International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-7, Issue-5C, February 2019

Apriori-based Frequent Symptomset Association Mining in Medical Databases

R. P. Ram Kumar, R. Jayakumar, A. Sankaridevi

Abstract*— Nowadays, healthcare organizations generate large volumes of data. An automatic way of retrieval is necessary when the volume of data is increased. Data mining is becoming very popular and has extensively used in various Healthcare organizations. With the use of various data mining algorithms, it is possible to extract many useful patterns. Data mining applications can highly benefit various parties in Healthcare organization. This paper proposes to enable healthcare organizations by predicting the number of patients affected by certain diseases with respect to their symptoms in medical databases. The pharmacists can use this discovered knowledge and avoid the run out of required drugs, so that the patients can be treated at the right time.

1. INTRODUCTION

Knowledge Discovery in Databases (KDD) is a distinct process consisting of several well-defined steps. Data mining is the core step in KDD, which helps in the discovery of buried, but useful knowledge from massive Healthcare databases. Data mining allows extraction of knowledge from heterogeneous healthcare databases, which in turn eliminates the manual tasks and retrieves the data exactly from electronic records. To carry out the mining process, the medical data can be preprocessed by various techniques like Cleaning, Integration and Transformation. Data preprocessing is one of the major problems when going for the ETL process.

2. PREPROCESSING OF MEDICAL DATA

Cleaning:

The cleaning method removes data inconsistencies by eliminating Duplicate Records, Data entry mistakes and filling missing values for improving the data quality which is highly recommended for mining to keep away from wrong conclusions. Messy data cleaning can be done either manually or by using some tools which are readily available in the market. Smoothing can also be done by techniques such as binning and regression.

Integration:

Once data cleaning has done, the medical data sources are integrated into a homogeneous one, which means the heterogeneity of the medical data is removed while merging data from multiple sources. Data integration is very tricky and various issues are encountered such as matching of the schema, objects when data are received from various sources. The data redundancy may cause if it is derived from another medical entity and it can be detected and solved by correlation analysis.

Transformation:

In data transformation phase the aggregation operations are applied, and a reduced form of clean data is obtained during the reduction phase. In the reduction phase, two main encoding methods like dimensionality reduction and numerosity reduction is used. These reduced data are aggregated in various structures such as tables or Multidimensional model. After these operations, the database is ready to mine. In this paper, a patient database of a particular healthcare organization has been taken as a sample.

3. SYMPTOM-BASED CATEGORIZATION

The medical oriented assessment is extremely specific which requires huge endeavor and determination to spot diseases accurately that shows symptoms. Identifying such symptoms is troublesome which is done by experienced doctors, and they usually categorize diseases based on various diagnosis methodologies. These support doctors to narrow down the cause of diseases that show symptoms and which can be done using facts and experience and later which will be confirmed by conducting various tests.

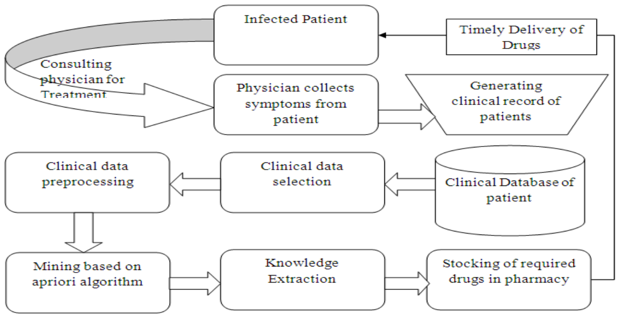
Infected Patient
Timely Delivery of Drugs
Consulting physician for Treatment
Physician collects symptoms from patient
Generating clinical record of patients
Clinical data preprocessing
Clinical data selection
Clinical Database of patient
Mining based on apriori algorithm
Knowledge Extraction
Stocking of required drugs in pharmacy

Figure 1 Architectural diagram of symptom categorization

The signs and symptoms which are self-assessed and reported by patients are crucial to identifying these disorders. Based on the symptoms of patients the physician can diagnose and treat them with adequate medicine.

Revised Version Manuscript Received on 22 February, 2019.

R. P. Ram Kumar, Professor, Department of Computer Science and Engineering, Malla Reddy Engineering College (Autonomous), Secunderabad, Telangana, India. (e-mail: rprkvishnu@gmail.com)

R. Jayakumar, Associate Professor & Head, Department Of Computer Applications, Mahendra Engineering College, Tamil Nadu, India.(E-mail: mymailjsjk@gmail.com)

A. Sankaridevi, Assistant Professor, Department Of Computer Applications, Mahendra Engineering College, Tamil Nadu, India.(E-mail: sankari.dv@gmail.com)

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The diseases such as malaria, Dengue, Zika, Ebola, chickenpox, Chikungunya, Typhoid and jaundice which are sometimes deadly and seasonal infections. Here these diseases and its symptoms are taken as a sample for frequent pattern mining based on patients report that is updated in Healthcare databases.

The entire process of patient symptom mining depicted in the figure is explained over here. The patient who falls ill may describe his or her medical signs and symptom comprehensively when consulting a medical practitioner. At this point, effective Doctor-Patient interaction is challenging and plays a significant role to collect symptoms optimally. The collected symptoms are updated on to the Medical database.

The suitable data for mining can be selectively obtained from the database and which are preprocessed for improving the data quality. Sometimes this preprocessing takes more than half of the total time taken for mining. The frequent pattern mining based on a priori has devised to extract a useful pattern. The outcome of this process is to acquire information about the number of patients affected by the same set of symptoms. Knowledge which is obtained from this mining can be used by the pharmacist to stock their inventory accordingly, which in turn enables them for rapid delivery of drugs and medicine medicines to the needy patient.

Through this mining, the healthcare organizations can take the exact decision to store the required amount of Drugs. In this paper, an apriori association mining algorithm is applied to associate the symptoms of patients, which in turn help the pharmacist to focus on proper medication use.

4. COMMON SYMPTOMS AND RELEVANT DISEASES

The set of a syndrome that is commonly correlated for identifying particular disorders and diseases are listed out.

Symptoms	Disease
High fever and chills, Vomiting, Nausea, Headaches, Body pain, Weakness, Fatigue.	Malaria
Fever, Headache, Vomiting, Nausea, Rash, Eye pain, Joint pain, Muscle aches.	Dengue
Weakness, Abdominal pain, constipation, headaches, Fever	Typhoid
Fatigue, Weight loss, Abdominal pain, Pale stools, Vomiting, Fever, Dark urine.	Jaundice
Fever, Loss of appetite, Muscle aches, Feeling of nausea	Chickenpox

Association Rule Mining:

Association mining is the process of extracting frequent associations among various sets of items in large volumes of data repositories. Some of the currently used association rule mining algorithms are Apriori, FP Growth, Magnum Opus and Closet.

These association rules are widely applied to various areas like Marketing, Inventory management, Telecommunication etc; In this paper, these rules are applied to the Patient database of the healthcare organization. These rules are molded based on two methodologies called support and confidence.

Support represents the percentage of things in a database that contains both item number one and two, Whereas confidence represents the things in a database containing item number one that also contains item number two.

Support (One => Two) = P (One U Two)
Confidence (One => Two) = P (One / Two)

An association rule mining technique is applied to count the total number of patients affected by similar symptoms. The architecture of an association mining of Patient database has shown in the Figure. Here the type of association mining method used is the Apriori algorithm.

Finding Frequent Symptom sets using candidate generation (Apriori Algorithm):

Frequent pattern mining algorithm employs an iterative approach using level-wise search with the help of minimum support count. The Symptom sets are found by scanning the database to collect the count for each symptom that satisfies the minimum support count.

Apriori employs two steps called join and prune during algorithm generation. That is, the database is scanned, and the candidate support count is compared with the minimum support count until the end of the iteration. The result is obtained after finishing full scan of the database. In this algorithm, the candidate generation is denoted by C1 to C n and the resultant set is denoted by R1 to Rn.

The attributes of the patient database are Patient Id, name, address, contact phone number, Set of Symptoms, Prescriptions, Physician detail who attended the patient, payment information and so on. Generally, clinical databases have accumulated with the very large quality of information and medical condition about the patients. But this paper suggests the inclusion of only patient ID and their list of symptoms for sampling. Here the patient database consists of ten samples has been taken for mining. The symptom ('S' represents symptom) sets of the patients are,

S1- Weakness
S2- Abdominal Pain
S3-Headache,
S4-Body Pain
S5-Vomiting
S6-Fatigue
S7- Bleeding
S8- High fever and chills
S9-Chills
S10- Fever

Table 1: Sample Patient Database (D)

P-Id	Symptoms
P1	S1, S3,S4, S5, S6, S10
P2	S2,S4, S5,S6, S9
P3	S2,S3, S8, S10
P4	S1,S5,S7,S8
P5	S1, S3,S4, S5, S6, S10
P6	S3, S7,S8
P7	S1,S3,S9
P8	S1, S3,S4, S5, S6, S10

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Patient database is scanned to count minimum support of each candidate (Symptom). After each scan, the support counts less than four is eliminated.

Iteration 1:

Table 2: Candidate C1

Symptom Set	Support Count
S1	5
S2	2
S3	6
S4	4
S5	5
S6	4
S7	2
S8	3
S9	2
S10	4

Table 3: Resulting Set R1

Symptom Set	Support Count
S1	5
S3	6
S4	4
S5	5
S6	4
S10	4

Iteration 2:

Table 4: Candidate C2

Symptom Set	Support Count
S1,S3	4
S1,S4	3
S1,S5	4
S1, S6	3
S1,S10	3
S3,S4	3
S3,S5	3
S3,S6	3
S3,S10	4
S4, S5	4
S4, S6	4
S4, S10	3
S5,S6	4
S5,S10	3
S6, S10	3

Table 5: Resulting Set R2

Symptom Set	Support Count
S1,S3	4
S1,S5	4
S3,S10	4
S4, S5	4
S4, S6	4
S5,S6	4

Iteration 3:

Table 6: Candidate C3

Symptom Set	Support Count
S1,S3,S5	3
S1,S3,S10	3
S1, S4, S5	3
S1, S5, S6	3
S4, S5, S6	4

Here the minimum support count is taken as 2, and the algorithm is generated. After three iterations the final result is obtained,

Table 7: Resulting Set R3

Symptoms	Support
{S1,S5,S6}	3

Table 7 shows the final result set obtained from the patient database. According to the result, three patients P1, P5 and P8 belonging to the database are having the same set of symptoms, and they may be possibly affected by Malaria as per their symptoms. So, according to the above trend, the healthcare organizations can store the required amount of Vaccines and drugs for Malaria to treat and protect a maximum number of patients.

CONCLUSION

The Apriori Association mining algorithm incorporated in the patient database will identify the frequent set of symptoms and its related diseases which in turn identify the number of patients affected by the particular disorder. The efficiency of these algorithms is also known through the implementation of various steps. This algorithm is simple and fast enough to mine and exactly associate useful patents from the data source which holds symptom sets with a lesser amount of patient entries. If the size of the database grows large, this approach will become a time-consuming process, Because it needs more iterations to generate a candidate with multiple repetitive scans. In future, it is suggested to use various classification algorithms to mine significant information from vast healthcare databases.

REFERENCES

1. Dhanya P Varghese and Tintu P B, “A survey on health data using data mining techniques”, International Research Journal of Engineering and Technology, Vol. 2, No. 7, pp. 713-720, 2015.
2. Ilayaraja & T. Meyyappan, “Mining medical data to identify frequent diseases using apriori algorithm”, In Proceedings International Conference on Pattern Recognition, Informatics and Mobile Engineering, pp. 194-199, 21-22 February, 2013.
3. Sheenal Patel and Hardik Patel, “Survey of data mining techniques used in healthcare domain”, International Journal of Information Sciences and Techniques (IJIST) Vol.6, No.1/2, pp. 53-60, 2016.
4. J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, “Mining infrequent causal associations in electronic health databases”, 11th IEEE International Conference on Data Mining Workshops, pp. 421-428, 2011.

5. R. Karthiyayini and J. Jayaprakash, "Association technique on prediction of chronic diseases using apriori algorithm", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Special Issue. 6, pp. 255-259, 2015.
6. Yanwei Xing, Jie Wang, Zhihong Zhao and Yonghong Gao, "Combination data mining methods with new medical data to predicting outcome of coronary heart disease", International Conference on Convergence Information Technology, pp. 868-872, 2007.
7. Shweta and Dr. Kanwal Garg, "Mining Efficient Association Rules Through Apriori Algorithm Using Attributes and Comparative Analysis of Various Association Rule Algorithm", IJARCSSE, Vol. 3, No. 6, pp. 306-312, 2013.