Analysis of Pathogens of Pneumonia in children Based on Aprior Algorithm

Mao Xiaojian
Pediatric Department of Guangzhou Children's Hospital, Guangzhou Women and Children's Medical Center, Guangzhou Medical College, Guangzhou, China 510623
Email: xjamao@tom.com

Wang Heyong
Department of E-Business, South China University of Technology, Guangzhou, China 510006
Email: wanghey@scut.edu.cn

Abstract—Objective To explore the relationship between the pneumonia clinical features and the pathogens of pneumonia in children. Methods Based on the clinical data of 6290 cases of pneumonia in children, to study the relationship between the pneumonia clinical features and the pathogens of pneumonia in children by making use of association rules. Results Through software analysis, the different association relationship can be got between the different clinical features of pneumonia in children such as gender, age and region, etc., and the pathogens of pneumonia. For example, between the same pathogen and different sex show the different association relationship. Due to the different association relationships between the pneumonia clinical features and the pathogens of pneumonia in children, so the differences methods in prevention and treatment of Guangzhou area child pneumonia should be adopted according to actual condition, for different clinical features of children take different means, in order to achieve the best effect. Conclusion The association rule can be adopted to analyze the relationship between the pneumonia clinical features and the pathogens of pneumonia in children, and to explore the multi-association relationship between the pneumonia clinical features and the pathogens of pneumonia in children.

Keywords- Pathogen, Pneumonia, Children, Association Rules, Clinical Features

I. INTRODUCTION

Pneumonia is the first cause of death under 5 year's old in the world's children, accounting for about 20% of deaths. In China, the annual total of about 21 million children with pneumonia, it has become leading cause of death under 5 year's old children, a considerable portion of which is caused by pneumococcus. According to WHO statistics, about 1 million children under 5 year's old children died of pneumonia caused by pneumococcus, meningitis and sepsis and other diseases every year in the worldwide[1]. Among the diseases which all can be vaccinated against, pneumococcal disease is the first disease, which caused the death under 5 year's old children. Therefore need to actively strengthen the prevention and treatment of clinical pneumonia in children, to reduce pneumonia in children with severe pneumonia morbidity and mortality.

Different pathogens have different susceptible population. During the epidemic, Pathogen types and categories also vary. And infections can be mixed with a few different pathogens. Pneumonia and other serious respiratory infections are caused by the respiratory syncytial virus (RSV), influenza virus (FV) and adenovirus (ADV) in children, and even lead to death. 36.5% children with RSV infection can cause pneumonia, the mortality rate 1 ‰ -3 ‰; 28% children with influenza virus can cause pneumonia, the mortality rate 3.10%; ADV infections can also cause severe pneumonia in children, in the deaths of children due to severe pneumonia autopsy specimens, ADV infection was detected in 9.14%. The three virus infection can cause some popular spread. But for children, RSV, FV, and pneumonia in children with ADV infection had no clear effects, safe treatments.

For now, antibiotics are the main drug to treat the pneumonia, which is the only effective treatment method. But, with the long-standing abuse of antibiotics, pneumococcal resistance to antibiotics is growing. Currently, the treatment of pneumococcal disease gets more and more difficult, the rising cost of the treatment cycle is getting longer and longer, mainly because of growing bacterial resistance. The pneumococcal vaccine can actually prevent the disease, so, it is important of pneumonia prevention in children. It is necessary to understand the pneumonia gender, age, season and annual distribution of clinical characteristics, it can be found the distribution and the law of pneumococcal disease through the analysis of disease characteristics of different regions pneumonia children. Association rules [2,3] is the most active data mining research methods, it can find the implicit knowledge present in the data. Therefore, this study used association rules and clinical features of pneumonia in children between different types of pathogenic association.

This paper is organized as follows: Section 2 shows the principal of association rules. Section 3 gives the clinical data and calculation the association rules. Experiment results are discussed in section 4. Finally section 5 concludes the paper.

II. APRIOR ALGORITHM

Association analysis focuses on discovering association rules which are interesting and useful hidden relationships that can be found in large data sets. The Apriori algorithm is the original association rule algorithm. The Apriori Algorithm [4] proposed by Agrawal et. al. in 1994, finds frequent items in a given data set using the anti-monotone constraint [5-7]. This algorithm embodies the following.

• Given a data set, the problem of association rule mining is to generate all rules that have support and confidence greater than user-specified minimum support and minimum confidence respectively.

978-1- 61284-704-7/11/$26.00 ©2011IEEE

• Candidate sets having k items can be generated by joining large sets having k-1 items, and deleting those that contain a subset that is not large (where large refers to support above minimum support).

• Frequent sets of items with minimum support form the basis for deriving association rules with minimum confidence. For A=>B to hold with confidence C, C% of the transactions having A must also have B.

The Apriori algorithm from is the basic candidate-generation-and-test algorithm. It is based on the Apriori principle, which says that the item set X' containing item set X is never large if item set X is not large. Based on this principle, the Apriori algorithm generates a set of candidate large item sets whose lengths are $k+1$ from the large k-item sets (for k≥1) and eliminates those candidates, which contain not large subset. Then, for the rest candidates, supports are counted and only those with support over minsup threshold are taken to be large $k+1$ -item sets.

Here's the general idea, use k-item sets to explore $k+1$ -item sets.

1. first find the set of frequent 1-itemsets, L_k :- this is done by scanning the database and accumulating the count for each item and then keeping those that meet the minimum support in a new set called L_k .
2. Then use L_k to find C_{k+1} (the set of candidate 2-itemsets):- This is a two step process that
3. first generates C_{k+1} based on L_k and then
4. secondly prunes C_{k+1} by getting rid of those C_{k+1} item sets using the Apriori principle.
5. find L_{k+1} : we do this by finding the support count for all the item sets in C_{k+1} and getting rid of those that are below the minsup.
6. continue step 2&3 until no new frequent $k+1$ -item set are found.

III. EXPERIMENT USING APRIORI ALGORITHM

1) Case Source

Cases from the Guangzhou Children's Hospital hospitalized with pneumonia in children from 2005 to 2009, a total of 6290 cases, including pneumonia gender, age, region, year, and other aspects of efficacy.

2) The establishment of cases database

The clinical data of the 6290 copies of the original are saved to Excel spreadsheet files, which retain the children's sex, age, season, year, place of residence and other clinical features of pneumonia pathogen information in children, analyze the correlation between the clinical characteristics of pneumonia and pathogen .

3) Calculation using the apriori algorithm

In this study, experiments are applied using Clementine 11.1. Analysis the pneumonia association rules in children according to gender, age, region, year, and pathogen pneumonia. Access association rules between the clinical features of pneumonia and pathogen in children using the minimum support (8%) and minimum confidence (25%).

4) Construction of the network map

Network diagram is a graphics description of correlation Degree on different things. Set the major frequency of symptoms appearing in the network diagrams, which are got using statistical software Clementine 11.1. In the network diagram, if a straight line connecting the two units, two kinds of units have a common appeared probability, the more coarse lines, said the higher co-occurrence probability.

IV. EXPERIMENTAL RESULTS

1) pneumonia pathogen statistic results

Study the distribution of different pathogens, the overall statistics of the data to get pneumonia in children of different pathogens, the results shown in table1 and Figure 1:

Table1: pneumonia pathogen Statistics

Type	Count	Percent
Asthmatic bronchial pneumonia	2095	33.31
Respiratory syncytial pneumonia	1377	21.89
Mycoplasma pneumonia	1221	19.41
Chlamydia pneumonia	503	8
Adenoviral pneumonia	461	7.33
Influenza pneumonia	437	6.95
Klebsiella pneumonia bacteria	137	2.18
Pseudomonas aeruginosa pneumonia	32	0.51
Severe pneumonia	27	0.43

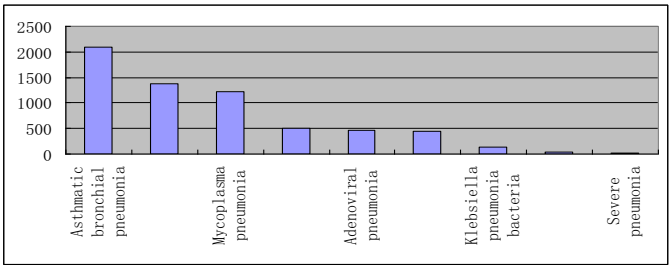

Figure 1 pneumonia pathogen Statistics

In Figure 1 it can be shown that 2095 cases of bronchial pneumonia without pathogens were detected in hospitalized children, accounting for 33.31% of the total number of children; Respiratory syncytial virus pneumonia is 1377 cases, accounting for 21.89% of the total number; Mycoplasma pneumonia is 1221 cases, accounting for 19.41% of the total number; and other pneumonias are a relatively small proportion in children. It can be described that it is more important factors that hospitalized children with pneumonia in asthmatic bronchial pneumonia, respiratory syncytial virus pneumonia, mycoplasma pneumonia.

2) Relationship between Gender and a Variety of Pathogens

In order to study the association relationship between gender and various pathogens in children, through the overall analysis of the hospitalized children data, the results is got shown in Figure 2:

From figure 2, it can be given that male children are the highest correlation asthmatic bronchial pneumonia, and women are associated with the highest degree of mycoplasma pneumonia. In addition, there is different association relationship between children of different gender and pathogens. Male patients in children should enhance the treatment with asthmatic bronchial pneumonia treatment. female patients in children should be strengthened the treatment and prevention with mycoplasma pneumonia, chlamydia pneumonia and respiratory syncytial virus pneumonia treatment.

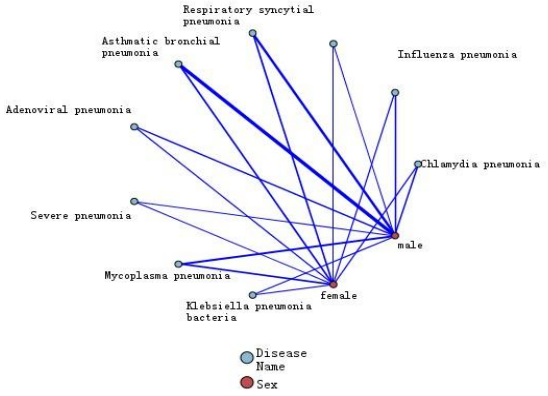

Figure 2 the association relationship between gender and a variety of pathogens

3）***Relationship between Age and a Variety of Pathogens***

To facilitate the study, the actual age of children is classified as follows 0-1 years (infants): 0 years <age <or equal to 1 year; 1-3 years (early childhood): 1 year <age <or equal to 3 years; 3-6 years (preschool): 3 years <age <or equal to 6 years; 6-14 years (school-age): 6 years <age <or equal to 14 years; It gets the results of Figure 3 between the age and the disease name by the degree of support and confidence.

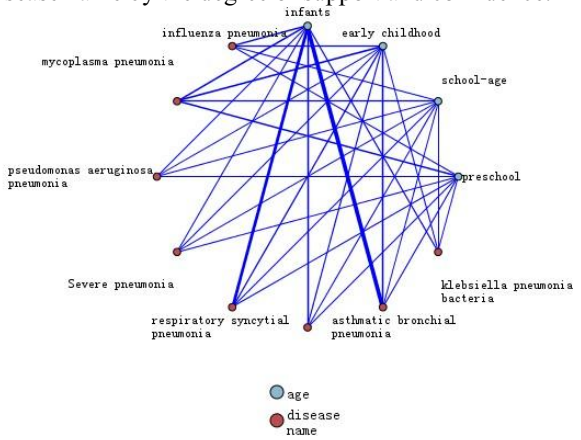

Figure 3 the relationship between age and various pathogens

It can be seen from Figure 3 that with the highest infant children is associated respiratory syncytial pneumonia, followed by asthmatic bronchial pneumonia. Early childhood with mycoplasma pneumonia has the highest correlation. Preschool children with mycoplasma pneumonia have the highest correlation, indicating that preschool children with mycoplasma pneumonia and asthmatic bronchial pneumonia are more worthy of attention. Mostly school-age children with mycoplasma pneumonia and Chlamydia pneumonia, asthmatic bronchial pneumonia and respiratory syncytial virus pneumonia are in the lower proportion of school-age children. Pneumonias in children of different age segments are pathogenic differences.

4） the relationship between various pathogens and year that children got pneumonia

According to the analysis of different years that children got pneumonia and various pathogens, it can be shown the results in figure 4:

In 2005, children with pneumonia, adenoviral pneumonia and mycoplasma pneumonia have strong association rules, which the asthmatic bronchial pneumonia is the most prominent among them, and there is a positive correlation. In 2006, mycoplasma pneumonia, respiratory syncytial virus pneumonia also has a strong association rules. In 2007, 2008 and 2009, it can be seen that the adenoviral pneumonia has been a strong association rules. From above, it can be got that pneumonia pathogens in different years are different.

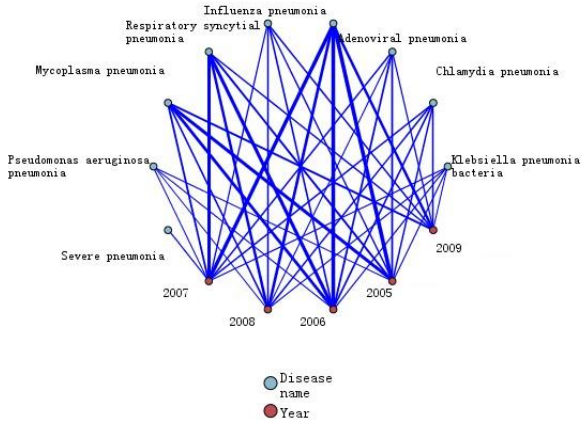

Figure 4 the relationship between various pathogens and year that children got pneumonia

5） the relationship between the region and a variety of pathogenic

For better analysis of various pathogens and the place of residence, to analysis the data to different regions of Guangzhou and the other around the city of Guangzhou, the results is shown in Figure 5:

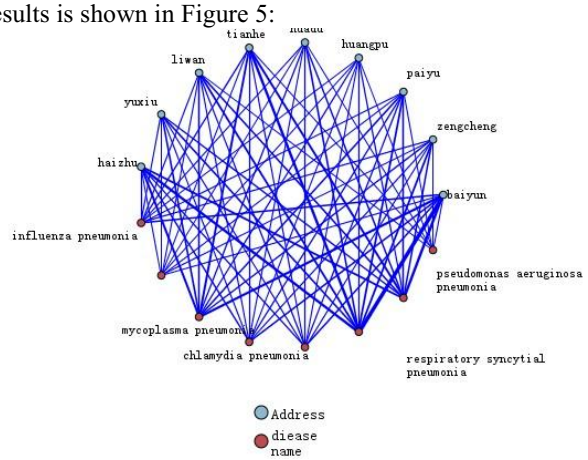

Figure 5 the relationship between region and variety of pathogenic

Baiyun District is the strongest association relationship with the respiratory syncytial pneumonia; for respiratory syncytial virus, it's confidence is more or less in the different regions, basically about 20%, indicating between it and the region there is no obvious correlation; it can be seen from the results that influenza viral pneumonia and adenoviral pneumonia are not correlation with regions; mycoplasma pneumonia is mainly concentrated in the Haizhu District and Baiyun District, Chlamydia pneumonia is concentrated in the Haizhu District. This shows that pneumonia in different pathogens have different associations with different regions.

V. CONCLUSION

Association rules can be found the implicit knowledge in the data, be able to identify multi-dimensional association rules in the data. In order to improve the model accuracy in practical applications, usually with minimum support, minimum confidence to measure the degree of association rules, only the support and confidence are greater than the specified minimum number that the association rules is in line with the requirements of an association rules model.

In this study, analysis the association relationship between pathogenic characteristics and pneumonia on hospitalized children in order to provide a new perspective on the pneumonia in children. The results showed that:

(1) Relationship between Gender and a Variety of Pathogens

Different types of infection, the absolute number of male children with pneumonia were higher than female patients, further it gives the validation that male children are more susceptible to infectious diseases than female children. This suggests that Male children should be focused on prevention and treatment as a target on pneumonia pathogens in Guangzhou.

(2) the relationship between age and various pathogens

The number of pneumonia in children increased with age is decreasing for different types of pathogens. It suggests that as age increases, children's ability to resist pathogen infection gradually is increasing. The peaks pneumonia in children, such as asthmatic bronchial pneumonia, respiratory syncytial virus pneumonia, mycoplasma pneumonia, are all in infancy and early childhood, it is time to focus on strengthening the infancy and early childhood pneumonia in children treatment and prevention. For preschool and school-age children, the main infection is mycoplasma pneumonia, so pneumonia pathogens should be aware of mycoplasma pneumonia infection in the preschool and school-age children.

(3) the relationship between various pathogens and year that children got pneumonia

Several pathogens are different in different years that children got pneumonia. Asthmatic bronchial pneumonia has been a major component of bronchial pneumonia in hospitalized pneumonia children each year, so it should have been in treatment and prevention as a priority.

The number of pneumonia in children, with The respiratory syncytial virus, mycoplasma pneumonia, is different in different years in five years , it suggests that it different pathogen infection prevention and control of clinical pneumonia in children should be according to an annual basis, and should enhance the infection of different pathogens annual monitoring.

(4) the relationship between the region and a variety of pathogenic

Several different distribution of pneumonia in children living in areas marked differences. Children with mycoplasma pneumonia is mainly in the Haizhu District and Baiyun District, Children with chlamydia pneumonia is concentrated in the Haizhu District, children with respiratory syncytial virus, influenza virus and adenovirus viral pneumonia is no obvious correlation with the residence. It suggests that in the prevention of pneumonia, the pathogen should be noted that according to different region.

ACKNOWLEDGMENT

This research was supported by Social Science Foundation for the Youth Scholars of Ministry of Education of China, Grant No. 10YJC630236, supported by the Fundamental Research Funds for the Central Universities, Grant No. 2011ZM0080, supported by Guangdong Medical Research Foundation of Science and Technology, Grant No. 2010472 and Dr start-up fund of Guangzhou Women and Children Medical Center Grant No.201008.

REFERENCES

[1] Rudan I, Boschi PC, Biloglav Z, Mulholland K,Campbell H. Epidemiology and etiology of childhood pneumonia ［J］. Bull World Health Organ, 2008, 86(5): 408-416.

[2] Gong yanbing, Luo zenggang, Gao Sihua, Ni Qing, Yi danhui, Xie Yanming, Wang Yongyan,Cross-sectional Study on the Relationship between Formulas and Syndromes of Type 2 Diabetes Based on Association Rules, JOURNAL OF TRADITIONAL CHINESE MEDICINE, 2010, 51(1) :64- 67.

[3] wu jianhu, He jia, He xianjian, Wu pin, ma xiuqiang, The Exploration of Application of Association Rules in the Medical Domain, JOURNAL OF MATHEMATICAL MEDICINE, 2005（3）:240-243.

[4] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, 1994: 487-499.

[5] Pang-Ning Tan, Vipin Kumar, and Jaideep Srivastava. Selecting the right objective measure for association analysis. Information Systems, 2004: 29(4):293-313.

[6] Yaron Gonen, Nurit Gal-Oz, Ran Yahalom and Ehud Gudes. CAMLS: A Constraint-Based Apriori Algorithm for Mining Long Sequences. Lecture Notes in Computer Science, 2010 5981(2010): 63-77.

[7] Zou liling, Zhao naiqing, Qing guoyou, Qian ji, Shao minhua, Analysis Method to Apply Association Rules in SNP Screening of Disease-related Locus and its Combination CHINESE JOURNAL OF HEALTH STATISTICS, 2009, 26(3) : 226-233.