of knowledge base could affect the stability and accuracy of the system directly. For the database of confirmed case, only part of the attribute or value of the database table has to be reserved. If we can reduce the unwanted attribute or delete the values, it will greatly enhance the clarity of the potential knowledge of the system, accelerate the speed of the subsequent inference module, and improve the statistical performance of system classification.

II. COMPUTER AIDED DIAGNOSIS & KNOWLEDGE BASE

A. Computer-aided Diagnosis

The application of Computer-aided diagnosis in medical can be traced back to 1950s. In 1959, Ledley, an American scholar, introduced mathematical model into clinical medicine for the first time, presenting a mathematical model of computer-aided diagnosis and diagnosing a group of lung cancer cases, which creates the precedent for computer-aided diagnosis. In 1966, Ledley first proposed the concept of “computer-aided diagnosis” (CAD). In the early 1980s, the computer-aided diagnosis system had a further development, among them, the expert system, which is applied in Chinese medicine field, is most compelling. The computer-aided diagnosis process includes gathering patient medical information and inspection data, quantification medical information, statistical analysis, until finally gets diagnoses. The more popular models at that time are Bayes theorem, the maximum likelihood model etc.. Since 1990s, the artificial neural network developed rapidly, which is a mathematical method to imitate the working principle of the human brain neurons. Because of its ability to self-learning, memory capacity, forecasting events, etc, artificial neural network has superior performance than traditional methods such as probability statistics, mathematical models, etc. It can be said that the artificial neural network is one of the most advanced artificial intelligence technology currently.

Since the 1960s, it has been reported that using computer to analyze medical image data. But after ten years, due to various reasons such as computer technology, the research on CAD once fell into a trough. On the one hand, because people expected too much on the CAD, hoping to achieve automatic diagnosis by means of computer, on the other hand the research result of CAD is not ideal. Until 1980s, due to the rapid development of computer technology as well as a variety
of mathematical and statistical technology, CAD got a fast development in the field of medical imaging in some developed countries, and has made gratifying achievements [14]. Currently, foreign scholars basically reached a consensus on the implication of CAD in medical image. That is the final diagnosis of CAD is determined by a doctor, just consulting the output of computer, which makes diagnosis result more objective and more accurate. At present, foreign scholars emphasize that the computer’s output is only as a second opinion. And these is different from the CAD idea in the originally the sixties and some people nowadays [15]. In medical imaging, the computer output result is obtained from quantitative analyzing the related image data, whose role is to help radiologists improve the diagnostic accuracy and the consistency of interpretation on image and disease.

The reason why CAD can improve the diagnostic accuracy of the doctor is that, the radiologist’s diagnostic is subjective, which would be limited and impacted with the doctor’s experience and knowledge level; Secondly, the doctor diagnosis is easy to miss some subtle changes, such as lung nodules, slight calcification within the breast etc [16]; Once more, different doctors have differences on reading the image. While, computers have a great advantage to correct these errors and deficiencies.

At present, the computer-aided is relatively mature in the lungs and breast lesions; while the research on brain tumor, cerebral perfusion, liver disease, virtual cavity mirror and Chinese medicine is at the start stage. B. The Functions of the Knowledge Base in Medical Assist Diagnosis Expert System

Computer aided diagnosis system is provided in the form of advisory system, whose purpose is to make the diagnosis process to be more objective, easier and effective. Besides, it could also be used to train new recruited inexperience doctors. At present, the practicality of the knowledge system in medical database is still in the research and development stage. Some knowledge discovery diagnosis systems identified many errors after discovery knowledge, and then verified and modified it. Besides, some other systems need further enrichment of the disease database in order to obtain more accuracy diagnosis rules and better predict disease. In addition, there must be inconsistency or contradiction between medical experts and the knowledge discovered from database, which requires the cooperation of development persons and medical experts to explain and eliminate the contradiction. This is the most important feature of the computer aided diagnosis system and the integration development direction of knowledge discovery diagnosis system and expert system in the future. In this paper, we developed a computer aided diagnosis system of breast cancer, which is a typical example in this domain. It developed a consistent rule base by the following steps: (1) find the data driven rules by database, rather than by asking experts; (2) analyze these new rules by medical experts with the confirmed case; (3) discovery the rules conflicted with the knowledge or understanding of he/she, which means two possibilities: (a) this rule is found using misleading disease cases and is must to be rejected, and we should also extend the training data; (b) experts may recognize that his/her ideas have no real basis, thus the system will increase the experience of experts [17].

III. DATA MINING

A. Data Mining Concept

Data mining is an emerging cross-disciplinary technology, involving artificial intelligence, database, statistics, machine learning, information retrieval, high-performance computing and some other fields. Data mining technology has evident advantages in massive data processing, potential information finding and association rules discovering. The effect of data mining is extracting interested knowledge from large databases. These knowledge is implicit, previously unknown and potentially useful information, and the form of extracted knowledge are conceptual, pattern, rule, etc.. In short, data mining refers to extracting the hidden, previously unknown and potentially useful information from database, and then offering the understandable knowledge, such as association rules, cluster patterns etc, so as to support users for decision-making. Now, data mining has been used in various fields such as business, finance, meteorology, astronomy, power, etc, showing more and more importance and superiority [18]. The process of data mining is shown in figure 1.

B. Data Mining Applications in Medicine

1) Applications of Sequence pattern analysis in DNA data analysis

Currently, a large number of biomedical researches have focused on the analysis of DNA sequences. The cause of one disease gene may be more than one. The recent results of DNA analysis have discovered many diseases and disabilities gene, and new drugs and methods for disease diagnosis, prevention and treatment. Typically, the sequence whose occurrence frequency in disease sample exceeds its in health sample can be considered as disease genes; the other hand, the sequence whose occurrence frequency in health sample exceeds its in disease sample can be considered as disease resistance genes. Different genes may play different roles at different stages of the disease. If the sequence of genetic factors can be found at different stages of disease development, there may be possible to develop therapeutic drug aimed at different stages of the disease, so as to obtain a more effective treatment result.

2) Application of neural network in clinical analysis
Based on the understanding of human brain structure and operation mechanism, neural network is one engineering system to simulate the human brain structure and intelligent behavior. According to the study results shown from the neurobiologist, the human brain has $1.4\times10^{11}$ neurons. The reason why brain can be able to handle extremely complex analysis and reasoning work, partly because of its large number of neurons, on the other hand lies in neurons can process the input signals in a non-linear way. Therefore, we can establish a neurons mathematical model closer to the engineering, which is a multi-input, single-output nonlinear element. Neural network is composed of a group of neurons, which determines the neural network has the ability of parallel processing. Each neuron of neural network uses M-P model, divided into input unit, output unit and hidden units these three categories. The input unit accepts the input signal from the outside world, and the hidden layer unit is the internal unit of the system, without any contact with the outside world. Figure 2 is the topology structure of three-layer feed forward neural network. Its characteristic is that neurons in each layer only have connection with neurons at adjacent layers, neurons in one layer have no connection and neurons in every layer have no feedback connection. The input signal first spreads forward to the hidden node, through the transformation function, sends the output information of hidden node to the output node, and then gives the output results. The transformation function of node is usually selected “Sigmoid” type.

For example: human heart disease has many factors, we have established prediction model of heart disease based on domestic and international references with sixteen input nodes, five hidden layer nodes, and one output node represented the induced likelihood of heart disease. We first trained neural network models, with the data of the eighty heart patients of Affiliate Hospital of Taishan Medical College in January to June of 2004. Then predicted the undiagnosed patients, and finally get a more satisfactory result.

3) Application of association analysis in medicine

With the increasingly serious problem of the population aging, people pay more and more attention on the diabetes study. Researchers began to build a physical database of patients with diabetes, analyzing physiological parameters such as blood glucose concentration, age, sex, bone density, ECG, blood pressure, muscle and fat tissue content, hoping to discover new medical knowledge. Association analysis is the most widely used and powerful tool in such topics of multidimensional data analysis in the diabetes database.

Feng Bo et al of Affiliated Hospital of Tongji University of Shanghai used statistical analysis on the relation between muscle tissue and fat tissue content and bone density of diabetes patients and got the conclusion that high muscle and high fat content can reduce the risk of patient hip fracture.

4) Application of data mining in disease diagnosis

There are many domestic and international examples in this regard. Such as using Bayesian classification methods to automatic diagnosis CT images of male and female patients; using machine learning methods to analyzing pressure-volume curve of patients’ breath in intensive care; the application of automatic detection in liver cancer heredity, and so on. All of these show the widely application of data mining in medical field.

Data mining technology develops rapidly and various algorithms emerge in endlessly. In the medical field, data mining has its own particularity, that is various information (such as laboratory results, B-ultrasonography, CT, X-ray, cases etc.) mutual interrelated and integration. How to make good use of this information is a topic worthy of further exploration. It is believed that, with the establishment of medical electronic files and the development of the biomedical engineering, data mining technology will play an increasingly important role [19].

IV. ASSOCIATION RULES

A. Association Rules

Association rule is a very important and one of the most active knowledge model of data mining [20]. It was firstly put forward by Agrawal et al. in 1993. Association rule represents the related relationship of a group of objects in database. It focuses on the relationship of different attribute domain, which can meet certain requirements. On one hand, association rules can provide high efficient method of pattern discovery and model recognition, verify the existing experiences and rules, find out new experiences and rules and provide important assistant tool for natural science research. On the other hand, combining with other decision-making method, association rules can provide scientific and effective basis for the decision makers. After association rule was proposed, many researchers including Agrawal himself conducted in-depth research on mining association rules, and improved and extended the initial algorithm in order to promote the practical application value of it.

The basic model of association rules is described as follows:

Set data set $D=\{T_1,T_2,\ldots,T_n\}$, while $T_j(j=1,2,\ldots,n)$ is called affair. The elements of $T$, $ik(k=1,2,\ldots,p)$ are called item. Let the set of all items in $D$ is $I=\{i_1,i_2,\ldots,i_m\}$. Obviously, $T \subseteq I$.

Concept 1: Support and Confidence. The support of association rule “$A\Rightarrow B$” refers to the proportion of the transaction sets which contain both item $A$ and item $B$ in all the transactions of data set $D$, as in (1). The confidence of association rule “$A\Rightarrow B$” refers to the
proportion of transactions which contain both item A and item B in all the transactions that contain item A, as in (2).

\[ \text{sup of } \langle A \Rightarrow B \rangle = P(A \cup B) = \frac{\text{count}(A \cup B)}{\text{count}(D)} \]  

(1)

\[ \text{confidence of } \langle A \Rightarrow B \rangle = P(B | A) = \frac{\text{count}(A \cup B)}{\text{count}(A)} \]  

(2)

Concept 2: the strong association rules. If there is an association rule \( \langle A \Rightarrow B \rangle \) whose support and confidence are satisfied minimum support threshold (\( \text{min\_support} \)) and minimum confidence threshold (\( \text{min\_conf} \)) that presented by users, we call it strong association rule. Strong association rules are of interest to users, and have the important guiding significance to users in discovering potential rules in large amount of data.

B. Apriori

Apriori algorithm use a layer by layer search iterative methods, which use frequent k items sets (the collection contains k items, and the appearance frequency of the combination of these k items is higher than the minimum support of the preset) to find frequent (k+1) items sets. Firstly, find out frequent 1-items sets, denoting as L1. Secondly, use L1 to find out frequent 2-items sets and denote them as L2. Then, use L2 to find L3, and so on, until Lk which could not meet the minimum support. That is the frequent k-items sets [21].

This algorithm generates the candidate sets as small as possible. In order to achieve it, the following two properties are used:

(1) Any non-empty subset of frequent itemsets must also be frequent itemsets.

(2) Any superset of non-frequent itemsets must also be non-frequent itemsets.

Apriori algorithm mainly includes the following three steps:

(a) Connecting step. Connect frequent (k-1) items sets Lk-1 to generate candidate set Lk of Lk. The connected condition is that, the (k-2) items of two (k-1) items sets are equal, and the (k-1) item of the first (k-1) items sets is smaller than that of the second one.

(b) Pruning step. Prun the candidate set Lk of k items sets with the apriori algorithm. The pruning rule is, if any (k-1)items of one k-items sets is not belong to frequent (k-1) items sets Lk-1, it would not become a frequent k-items sets, which should be removed from Ck.

(c) Counting step. Scan the transaction database, accumulating the number of candidate set k in database. For one candidate set, if a transaction record contains it, then the number of it plus 1. Finally, according to the pre-given minimum support, generate frequent k-items sets Lk.

V. THE IMPROVEMENT OF APRIORI IN KNOWLEDGE BASE BUILDING

The classical association rules focuses on the association relationships between a group of the objects in the transaction database. For example, the rule “85% of people, who had bought A and B, will buy C and D” can be expressed as “\( A \land B \land C \land D \)”. This rule pattern mainly is suitable for drawing up the targeted marketing strategy and goods shelves designing, etc. While for breast cancer diagnosis expert system, the rules we need are like “\( A \land B \land C \land D \)”, that means there must be only one attribute in the right of the arrow, and we call that “decision rule”. This requires some modification for the classic Apriori algorithm.

Five breast disease data sets (table 1) were selected for rule mining. Set the minimum support threshold as 40% and minimum confidence threshold as 80% [22-23].

<table>
<thead>
<tr>
<th>TABLE I. ORIGINAL MEDICAL DATA TABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>no</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>001</td>
</tr>
<tr>
<td>002</td>
</tr>
<tr>
<td>003</td>
</tr>
<tr>
<td>004</td>
</tr>
<tr>
<td>005</td>
</tr>
</tbody>
</table>

A. The Establishment of Mining Data Items

In the classical apriori algorithm, we should set “age”, “position”, “calcify”, “tissue”, “class” as the basic data items. However, in medicine, each attribute will have several different values. For example, “tissue” attribute has three values, i.e., “dense”, “fatty” and “glandular”. Here, we use the form like “attribute.attribute value” as the final data item. The data items are shown in table 2.

<table>
<thead>
<tr>
<th>TABLE II. ATTRIBUTE TABLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>1.1</td>
</tr>
<tr>
<td>1.2</td>
</tr>
<tr>
<td>2.1</td>
</tr>
<tr>
<td>2.2</td>
</tr>
<tr>
<td>3.1</td>
</tr>
<tr>
<td>3.2</td>
</tr>
<tr>
<td>4.1</td>
</tr>
<tr>
<td>4.2</td>
</tr>
<tr>
<td>4.3</td>
</tr>
<tr>
<td>5.1</td>
</tr>
<tr>
<td>5.2</td>
</tr>
</tbody>
</table>

According to the attribute table, we can get the mining database after converting the original database. The mining database is shown in table 3.

<table>
<thead>
<tr>
<th>TABLE III. MINING DATABASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>001</td>
</tr>
<tr>
<td>002</td>
</tr>
<tr>
<td>003</td>
</tr>
<tr>
<td>004</td>
</tr>
<tr>
<td>005</td>
</tr>
</tbody>
</table>

B. Mining Frequent Item Sets

The first step of Apriori algorithm is to generate frequent 1-items sets, as shown in Table 4.

In the procession of generation candidate 2-items sets from frequent 1-items sets, it would got candidate item sets such as “1.1, 1.2”, “4.1, 4.2”, “4.1, 4.3”, “4.2, 4.3” ect. To “2.1, 2.2” for example, this item is meaningless in practice. We can see from table 2, “2.1” and “2.2” respectively are “L” and “R”, which are the different
values of attribute “position”. They are mutually exclusive relationship, and can not coexist. Thus, even it is in candidate 2-itemsets, it can not be the frequent item sets. Therefore, this situation is not connected to generate a new set of candidates.

Mining table 3 according to this improved method, we can get the final frequent 4-itemset “1.1,2.2,3.1,5.1”.

### C. Generate rules

In classic Apriori algorithm, the method of generating association rules is: generating all the non_empty subsets of frequent item set L, and calculating the confidence of each non_empty subset S. If the confidence is not less than the minimum confidence threshold given by the users, it will generate a association rule “S->(L-S)”.

According to this method, the right of the rules is likely to be more than one attributes, such as: 1.1 ∨ 2.2 ∨ >3.1 ∧ 5.2.

However, in the breast cancer diagnosis system, we hope the rule could help doctors to make their decision, rather than only supply a few attributes association. That is to say, firstly, the attribute “class” (5.1 or 5.2) can only in the right hand of the rule. Secondly, there could be no other attribute in addition to the attribute “class” in the right hand of the rule. The rule conforms these two conditions is significant. According to these requirements, we can only calculate the confidence of the subsets excepted attribute “class” when we make rules (that is to ensure the right of the rule “L-S” is “class”). For example, for the final frequent 4-itemsets “1.2,2.2,3.1,5.1” mined in 2.2, we calculated the confidence of all the subsets removing attribute “class”. Only the subset “1.2,2.2,3.1” could meet the condition “whose confidence is not less than the minimum confidence threshold”. Thus, we generate the rule “1.2,2.2,3.1->5.1”. At last, combined with the initial attribute value (table 2), the rule is converted into intuitive form: “age=40 and position=R and calcify=1 Y class=abnormity (support is 40% and confidence is 100%)”. Such rules could help doctor for diagnosis.

### VI. CONCLUSIONS

In this paper, association rules mining is applied to case diagnosis of the breast cancer in hospital to find out the association of factors from volumes of case recordings. By improving the Apriori algorithm of association rules, this study mined decision rules from breast cancer diagnosis database and found out the relationship between breast cancer and the factors such as age, position, calcify, etc. Combined with the clinical experience of doctor, data mining technique is also used to set up a relatively integrated assisted diagnostic knowledge database and set groundwork for subsequent consequence and explanation program.

### ACKNOWLEDGMENT

This work is supported by Scientific research fund of Hebei Education Department (Q2012139) and by the science technology research and development projects of Baoding in 2012 (12ZG004, 12ZG002, 12ZG003), science and technology research projects of hebei construction (2012-132), philosophy and social sciences planning project of Baoding (201201157, 201201134).

### REFERENCES


