International Journal of Applied Science and Engineering

Analysis of association among various attributes in medical data of heart patients by using data mining methods

O. Jamsheela*

Department of Computer Science, EMEA College of Arts and Science, Kondotti Kerala, India

ABSTRACT

Large quantities of information about patients have been collected in clinical databases. The data collection is started for the use of doctors but actually these information are a good input for data mining. However, the collected data contains much valuable information when it is treated as a single set of data. From the data many useful knowledge can be extracted. Most of the hospitals are nowadays keep the data of every visited patient. This data can be used to extract new knowledge associated with medical field. The association rule mining, which is one of the popular data mining techniques, can be applied on clinical databases to extract novel and potential knowledge. In this paper, analysis of results obtained from application of association rule mining on a database consisting of details of 10000 patients is presented.

Keywords: Data mining, Analysis of patients' data, Data mining on medical data, Frequent itemset mining, Association rule mining on medical data.

1. INTRODUCTION

Clinical databases became an integral part of modern medicine in the last two decades. Initially such databases were used by the physicians to track treatment history of patients and also used by other departments such as pharmacy, laboratory and front office. With the enormous growth of clinical data, it became a resource for data mining, especially association rule mining, there by generating potential and novel knowledge. Such knowledge includes the associations between different diseases, the association between factors like age, gender, place of living and food habits and different diseases. The knowledge can help in the treatment process, prediction of diseases, measures to check outburst of diseases and so. It can be used to find out the relationships among the data such as the set of diseases which are frequently appeared together, which age group is effected more, some hidden causes of the disease and effectiveness of the treatment etc. The extracted information can be used to predict deceases and can be used to prevent the chance of getting a patient etc. (Gandhi et al., 2020). Unfortunately, very few techniques have been developed and applied to discover this kind of hidden knowledge. The interpretation of the results of mining over medical datasets requires significant domain expertise. The need for strong cooperation between the IT researcher and the medical domain expert(s) is thus higher (Roddick et al., 2003). Clinical data mining is one of the most difficult domains for the KDD field and it is also one of the most challenging field. Only the domain experts can validate and use the generated knowledge. In this study, two popular data mining techniques viz. frequent itemset mining and association rule mining are implemented and applied on a large clinical database to extract and analyze the knowledge generated.

OPEN ACCESS

Received: September 13, 2020

Accepted: January 21, 2021

Corresponding Author:
O. Jamsheela
ojamshi@gmail.com

Publisher:
Chaoyang University of Technology
ISSN: 1727-2394 (Print)
ISSN: 1727-7841 (Online)

2. RELATED WORKS

Earlier, the application of data mining was limited to very few fields like e-business, marketing and retail. Recently, because of the effectiveness of data mining, the techniques are widely used in other sectors including medical sector. Numerous research papers have been published related to application of data mining in health care industry. A brief account of related prominent works are presented below:

Soni et al. (2011) provide a survey of data mining techniques on medical research in heart disease prediction. The authors states that the decision tree is the best method though in most of the cases bayesian classifiers also perform well. The performance of KNN, neural networks and classification based on clustering are reported to be suboptimal. The authors did not consider association rule mining in their work. A new method has been developed by Ilayaraja and Meyyappan (2015) to predict the patients under risk based on some selected symptoms and also to find the risk level of each patient. Ischemia is a restriction in blood supply to tissues, causing a shortage of oxygen and glucose needed for cellular metabolism (to keep tissue alive). When blood flow is completely blocked to the heart, ischemia can lead to a heart attack (Cohen, 2005). To support medical decision-making, a group of researchers highlighted the feasible utilization of matrix-based feature representation and CNN-based feature extraction for incomplete clinical data mining (Wang, 2020). This will have an impact in the medical field to as. Another work is proposed by Gandhi et al. (2020), the authors have developed a specialist framework to support the doctors and authorities to take decision at the time of operation and diagnosing the patients based on the symptoms. An association rule mining algorithm is applied and a rule-based classification model is created for the detection of ischemic beats in long duration ECGs in a research paper by Exarchos et al. (2006). The authors suggest that by using the method the beats can be classified as ischemic or non-ischemic. Ordonez et al. (2000) suggested a method to make association rules more useful for medical data. The authors used a new method to solve the problem of mapping complex medical data to items. The authors claim that most of the proposed improvements are simple but useful. Koh and Tan (2011) presented a paper with an extensive study of data mining applications in health care. The authors states that the data mining techniques, when applied in health care, produce useful information to all parties involved in health care industry. Bellazzi and Zupan (2008) published a review paper on prediction of heart disease by using data mining.

3. MOTIVATION AND CHALLENGES

The World Health Organization has estimated that 12 million deaths occur worldwide every year due to the heart diseases. It is the chief reason of deaths in numerous developing countries (Soni et al., 2011). In this context, we have chosen a medical database relating to heart diseases and applied recent frequent pattern mining and association rule mining algorithms to extract the rules.

4. THE DATASET

The data is collected from the cardiology department of a reputed hospital in Kerala. Details of 10000 patients have been included in this study. The age, gender, place, first consultation date and the diagnosed diseases are the main attributes. The data set contains more than 200 various

Table 1. Highest frequent itemsets out of 10000 transactions

Slno	Itemsets	Frequency count
1	male patient	frequency: 6840
2	female patient	frequency: 3152
3	heart attack	frequency: 5679
4	calicut	frequency: 4934
5	Hypertension	frequency: 2839
6	2011	frequency: 2925
7	2010	frequency: 2835
8	2012	frequency: 2574
9	malappuram	frequency: 2396
10	diabetes mellitus	frequency: 2473
11	heart attack, calicut	frequency: 2804
12	heart attack, male patient	frequency: 4439
13	calicut, male patient	frequency: 3300
14	male patient, 2010	frequency: 1992
15	male patient, 2011	frequency: 1986
16	heart attack, diabetes mellitus	frequency: 1989
17	heart attack, hypertension	frequency: 1884
18	heart attack, calicut, male patient	frequency: 2134

diseases. Same kind of diseases have been grouped together and formed 70 different groups to get more accurate rules. The dataset have been processed to remove the unnecessary data and to make it as a suitable data set for mining purpose. Due to ethical issues most of the hospitals keep the treatment history of patients in a highly confidential manner. Availability of clinical data for research is limited and hence mining of clinical data is a potential area for data mining researchers.

5. FREQUENT ITEMSET MINING ON CLINICAL DATA

The frequent itemset mining find out the items that frequently appears together in a dataset. By applying this method on a clinical dataset, the most frequent items or itemsets can be extracted. Some of the extracted results are included in Table 1. Table 2 shows the year wise details of patients and Table 3 contains the number of patients from 4 major cities.

Table 2. Year-wise details of patients

year	male	female	total
2010	1992	843	2835
2011	1986	939	2925
2012	1743	830	2573
2013	1127	539	1666
Total	6848	3151	9999

6. ASSOCIATION RULE MINING ON CLINICAL DATA

Association rule mining is used to extract the relationship among the items in database. By applying association rule mining method on a clinical database, the correlations of different attributes can be found out such as which diseases are co-occurring together, in which age group the disease is more affected etc.

6.1 Analysis of the Extracted Rules

The database contains the information of every visited patient in the cardiac department. Heart attack is the 3rd most frequent item in Table 1. Therefore most of the item sets with highest support value include heart attack. Hence in this study we have concentrated on the variations of the confidence value other attributes are associated with heart attack. The important association rules mined from the database are shown in Table 5 and 6. The details are presented below. The uncommon medical terms are explained with a short description.

6.1.1 Left Ventricular Dysfunction

Left ventricular dysfunction (LVD) is the result of many cardiac disorders causing a mechanical alteration of cardiac performance (Iliceto, 1997). LVD with subsequent congestive heart failure (CHF) constitutes the final common pathway for a host of cardiac disorders. Coronary artery narrowing or ischaemic heart disease is the dominant cause of heart failure and is often associated with acute or prior myocardial infarction (Armstrong, 2000). The left ventricle loses its ability to relax normally because the muscle has become stiff.

Table 4 displays the frequency of LVD. The Rules 19, 29 and 25 in Table 6 and Rules 1, 2 and 12 in Table 5 show the relationship of heart attack and 'left ventricular dysfunction'(LVD). Many past studies consider the severe LVD as the important cause of heart failure. However, the table shows that it is not the main reason of heart attack. Only 8% of the total heart attack patients have severe LVD and 16% of total heart attack victims have a normal LV function. But it is also noted that the confidence of severe LVD and mild LVD to heart attack is 1, which shows that all visited patients with severe LVD or mild LVD have suffered a heart attack. Therefore we can conclude that the patients with LVD should take the precautions against heart attack because they have a higher risk of heart attack.

6.1.2 Valve Disease

All the valve diseases such as trivial MR, mild AR, mild to moderate AR, mitral stenosis, Bicuspid aortic valve etc. have been grouped together into a single class. Rules 18, 20, 21, 22, 24, 26, 27 in Table 6 and Rule 14 in Table 5 show the associations of the valve diseases with heart attack. Only 8% of the total heart attack victims have a valve problem but 50% of the patients with a valve problem have suffered a heart attack. Hence valve problem is not very serious as LVD.

LVD and valve problem do not have a close relationship. Only 15% of the severe LVD patients have a valve problem and 16% of the people with normal LV function have a valve problem. Out of total valve problem patients 7%

Table 3. Place and details of patients

place	total	male	female	2010	2011	2012	2013
kozhikode	4946	3313	1633	1513	1461	1196	776
malappuram	2397	1699	698	666	724	631	376
wayanad	937	615	321	254	268	263	152
kannur	953	656	297	232	260	257	204
total	9219	6284	2949	2665	2713	2347	1508

Table 4. LV function with frequency

item	frequency
Normal LV function	1177
Sever LVD	464
Mild LVD	107

patients have a severe LVD. But hypertension (HTN) and valve disease are closely related. 21% of total valve problem patients have hypertension also.

6.1.3 Stent

A stent is used to remove the block of the heart valves. Stent is a metal or plastic tube designed to be inserted into a vessel or passageway to keep it open. Stents are inserted into narrowed coronary arteries to help keep them open after balloon angioplasty. The stent then allows the normal flow of blood and oxygen to the heart (Stent, 2011). The placement of such a device is called stenting or stent. Rules 23, 28 and 30 in Table 6 and Rule 11 in Table 5 show an interesting relationship of stent with female patients. The majority of the heart patients are males after the age of 27. Therefore, it is quiet natural that the total number of patients treated with stent is male patients. The support of Rule 11 is 362 and the confidence is 0.89. However, the confidence of Rule 23 is 0.11 only. The rules reveals that compared with the male patients, very few female patients have used stent.

6.1.4 Angina

Angina is chest pain or discomfort caused when heart muscle doesn't get enough oxygen-rich blood. It may feel like pressure or squeezing in the chest. However, angina is not a disease. It is a symptom of an underlying heart problem, usually coronary heart disease (Angina, 2016). From the Rule-9 in Table 5, we can find that angina is a strong symptom of heart attack but not all angina lead to heart attack. Here the confidence is 0.94 only. Exertional angina is the angina during exercise. Rule-8 in Table 5 reveals that patients having exertional angina have more chance of heart attack than the patients having normal angina because the confidence is 0.97.

The Rules 3 and 4 in Table 5 state that all the male patients in 2010 and 2011 with exertional angina had heart attack but the confidence of Rule 7 in Table 5 is only 0.98 which reveals that not all male patients with exertional angina reported in 2012 had heart attack. Rule-5 in Table 5

Table 5. Selected association rules mined from clinical database

	antecedent→ consequent[Support;Confidence]
1	Mild LVD → heart attack [support = 107 ; confidence = 1]
2	Sever LVD → heart attack [support = 464 ; confidence = 1]
3	Exertional angina: male patients: 2011 → Heart attack [support = 262 ; confidence = 1]
4	Exertional angina: male patients: 2010 → Heart attack [support = 297; confidence = 1]
5	Hypertension:Exertional angina → Heart attack [support = 444 ; confidence = 1]
6	Diabetes Mellitus:Exertional angina → Heart attack [support = 442 ; confidence = 1]
7	Exertional angina: male patients: 2012 → Heart attack [support = 213 ; confidence = 0.98]
8	Exertional angina → Heart attack [support = 1146 ; confidence =0.971]
9	angina → Heart attack [support = 472 ; confidence = 0.944]
10	Diabetes Mellitus:Normal LV function → Heart attack [support = 327; confidence = 0.93]
11	stent → male patients [support = 362 ; confidence =0.887]
12	Normal LV function → heart attack [support = 946 ; confidence = 0.804]
13	Diabetes:Hypertension → heart attack [support = 915 ; confidence = 0.78]
14	Valve Problem → Heart attack [support = 496 ; confidence = 0.5]
15	Diabetes → Hypertension [support = 1178 ; confidence = 0.48]

reveals that all the visited 444 patients with hypertension and exertional angina have suffered a heart attack. By analyzing Rule-6 in Table 5, the possibility to have a heart attack of a diabetic patient with exertional angina is very high because the confidence value is 1.

The remaining Rules in Table 5 and 6 are included to show the correlation between diabetes, hypertension and heart attack. Hypertension and diabetes appeared together in 1178 patients' record. Total number of diabetes patients and hypertension patients are 2473 and 2839 respectively. Rules 15 and 16 show that diabetes is closely associated with hypertension because the confidence values are above 0.40. But Rule 13 reveals that the combination of both diabetes and hypertension is not very risky as exertional angina.

The above rules unveil many hidden information and it is not possible to find these knowledge directly from the dataset. The information is extracted by using the association rule mining and frequent itemsets mining methods. Therefore, these two tools of data mining can be efficiently applied on clinical dataset to find many hidden knowledge. With lowest minimum support and confidence values, more than four lakh rules are mined. Very few rules are presented and discussed. Many more relevant information can be extracted if a detailed study carries out on those mined rules.

7. CONCLUSION

In this paper the frequent itemset mining methods have been used to extract frequent items from a medical dataset and subsequently applied association rule mining. The study carried out with a data of 10000 patients associated with cardiology department demonstrated the strength of association rule mining in producing novel and potential knowledge. However, the study is limited in its scope but paves way for intensive and in-depth research in the area of application of association rule mining in medical data.

Table 6. Selected association rules mined from clinical database

	antecedent → consequent[Support;Confidence]
16	Hypertension → Diabetes [support = 1178 ; confidence = 0.41]
17	Diabetes → heart attack:Hypertension [support = 915 ; confidence = 0.37]
18	Valve Problem → HTN [support = 216 ; confidence = 0.218]
19	heart attack → Normal LV function [support = 946 ; confidence = 0.167]
20	Normal LV function → Valve Problem [support = 195 ; confidence = 0.166]
21	Sever LVD → Valve Problem [support = 74 ; confidence = 0.159]
22	Valve Problem → Heart attack, Normal LV [support = 116 ; confidence = 0.117]
23	stent → Female Patients [support = 46 ; confidence = 0.113]
24	Heart attack → Valve Problem [support = 496 ; confidence = 0.087]
25	heart attack → Sever LVD [support =464 ; confidence = 0.082]
26	Valve Problem → Sever LVD [support = 74 ; confidence =0.074]
27	Valve Problem → Heart attack, Sever LVD [support = 69 ; confidence = 0.07]
28	male patients → stent [support = 362 ; confidence = 0.053]
29	heart attack → Mild LVD [support = 107 ; confidence = 0.019]
30	Female Patients → sent [support =46 ; confidence = 0.015]

REFERENCES

Angina, 2016. URL http://www.heart.org/HEARTORG/Conditions/HeartAttack/SymptomsDiagnosisofHeartAttack/Angina-Chest-Pain UCM 450308 Article.jsp #V6BGPtR96t9, 18.

Armstrong, P.W. 2000. Left ventricular dysfunction: causes, natural history, and hopes for reversal. Heart 84, Nr. Suppl 1, S. i15–i17.

Bellazzi, R., Zupan, B. 2008. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics 77, Nr. 2, S. 81–97.

Cohen, E. 2005. Method and system for diagnosis of cardiac diseases utilizing neural networks. U.S. Patent Application 11/718,840.

Exarchos, T.P., Papaloukas, C., Fotiadis, D.I., Michalis, L.K. 2006. An association rule mining based methodology for automated detection of ischemic ECG beats. IEEE Transactions on Biomedical Engineering 53, Nr. 8, S. 1531–1540.

Gandhi, Y., Singh, A., Jani, R. 2020. Big data and its application in healthcare and medical field. Data Science and Intelligent Applications. Springer, S. 161–166.

Ilayaraja, M., Meyyappan, T. 2015. Efficient data mining method to predict the risk of heart diseases through frequent itemsets. Procedia Computer Science 70, S. 586–592.

Iliceto, S. 1997. Left ventricular dysfunction: which role for calcium antagonists? European heart journal, 18, Nr. suppl A, S. 87–91.

Koh, H.C., Tan, G. 2011. Data mining applications in healthcare. Journal of Healthcare Information Management, 19, Nr. 2, S. 65.

Ordonez, C., Santana, C.A., Braal, L. De. 2000. Discovering interesting association rules in medical data. ACM SIGMOD workshop on research issues in data mining and knowledge discovery Citeseer (Veranst.), S. 78–85.

Roddick, J.F., Fule, P., Graco, W.J. 2003. Exploratory medical knowledge discovery: Experiences and issues. ACM SIGKDD Explorations Newsletter 5, Nr. 1, S. 94–99.

Soni, J., Ansari, U., Sharma, D., Soni, S. 2011. Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17, Nr. 8, S. 43–48.

Stent, 2011. URL http://medicine.academic.ru/7766. -2011