COVID-19 patient diagnosis and treatment data mining algorithm based on association rules

Zicheng Shan | Wei Miao

Abstract
Association rules are used in different data mining applications, including Web mining, intrusion detection, and bioinformatics. This study mainly discusses the COVID-19 patient diagnosis and treatment data mining algorithm based on association rules. General data The key time interval during the main diagnosis and treatment process (including onset to dyspnea, first diagnosis, admission, mechanical ventilation, death, and the time from first diagnosis to admission, etc.), the cause of death by laboratory examination, and so forth. The frequency of drug use was counted and association rule algorithm was used to analyse and study the effect of drug treatment. The results could provide reference for rational drug use in COVID-19 patients. In this study, in order to improve the efficiency of data mining in data processing, it is necessary to pre-process these data. Secondly, in the application of this data mining, the main objective is to extract association rules of COVID-19 complications. So its properties for mining should be various diseases. Therefore, it is necessary to classify individual disease types. During the construction of association rules database, the data in the data warehouse is analysed online and the association rules data mining is analysed. The results are stored in the knowledge base for decision support. For example, the prediction results of the decision tree can be displayed at this level. After the construction of the mining model, the display interface can be mined, and the decision-maker can input the corresponding attribute value and then predict it. 0.76% of people had both COVID-19, CHD and hypertension, while 46.5% of people with COVID-19 and CHD were likely to have hypertension. This study is helpful to analyse the imaging factors of COVID-19 disease.

KEYWORDS
association rules, COVID-19 patients, data warehouse, diagnosis treatment data mining, online analytical processing

1 | INTRODUCTION

In recent years, great advances have been seen in the ability to perform effective association rule mining (ARM). The birth of artificial intelligence has provided many more effective new technologies for data mining, and has made great progress in data mining, which has alleviated the phenomenon of “massive data.” At present, data mining has important application value in many aspects. By describing the existing data, it can effectively predict the future pattern of data.
There are more than 3000 known viruses, only a small fraction of the total number of viruses in nature. Although not all viruses can cause serious infectious diseases in human populations, there are still a large number of unknown viruses in nature that may cause great harm to human populations. The boundary between human society and nature is gradually blurring, resulting in many viruses that have long lived in nature are entering people for the first time. Take, for example, the recent outbreak of viral pneumonia caused by the 2019-ncov virus (collectively referred to as COVID-19).

Using association rules in data mining is one of the most relevant tasks in modern society. Dahbi A believes that one of the main problems associated with discovering these associations (which decision makers can face) is the extraction of a large number of association rules. The knowledge post-processing stage becomes very challenging in terms of ranking and selecting the most interesting AR. He has proposed various interest measures as a post-processing stage. But the richness of these measures presents a new problem, for there is no best measure and no measure that is better than any other measure. To overcome this challenge, he proposes a new algorithm based on dominance relations, which aims to find a good compromise without favouring or ruling out any measures. Although he conducted numerical experiments on the benchmark data set and related data and compared them with other methods, no specific research results were obtained (Dahbi, 2020). Huiyu's research involves the implementation of genetic network programming (GNP) and ant colony optimization (ACO) to solve the problem of mining the order rules of business recommendations in time-dependent transaction databases. He believes that an excellent recommendation system should be able to detect customers' preferences in an active and effective way, which requires accurate and timely methods to explore customers' potential needs. Due to the changing nature of customer preference and the difference from the traditional "all first, prune later" method, he extracted interesting time association rules through the GNP method based on meta-heuristic and genetic algorithm. In addition, he used acquired rules to predict future customer needs and used ACO methods to continuously develop online recommendation systems to build useful models. By analysing the customer database of the online supermarket, he has conducted an experimental evaluation of the method in practical application, but the evaluation accuracy is not high (Huiyu, 2019). Gayathiri P proposed a new technique of sensitive rule selection (GS-SRS) based on gravity search to select sensitive rules and hide them to improve the privacy protection of transactional databases. He introduced the GS-SRS technique to select sensitive rules from derived association rules by conditional probability. Sensitive rules contain sensitive information about the transactional database. He has identified sensitive rules for many applications. One application of sensitive rule identification is to protect the privacy of an organization or individual in the following ways. Although the method proposed by him is very confidential, it does not give the actual research method (Gayathiri, 2018). Al-Mamory S argues that large rules cause analysts to spend more time searching for large rules to find interesting rules. One way to solve this problem is to combine one of the association rule visualization methods and the generalization method. His generalization method is attribute oriented inductive algorithm (AOI). The combing AOI is called modified AOI because it removes and changes the steps of the traditional AOI. The carded graph technique is also known as the grouping graph method because it shows the aggregate result rules from AOI. His result is a compression ratio that can make visualizations clearer. His research results provide the ability to test and study rules in depth or to understand and summarise them, but the research process is too cumbersome (Al-Mamory, 2016).

Association rules are used in different data mining applications, including web mining, intrusion detection and bioinformatics. This study mainly discusses the COVID-19 patient diagnosis and treatment data mining algorithm based on association rules. General information on key time intervals during the primary treatment process. The frequency of drug use was counted and association rule algorithm was used to analyse and study the effect of drug treatment. The results could provide reference for rational drug use in COVID-19 patients. In this study, in order to improve the efficiency of data mining in data processing, it is necessary to pre-process these data. Secondly, in the application of this data mining, the main objective is to extract association rules of COVID-19 complications. So its properties for mining should be various diseases. Therefore, it is necessary to classify individual disease types. During the construction of association rules database, the data in the data warehouse is analysed online and the association rules data mining is analysed. The results are stored in the knowledge base for decision support (Jain, 2019; Jain, 2020). For example, the prediction results of the decision tree can be displayed at this level. After the construction of the mining model, the display interface can be mined, and the decision-maker can input the corresponding attribute value and then predict it. This study is helpful to analyse the imaging factors of COVID-19 disease.

2 | DATA MINING FOR DIAGNOSIS AND TREATMENT OF PATIENTS WITH NEW CORONAVIRUS PNEUMONIA

2.1 | Association rules

With the vigorous development of social media and health informatics, there is an urgent need for a powerful tool to maintain a comprehensive analysis of public and personal health information. In particular, it should be able to maximize the discovery of association rules between data items and handle the rapidly growing data scale. FP-Growth algorithm is a remarkable method for learning association rules, which can be used to explore potential relationships in databases that may lack prior knowledge. It has the advantages of low time and space complexity, but it cannot handle the negative association rules necessary for comprehensive mining of health data. ARM is an important topic in data mining. Mining association rules is to find rules of the form X → Y from the rule base where X and Y satisfy certain constraints. Class Association Rules (CAR) is a
special type of association rules suitable for classification problems. The research on ARM and CARM (CAR mining) can be traced back to the early nineties. Since then, many algorithms have been proposed. However, all existing algorithms will encounter inefficiencies when dealing with frequently updated data sets (rules), because any update requires recalculation of the rules, so it takes a long time (Meng, 2019; Ms. M A, 2018). ARM is the process of identifying frequent items and association rules in the market sub-data analysis of a large transaction database set. This leads to the need for SRS to enhance the privacy protection of data transactions. Bacteria is a kind of prokaryote, and most of them reproduce asexually through two divisions. At present, we have good treatments for most infectious diseases caused by bacteria. At present, there is a lack of specific medicines for the treatment of viruses. Most of them can only be cured by the patient. A small number of interferons can inhibit the replication of the virus, but overall there is a lack of treatment (Al-Daher, 2017). Assuming that \( n \) is the number of samples belonging to category \( c \) in the data set \( X \), and the total number of samples in \( X \) is total, then the prior probability of each category is (Zhang, 2017; Samantaray & Singh, 2016):

\[
P(B_i) = \frac{n_i}{\text{total}}; i = 1, 2, ... n
\]  

(1)

For the data set \( x \), the expected information is calculated as (Sumangali, 2016):

\[
F(n_1, n_2, ..., n_m) = -\sum_{j=1}^{m} P(B_j) \log_2 (P(B_j))
\]  

(2)

The entropy obtained by dividing the data set \( X \) by the description attribute \( F \) is:

\[
E(F) = \sum_{x=1}^{m} \frac{n_{1x} + n_{2x} + ... + n_{mx}}{X} (n_{1x}, n_{2x}, ..., n_{mx})
\]  

(3)

Among them:

\[
E(n_{lx}, ..., n_{mx}) = -\sum_{j=1}^{m} \rho_j \log_2 (\rho_j)
\]  

(4)

The information gain when \( E(f) \) divides the data set can be obtained as (Wang, 2017; Rauch, 2019):

\[
\text{Gain}(F) = E(n_1, n_2, ..., n_m) - E(F)
\]  

(5)

If the current data point is null or noisy data, use the average value of \( n \) non-null data points before (after) the current point to replace (Won, 2020).

\[
C_i = \text{Max} \left( \sum_{j=1}^{i-1} C_j + \frac{\sum_{j=1}^{i-1} C_j}{\sum_{j=1}^{i-1}} \right)
\]  

(6)

Among them, \( C_i \) represents the value of the current data point, and \( C_j \) represents the point that is not empty before (after) the current data point (Ma, 2016).

\[
C_i = \left( \sum_{j=N}^{i-1} V_j \times C_j + \frac{\sum_{j=N}^{i-1} V_j + C_j}{\sum_{j=N}^{i-1} V_j} \right) \div \left( \sum_{j=N}^{i-1} V_j - \sum_{j=N}^{i-1} V_j \right)
\]  

(7)

\[
M \text{ the amount of information in data warehouses is often very large, and queries may involve multiple complex join and aggregation operations at the same time (Qiang, 2016; Zhu, 2016).}
\]

\[
T = B \left( \frac{\sum_{j=1}^{i-1} |V_j|}{r} + B \right)
\]  

(8)

As the data set increases, the communication cost between the Mapper interface and the Reducer interface will also increase (Han, 2016; Swetapadma, 2016). Current data mining systems or tools rarely allow users to participate in the mining process. It is an important but unsolved problem to integrate knowledge of related fields into the data mining system (Ma, 2019).
Among them, $T_c$ is the communication cost time. Data mining research has a wide range of application prospects. It can be applied to decision support, and it can also be applied to database management systems. As a decision support tool, data mining can be used to construct data mining in a knowledge database. Semantic query optimization, integrity constraints and inconsistency checking (Singhal, 2016). In the field of statistics and machine learning, there are many data mining systems (Jain, 2020). Some people think that the combination of data warehouse, OLTP, OLAP and data mining technology is a trend in recent database development. Data mining has been widely used in the field of statistics. As a rapidly developing branch of logic programming, logic programming is closely related to data mining (San I, 2016; Necir, 2017).

2.2 | COVID-19

Viral pneumonia is a kind of disease that seriously endangers human health. Prior to the current COVID-19 epidemic, influenza viruses were the common pathogen of viral pneumonia, affecting approximately 3–5 million cases of severe influenza worldwide each year, of which approximately 290–650,000 died; Statistics since 2003 show that SARS and MERS, caused by coronavirus infections, have also caused more than 10,000 deaths, with mortality rates of 10% and 37% respectively, posing a huge economic and social burden around the world. Scientists have put a lot of effort into developing a vaccine to fight viral infections, but vaccines can only prevent infections, not eradicate the virus. Currently, the antiviral drugs used in clinical practice cannot completely block or inhibit the life cycle of the virus, and the virus has strong variability and drug resistance often occurs, so there is no real antiviral specific drug at the present stage. There is also no clear treatment plan for viral pneumonia infected by different pathogens. As ACE2 receptor content is most abundant in lung tissues and small intestinal epithelial tissues in human body, most patients after infection with 2019-NCoV show COVID-19, and some of them are accompanied by digestive tract symptoms. The presence of an overactivated immune response in the body was confirmed in autopsy reports of patients who died of COVID-19. Therefore, how to suppress the inflammatory storm is the key to control the transformation from light and common type to severe and critical type (Pérez-Palacios, 2017).

The structure of Novel Coronavirus under electron microscope is shown in Figure 1.

The information entropy is (Rahmati, 2017; Tzanis, 2017):

$$H(Y) = - \sum_{i=1}^{n} P(Y_i) \log_2 P(Y_i)$$

(10)

Among them, $i$ is the number of possible symbols for the source $Y$ (Chinchuluun, 2017).

$$H(X/Y) = - \sum_{i=1}^{n} \sum_{j=1}^{m} P(X_i,Y_j) \log_2 P(X_i/Y_j)$$

(11)
Let $|T|$ be the sample size of the data set $T$ (Figueiredo, 2016; Ka, 2016):

$$M(T) = - \sum_{i=1}^{k} \left( \frac{freq(C_i,T)}{|T|} \right) \times \log_2 \left( \frac{freq(C_i,T)}{|T|} \right)$$  \hspace{1cm} (12)

The data set $T$ is split according to the attribute $V$, and the expected information calculation formula is (Kasperczuk, 2016):

$$N(T) = - \sum_{i=1}^{n} \left( \frac{|T_i|}{|T|} \right) \times \text{Info}(T_i)$$  \hspace{1cm} (13)

The information gain is:

$$Gain(V) = \text{Info}(T) - \text{Info}(V)$$  \hspace{1cm} (14)

The information gain rate is:

$$Gain\;\text{ratio}(V) = \frac{Gain(V)}{\text{Split Info}(V)}$$  \hspace{1cm} (15)

Among them, $Gain\;\text{ratio}(V)$ is the information gain rate.

### 3 | ASSOCIATION RULE MINING EXPERIMENT OF COVID-19 PATIENT DIAGNOSIS AND TREATMENT DATA

#### 3.1 | Research objects and criteria

General information mainly includes gender, age, underlying disease, contact history, and so forth. Clinical data mainly included first symptoms and signs, Mulbsta score, critical time interval during diagnosis and treatment (including onset to dyspnea, first diagnosis, admission, mechanical ventilation, death, and time from first diagnosis to admission, etc.), laboratory examination, complications and main treatment conditions, and cause of death, and so forth. The frequency of drug use was counted, and association rule algorithm was used to analyse and study the effect of drug treatment. Through the study and analysis, the results obtained can provide reference for rational drug use of COVID-19 patients, reduce the cost of disease treatment and reduce the disease pain of patients.

A retrospective analysis was conducted on 49 cases of COVID-19 deaths diagnosed on January 29, 2020, BBB 0 and March 6, 2020 in our hospital. Inclusion criteria: All the enrolled patients met the diagnostic criteria of confirmed cases in Novel Pneumonia Diagnosis and Treatment Protocol for Coronavirus Infection (Trial Seventh Edition) issued by the National Health Commission on March 3, 2020. Clinical manifestations: fever and/or respiratory symptoms; COVID-19 imaging features: Multiple small patches and interstitial changes were present in the early stage, with obvious extrapulmonary zone; Further development for double lung ground glass shadow, infiltrating shadow, serious cases may appear lung consolidation. Exclusion criteria: excluded successfully treated cases and excluded undeniably diagnosed cases. The study was non-interventional and did not require patients to sign informed consent.

#### 3.2 | Observation indexes

According to the course records, the laboratory examination results on admission ($D1 + 1$), 4 + 1 day ($D4 + 1$), 7 + 1 day ($D7 + 1$) and 14 + 2 days ($D14 + 2$) were recorded, including routine blood, blood gas analysis, PCT, hypersensitive C-reactive protein (HSCRP), myocardial enzymes, liver enzymes, renal function, coagulation indexes, electrolytes and etiological data.

#### 3.3 | Data pre-processing

In this study, the data were obtained from the regional health information platform based on health records. In the final analysis, it belongs to the medical information system, which is closely related to the real world. In order to improve the efficiency of data mining in data processing, it is necessary to pre-process these data. In the data table of personal basic information, in addition to previous history records, there are other fields
unrelated to the research, such as the person who built the file, the date of the file, the medical institution, and so forth. In this application, only the past history records are needed. Therefore, there is no need to pre-process these irrelevant fields, only the past history fields are processed.

Secondly, in the application of this data mining, the main objective is to extract association rules of COVID-19 complications. So its properties for mining should be various diseases. Therefore, it is necessary to classify individual disease types. In the data storage of a person’s disease history, it is often a personal disease history composed of multiple diseases, so it needs to be classified and labelled. For example, in the database, the data in the past history column of “Zhang San” is “hypertension, COVID-19,” indicating that “Zhang San” had suffered from hypertension and COVID-19 before. Therefore, in the information column of “Zhang San,” the column of “Hypertension” is marked as “A”, and the column of “COVID-19” is marked as “B”.

### 3.3.1 Data cleaning

The data cleaning process is to remove the noise design in the original data and some data that is not relevant to the data mining of association rules, and also to process the missing data. Mainly includes missing data processing and error data processing, and complete some data type conversion work.

Due to the large amount of data in electronic health records, which are generated in different places, and the complicated process of generation, it is inevitable that there will be data loss, duplication, and even wrong data. So the data is cleaned.

Fill the void value: Because some attributes in a record may be related to a certain degree of Novel Coronavirus, but its record is empty, so it needs to fill the void value. Filling the void value can be handled by: Ignore the record: When some data rows in the data lack the class label required for their classification, this row can be ignored and the data can be deleted. If the number of tuples missing a class label is very large, this approach will be difficult to work with. Manually fill in missing values: This method compares the cost of time. Especially if the data set is very large. Global constant padding: This method is to populate the records for which some of the attributes are missing with a uniform constant. Although this is an easy way to do it, it is not safe. Mean padding: Calculates the average value of an attribute so that records with missing values in that attribute can be filled in with this average value.

Modify error value: because a lot of data in the medical information system are entered artificially by medical workers, there are some errors in some values, so they need to be modified. Values of data attributes that belong to the canonical standard can be modified by the range standard.

### 3.3.2 Data conversion

For the original data, after data cleaning, cannot be directly used. You also need to convert some of the attributes into the required form. In the original data, the age of an individual is not stored, only the date of birth is stored. Therefore, the age of an individual will be determined according to the date of birth and the date of filing. But the format of these two dates is not the same in some records, some use “year - month - day” format, and some use “year - month - day” format, in order to deal with the convenience, all use “year, month, day” format; An individual's age is then calculated from the difference between the date of birth and the date of filing. The calculated age belongs to continuous attribute, which is not good for the classification of discrete attribute. So you need to discretize. The transformation of age attributes is shown in Table 1.

### 3.4 Construction of association rules database

The system is based on the central database of health records of the regional health information platform. In the health record center data, the repository integrates the management platform of data from different medical information. The overall architecture of the system. The functions of each part are described as follows:

| Age level coding | Interval       |
|------------------|---------------|
| 1                | Under 30      |
| 2                | 30–50 years old|
| 3                | 50–80 years old|
| 4                | >90           |
1. Regional health information platform database: It is a basic database for storing health records, and its information comes from medical institutions at all levels. It includes personal basic information, physical examination information, maternal and child health information, as well as disease control, disease management, and medical services related information content.

2. Data extraction and processing: The health archive database from the regional health information platform database is extracted into the data warehouse according to the subject content of the data warehouse. At the same time, non-standardized data should be processed. This process is called ETL processing. That is, we can write the corresponding handler to process the data according to the need, or we can load and extract the data through the ETL tool of SSAS.

3. Data Warehouse: The health archive data stored for many years is the underlying database of the decision support system. The data is aggregated by topic. Data warehouse is a multi-dimensional database, which is divided into fact table and dimension table. Decision makers can analyse and observe the data in the fact table through dimensions, which is conducive to statistical analysis and enables decision makers to analyse the data from multiple perspectives.

4. Mining application interface: Online analysis and processing of data in the data warehouse and data mining and analysis of association rules. The results are stored in the knowledge base for decision support. For example, the prediction results of the decision tree can be displayed at this level. After the construction of the mining model, the display interface can be mined, and the decision-maker can input the corresponding attribute value and then predict it.

4 | RESULTS AND DISCUSSION

4.1 | Data mining and analysis of association rules

In previous disease analysis, many researchers focused on association rules with high support and confidence. However, in this study, the threshold of support and confidence should not be set too high. The main reason is that the probability of having multiple diseases at the same time is relatively small and the variety of diseases is relatively large. So the frequency of the possible combinations is relatively small. If the threshold value is selected relatively high, some association rules that may exist will be omitted, or even the situation where no association rules can be found in such data may occur. Considering this situation, the minimum support and minimum confidence are set at 0.4% and 27%, respectively. The improved Apriori algorithm based on frequent matrix is used to mine the data. About 1% of the population also suffers from diarrhoea, colds and high-blood pressure. People with diarrhoea and colds were 40.5% more likely to have high-blood pressure and 60.4% more likely to have COVID-19. One percent of the population had retinopathy, colds, and COVID-19, while those with retinopathy, colds, were 60.4% more likely to have COVID-19 and 40.5% more likely to have high-blood pressure. 0.76% of people had both COVID-19, CHD and hypertension, while 46.5% of people with COVID-19 and CHD were likely to have hypertension. Through the analysis of all association rules, it can be concluded that dysentery, hypertension, COVID-19, coronary heart disease, fever, immune deficiency, cold and other diseases have a strong relationship. The relationship between heart disease and psychosis and COVID-19 or high-blood pressure was not strong. Some diseases associated with COVID-19 include immune deficiency, dysentery, hypertension, coronary heart disease, and so forth. Some of the resulting association rules are shown in Table 2. The lung CT before and after admission was shown in Figure 2.

4.2 | Correlation analysis of critical time intervals

The duration from onset to first diagnosis was 0–15 days, with a median course of 4.0 days. The course of disease from first diagnosis to admission was 0–25 days, with a median course of 4.0 days. The course from onset to admission was 2–25 days, with a median course of 7.0 days. The

| TABLE 2 | Part of the obtained association rules |
| Association rules | Attributes | Support (%) | Confidence |
|-------------------|------------|-------------|------------|
| E,H → A           | Heart disease, immune deficiency, high-blood pressure | 1.0 | 40.5 |
| G,C → B           | Dysentery, fever, COVID-19 | 1.0 | 60.4 |
| H,D → A           | Immune function deficiency, coronary heart disease, high-blood pressure | 1.0 | 40.5 |
| B,D → A           | COVID-19, coronary heart disease, high-blood pressure | 0.76 | 46.5 |
| C,B → A           | Fever, COVID-19 high-blood pressure | 0.667 | 388 |
| C,A → B           | Fever, high-blood pressure, COVID-19 | 0.571 | 54.6 |
| B,A → D           | COVID-19, hypertension, coronary heart disease | 0.48 | 49.7 |
**FIGURE 3**  Statistical analysis of disease course

**TABLE 4**  Characteristics of clinical manifestations

| Symptom                  | Number of cases (cases) | Proportion (%) |
|--------------------------|-------------------------|----------------|
| Difficulty breathing     | 33                      | 67.3           |
| Fatigue                  | 25                      | 51.0           |
| cough                    | 23                      | 46.9           |
| Chills/chills            | 19                      | 38.8           |
| Expectoration            | 18                      | 36.7           |
| feel sick and vomit      | 9                       | 18.4           |
| Abdominal pain/diarrhoea | 9                       | 18.4           |
| Muscle ache              | 4                       | 8.2            |
| Tachycardia              | 4                       | 8.2            |
The course from onset to dyspnea was 0–17 days, with a median course of 2.0 days. The course of disease from onset to mechanical ventilation was 0–24 days, with an average course of 9.4 ± 5.9 days. The course of disease from onset to death was 7–49 days, with a median course of 20.0 days. The length of hospital stay ranged from 1 to 41 days, and the mean course of disease was 12.37 ± 8.4 days. The statistics of disease course are shown in Table 3. Statistical analysis of the course of disease was shown in Figure 3.

Clinical manifestations: the primary symptom was fever, and 95.2% of the patients presented fever. The highest body temperature was 40°C, and the average body temperature was 38.3 ± 0.8°C. With dyspnea (66.7%); Followed by fatigue (50.0%), cough (40.5%), chills/chills (40.5%), expectoration (35.7%), and so forth; There were 15 cases (35.7%) with fever, cough and dyspnea. Radiographic records were recorded in all 49 patients with a wide range of pulmonary lesions. The first imaging findings showed that only one case (2.0%) had unilateral pulmonary lesion,
and 48 cases (98.0%) had bilateral lesion. The clinical features were shown in Table 4. The analysis of clinical features was shown in Figure 4. The lung image of the patient is shown in Figure 5.

4.3 | Laboratory examination and analysis

The laboratory test results are shown in Table 5. Check the test analysis as shown in Figure 6. Blood routine of most of the deaths on admission showed leukocytosis and lymphocytopenia; Inflammatory indexes such as procalcitonin and HSCRP were increased. Arterial blood gas analysis suggested hypoxemia; 21 of the deaths had high levels of D-Dimer on admission.

![Graph showing laboratory test results](image)

**FIGURE 6** Analysis of laboratory inspection test results

| Cytokine | Mean | SE |
|----------|------|----|
| IL-1β    | 4.08 | 1.02|
| IL-1ra   | 1478.23 | 917.46|
| IL-2     | 6.44 | 0.84|
| IL-3     | 1.32 | 0.15|
| IL-4     | 0.10 | 3.54|
| IL-5     | 9.98 | 2.79|

**TABLE 6** Plasma cytokine levels in patients with bacterial infection with pneumonia

![SEM images of virus particles](image)

**FIGURE 7** Virus drug screening under electron microscope
In the correlation analysis of plasma cytokine M-CSF level and CT value of respiratory tract samples of COVID-19 patients, Spearman correlation coefficient was \(-0.728\), with \(p\) value less than 0.01. In the correlation analysis between the levels of plasma cytokines IL-10, IFN-A2, IL-13 and IL-17 of COVID-19 patients and the CT values of respiratory tract samples with viral load index, Spearman correlation coefficient was between \(-0.7\) and \(-0.6\), which were \(-0.685\), \(-0.653\), \(-0.636\) and \(-0.608\), respectively, with \(p\) values less than 0.01. The results showed that the plasma cytokines M-CSF, IL-10, IFN-A2, IL-13 and IL-17 in patients with COVID-19 had a strong linear negative correlation with the virus CT value. The lower the viral CT value of COVID-19 patients, the higher the viral load, and the higher the plasma levels of M-CSF, IL-10, IFN-A2, IL-13 and IL-17. Plasma cytokine levels in patients with bacterial pneumonia are shown in Table 6. The drug screening of the virus under electron microscopy is shown in Figure 7. Correlation analysis of M-CSF, IL-10, IFN-A2, IL-13, IL-17 and Novel Coronavirus CT values was shown in Figure 8.

5 | CONCLUSION

Association rules are used in different data mining applications, including Web mining, intrusion detection, and bioinformatics. This study mainly discusses the COVID-19 patient diagnosis and treatment data mining algorithm based on association rules. General data The key time interval during the main diagnosis and treatment process (including onset to dyspnea, first diagnosis, admission, mechanical ventilation, death, and the time from first diagnosis to admission, etc.), the cause of death by laboratory examination, and so forth. The frequency of drug use was counted, and association rule algorithm was used to analyse and study the effect of drug treatment. The results could provide reference for rational drug use in COVID-19 patients. In this study, in order to improve the efficiency of data mining in data processing, it is necessary to pre-process these data. Secondly, in the application of this data mining, the main objective is to extract association rules of COVID-19 complications. So its properties for mining should be various diseases. Therefore, it is necessary to classify individual disease types. During the construction of association rules database, the data in the data warehouse is analysed online and the association rules data mining is analysed. The results are stored in the knowledge base for decision support. For example, the prediction results of the decision tree can be displayed at this level. After the construction of the mining model, the display interface can be mined, and the decision-maker can input the corresponding attribute value and then predict it. This study is helpful to analyse the imaging factors of COVID-19 disease.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Wei Miao ▼ https://orcid.org/0000-0001-7055-4664

REFERENCES

Al-Daher, A. H., & Shkoukani, M. (2017). A proposed dynamic algorithm for association rules mining in big data. Journal of Theoretical and Applied Information Technology, 95(13), 2973–2980.
Al-Mamory, S., & Abdullah, Z. (2016). Combining the attribute oriented induction and graph visualization to enhancement association rules interpretation. Iraqi Journal for Computers and Informatics, 42(1), 10–22.
Chinchuluun, A., Xanthopoulos, P., & Tomaino, V. (2017). Data mining techniques in agricultural and environmental sciences. *International Journal of Agricultural & Environmental Information Systems*, 1(1), 8–12.

Dahbi, A., Jabri, S., & Balouki, Y. (2020). Selecting, sorting and ranking association rules with multiple criteria using dominance relation. *Advances in Mathematics Scientific Journal*, 9(11), 9489–9508.

Figueiredo, M., Esteves, L., & José, N. (2016). A data mining approach to study the impact of the methodology followed in chemistry lab classes on the weight attributed by the students to the lab work on learning and motivation. *Chemistry Education Research and Practice*, 17(1), 156–171.

Gayathiri, P., & Poorna, B. (2018). Gravitational search algorithm for effective selection of sensitive association rules. *Journal of Theoretical and Applied Information Technology*, 96(10), 3047–3060.

Han, B., Wang, Z., & Jin, B. (2016). Anomaly detection algorithm for taxis based on trajectory data mining and online real-time monitoring. *Journal of University of ence and Technology of China*, 46(3), 247–252.

Huiyu, Z. (2019). Kotaro, et al. evolving temporal association rules in recommender system. *Neural Computing & Applications*, 31(7), 2605–2619.

Jain, D. K., Jain, R., Lan, X., Upadhyay, Y., & Thareja, A. (2020). Driver distraction detection using capsule network. *Neural Computing and Applications*, 33(3), 1–14.

Jain, D. K., Jain, R., Upadhyay, Y., Kathuria, A., & Lan, X. (2019). Deep refinement: Capsule network with attention mechanism-based system for text classification. *Neural Computing and Applications*, 32(1), 1839–1856.

Jain, D. K., Mahant, A., Shamsolmoali, P., & Manikandan, R. (2020). Deep neural learning techniques with long short-term memory for gesture recognition. *Neural Computing and Applications*, 32(4), 16073–16089.

Ka Elan, V., Ka Elan, L., & Novovi Buri, M. (2016). A nonparametric data mining approach for risk prediction in car insurance: A case study from the Montenegro market. *Ekonomska Istraivanja*, 29(1), 545–558.

Kasperczuk, A., & Agnieszka, D. A. R. D. Z. I. (2016). Comparative evaluation of the different data mining techniques used for the medical database. *Acta Mechanica Et Automation*, 10(3), 233–238.

Ma, J., Tang, H., & Hu, X. (2016). Identification of causal factors for the Majigou landslide using modern data mining methods. *Landslides*, 14(1), 1–12.

Ma, Z., Wang, X., Jain, D. K., Khan, H., & Wang, Z. (2019). A blockchain-based trusted data management scheme in edge computing. *IEEE Transactions on Industrial Informatics*, 14, 1353–1354.

Meng, X. (2019). Efficient method for finding class association rules in dynamic datasets with record deletion. *Computing Reviews*, 60(6), 262–263.

Ms, M. A., & Dr, R. S. (2018). A novel predictive data mining technique for predicting Slp using association rules and Kmeans clustering (Armkm). *International Journal of Engineering and Technology*, 10(1), 29–32.

Necir, H. (2017). A data mining approach for efficient selection bitmap join index. *International Journal of Data Mining Modelling & Management*, 2(3), 238–251.

Pérez-Palacios, T., Caballero D., Antequera T. optimization of MRI acquisition and texture analysis to predict Physico-chemical parameters of loins by data mining. *Food & Bioprocess Technology*, 10(4), 1–9.

Qi, Y., & Lam, S. N. (2016). The impact of hurricane Katrina on urban growth in Louisiana: An analysis using data mining and simulation approaches. *International Journal of Geographical Information ence*, 30(9–10), 1–21.

Rahmati, O., & Pourghasemi, H. R. (2017). Identification of critical flood prone areas in data-scarce and ungauged regions: A comparison of three data mining models. *Water Resources Management*, 31(5), 15–15.

Rauch, J. (2019). Expert deduction rules in data mining with association rules: A case study. *Knowledge and Information Systems*, 59(1), 167–195.

Samantaray, S. D., & Singh, P. (2016). Extracting association rules in spatial databases of agriculture domain for land use planning. *Journal of the Indian Society of Agricultural Statistics*, 70(2), 167–172.

San, I., At, N., & Yakut, I. (2016). Efficient pailler cryptoprocessor for privacy-preserving data mining. *Security and Communications Networks*, 9(11), 1535–1546.

Singh, S. P., & Vikas, K. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, 35(3), 38–54.

Sumangali, K., & Singaraiah, J. N. (2016). Determining association rules on optimized XML document. *International Journal of Pharmacy and Technology*, 8(4), 26222–26227.

Swetapadma, A., & Yadav, A. (2016). Data-mining-based fault during power swing identification in power transmission system. *let ence Measurement & Technology*, 10(2), 130–139.

Tzanis, G. (2017). Biological and medical big data mining. *International Journal of Knowledge Discovery in Bioinformatics*, 4(1), 42–56.

Wang, B., Chen, D., & Shi, B. (2017). Comprehensive association rules Mining of Health Examination Data with an extended FP-growth method. *Mobile Networks & Applications*, 22(1), 1–8.

Won, E. S., & Kim, S. Y. (2020). An analysis of consumers purchasing patterns for fresh food products using association rules. *Journal of Agriculture & Life Science*, 54(4), 111–122.

Zhang, Z., & Guo, C. (2017). Association rules evaluation by a hybrid multiple criteria decision method. *International Journal of Knowledge & Systems Science*, 2(3), 14–25.

Zhu, F., Kalta, A., & Saif, T. (2016). Parametric analysis of the biomechanical response of head subjected to the primary blast loading – A data mining approach. *Computer Methods in Biomechanics and Biomedical Engineering*, 19(9–12), 1053–1059.

**AUTHOR BIOGRAPHIES**

**Zicheng Shan** is a Researcher from Artificial Intelligence Research Institute, Donghua University. His scientific interests include artificial intelligence, data mining algorithms, economic data models. During the COVID-19 epidemic, he worked with hospitals in Wuhan to apply data science related models and algorithms to medical research.
Wei Miao is a Researcher from Artificial Intelligence Research Institute, Donghua University. Dr. Wei Miao graduated from College of Behavioral and Social Sciences, University of Maryland. His scientific interest lies in the application of data analysis and computer modeling in the field of sociology, he has done many years of research in the fields of credit evaluation and crime rate prediction using data science technology.

How to cite this article: Shan, Z., & Miao, W. (2023). COVID-19 patient diagnosis and treatment data mining algorithm based on association rules. Expert Systems, 40(4), e12814. https://doi.org/10.1111/exsy.12814