Relevance of Data Mining for Presumption of Anarchy

Dr.Aneeshkumar A.S.

Assistant Professor, Department of Computer Science, Alpha Arts and Science College, Chennai, India
aneeshkumar.alpha@gmail.com

Abstract— Association in Data Mining is used to identify frequent item sets and its correlation. The broad application of association mining in research, market analysis and disease predictions are proved. But sometimes the formation of association rule is very peculiar work in medical diagnosis. Early detection of disease is always useful to reduce the complexity and after effects. It also reduces time and money expenses. In this paper, I propose a case study for disease prediction model.

Keywords— *Data Mining; Association Rule; heart Disease.*

1. Introduction

Data mining or knowledge discovery is a process of identifying valid and useful information from a large multi dimensional dataset. So it combines data bases, information retrieval, statistics, machine learning and algorithms. Association mining is one of the descriptive methodologies to generate associative rules for the frequent sets among the transactional data. The efficient discovery of such rules has been a major focus in the data mining research community [1].

Association rules are defined as $X = \{X1, X2, \dots, Xm\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq X$. Each transaction is associated with a unique identifier *TID*. A transaction T is said to contain Q, a set of items in X, If $\subseteq T$.

An association rule is an implication of the form $"X \Rightarrow Y"$, where $Q \subseteq X$, $R \subseteq X$, *and* $Q \cap R = \emptyset$. The rule $Q \Rightarrow R$ has a support s in the transaction set D if s% of the transaction in D contains $Q \cup R$. That is known as, the support of the rule is the probability that Q and Y hold together among all the possible presented cases i.e., $P(Q \cup R)$. It is said that the rule $Q \Rightarrow R$ holds in the transaction set D with confidence c, if $c\%$ of transactions in D that contain Q and also contain R. That is, the confidence of the rule is the conditional probability that the consequent Y is true under the condition of the antecedent X.

The problem of discovering all association rules from a set of transaction D which consists of the rules that have a *support* and *confidence* greater than given thresholds. These rules are called *strong rules*.

2. Case Study: Heart Disease

Heart disease is one of the major causes of mortality in the world. Each year about 500,000 people die from heart attacks. An additional 500,000 undergo coronary artery bypass surgery or balloon angioplasty for advanced heart disease. Early recognition and treatment of heart disease is vital to prevent some of these events [2]. In case of any disease, early intervention is better than providing treatment after identifying the disease. According to surveillance of mortality and cardiovascular disease(CVD) related morbidity in industrial settings, almost 2.6 million Indians are predicted to die due to coronary heart disease(CHD), which constitutes 54.1% of all CVD deaths in India by 2020. Additionally, CHD in Indians has been shown to occur prematurely, that is, at least a decade earlier than their counterparts in developed countries. Heart diseases may see in all categories of peoples without any age categories in India. It is due to the lack of physical activity, modern improper food habits, smoking and alcohol consumption. Sadly we can say these are the part of Indian youth’s life style.

The ultimate goal of risk factor prevention, detection, and control is to prevent acute events. In India, most cases felt the first occurrence of heart disease is without prior knowledge of heart disease. Sudden death and out of hospital deaths, due to heart disease are also without the prior evidence. So in this paper, we try to identify the cases of heart disease by using associative symptoms. Risk factors are widely classified into two categories, which are major and contributing. Major risk factors are here used to prove the risk of heart diseases. Contributing risk factor are those that doctors think can lead to an increased risk of heart disease, but their exact role has not been defined. The more risk factors you have, the more likely to have heart disease development. Some risk factors can be changed. But as a whole, it can be considered as evidence. If in case of earlier identification, most of the risk factors can be changed or treated and also controlled as possible through lifestyle changes and medicines. Based on Table I data, we are considering the minimum support as 2.

i.e. min_sup_count= 2/10 =20%

Then the maximum support count in C_1, which is identified as 8 for Cholesterol and other support counts are C_2 = 7, C_3= 6, C_4=4, C_5=3,C_6=2. Strong association rule satisfy both minimum support and minimum confidence. So the confidence is,

$$Confidence(A => B) = P(B|A) = \frac{Support\_count(A \cup B)}{Support\_count(A)}$$

The conditional probability is expressed in terms of itemset support count, where support count $(A \cup B)$ is the number of transactions that contains the itemset $A \cup B$, and support count *(A)* is the number of transactions contains the itemset *A*.

Confidence(H D=>Cho) = 9/10 = 90%

Confidence(H D=>Chol^Dia) = 7/10 = 70%

Confidence(HD=>Cho^P I) = 7/10 = 70%

Confidence(H D=>Cho^Dia^P I) = 6/10 = 60%

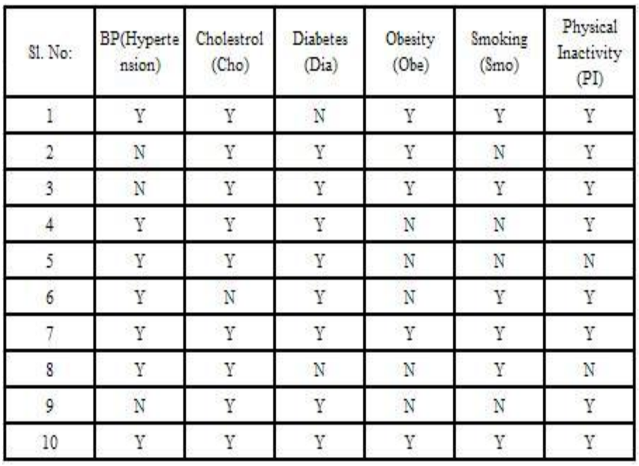
Confidence(H D=>Cho^Dia^Obe^P I) = 4/10 = 40%

Confidence(H D=>Cho^Dia^Obe^Smo^PI) = 3/10 = 30%

Confidence(H D=>BP^Cho^Dia^Obe^Smo^P I) =2/10= 20%

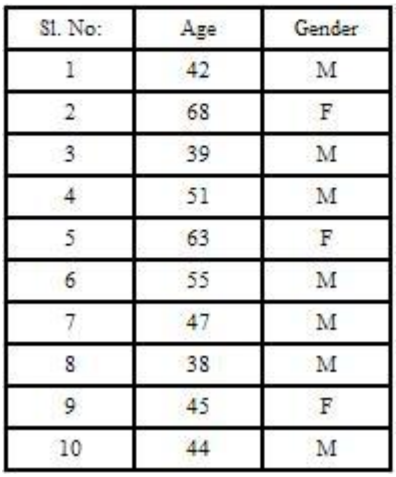
If the minimum confidence threshold is 70%, then the first three rules are generating strong association.

Table 1: Symptom of Heart disease



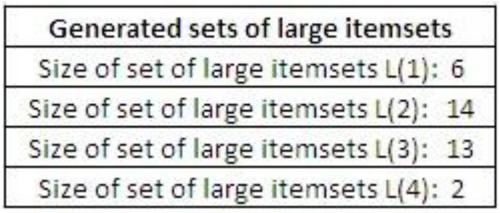
Sl. No:	BP(Hypertension)	Cholestrol (Cho)	Diabetes (Dia)	Obesity (Obe)	Smoking (Smo)	Physical Inactivity (PI)
1	Y	Y	N	Y	Y	Y
2	N	Y	Y	Y	N	Y
3	N	Y	Y	Y	Y	Y
4	Y	Y	Y	N	N	Y
5	Y	Y	Y	N	N	N
6	Y	N	Y	N	Y	Y
7	Y	Y	Y	Y	Y	Y
8	Y	Y	N	N	Y	N
9	N	Y	Y	N	N	Y
10	Y	Y	Y	Y	Y	Y

Table 2: Age of the patients



Sl. No:	Age	Gender
1	42	M
2	68	F
3	39	M
4	51	M
5	63	F
6	55	M
7	47	M
8	38	M
9	45	F
10	44	M

Table 3: large itemset generated in each list

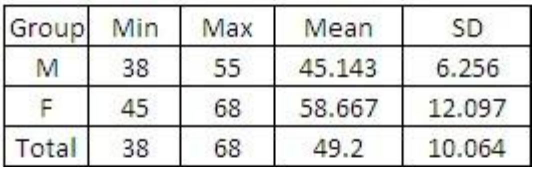


Generated sets of large itemsets
Size of set of large itemsets L(1): 6
Size of set of large itemsets L(2): 14
Size of set of large itemsets L(3): 13
Size of set of large itemsets L(4): 2

3. Result and Discussion

Apriori algorithm is used here to determine the association of symptoms and generate association rules. Ten sample data of heart disease (HD) with its associative symptoms are shown in Table 1. The maximum confidence of this dataset is identified as 0.9. Table 2 is used to show the age of ten patients.

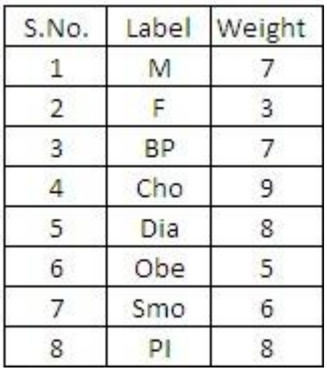
Table 4: Mean and Standard deviation



Group	Min	Max	Mean	SD
M	38	55	45.143	6.256
F	45	68	58.667	12.097
Total	38	68	49.2	10.064

Table 3 shows the largest itemset which generated from 4 lists by the apriori algorithm. L(2) produced 14 itemsets as largest and L(4) produced smallest as 2. Table 4 shows the average age group for heart disease. According to male group the minimum age for occurring heart disease is 38 and for female it is 45. The average age group for male is 45 and female is 58.7. In recent years, the average of both male and female is reducing because of modern food habits and life style. Table 5 represents the weight of each attribute, given as the predicting factor of heart disease.

Table 5: Weight of the attributes



S.No.	Label	Weight
1	M	7
2	F	3
3	BP	7
4	Cho	9
5	Dia	8
6	Obe	5
7	Smo	6
8	PI	8

From table 2, it can be concluded that the age and sex are major dependent factors for heart disease prediction. So the condition of rule based multidimensional association is written as,

```
IF (Confidence≥ 70)
        IF (Gender==Male)&&(age>35)
        THEN ("The Risk is High");
        ELSE-IF (Gender==Male) && (age<35)
                THEN ("The Risk is Normal");
        ELSE_IF (Gender==Female)&&(Age>45)
                THEN ("The Risk is High");
        ELSE ("The Risk is Normal");
END-IF;
```

4. Conclusion

This paper introduced a case study for associative mining and which may useful for experts to predict heart diseases and make consciousness in surveillance of mortality due to heart disease or its related issues. The relation of age, sex and suspected symptoms like hyper tension or blood pressure, cholesterol, diabetes, obesity, smoking and physical inactivity for the prediction of heart disease is proven in this work. In India, the average time of more than one hour is needed for a patient to reach a hospital in any emergency situation. So Indian medical practitioners are demonstrating awareness of evidence based treatment and therefore this application will give more support to such society for their future work.

References

[1] Maria-Luiza Antony , Osmar R. Zaiane, Text document categorization by term association.

[2] Lawrence s. cohen, Heart disease symptoms, chapter 9, Yale University School of Medicine Heart Book.

[3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Published by Elsevier, second edition – 2006.

[4] Xiaoxin Yin, Jiawei Han, CPAR: Classification based on Predictive Association Rules.

[5] Shantakumar B. Patil, Y.S.Kumaraswamy, Intelligent and effective heart attack Prediction System using Data mining and Artificial neural network, European journal of Scientific research, Volume 31 Issue 4, 2009, pp.642-656.

[6] K. Rajeswari, Dr. V.Vaithiyanathan, Dr. P. Amirtharaj, Prediction of risk score for heart disease in India using machine Intelligence, International Conference on Information and Network Technology, Volume 4, 2011, IACSIT Press, Singapore.

[7] Sunita Soni, O.P.Vyas, Using Associative classifiers for Predictive Analysis in Healt care Mining, International Journal of Computer Applications, Volume 4, Issue 5, July 2010.

[8] Aneeshkumar A.S and Jothi Venkateswaran C, Estimating the Surveillance of Liver Disorder using Classification Algorithms, International Journal of Computer Application, Volume 57, Issue 6, pp.39-42, November 2012.

[9] Latha Parthiban and R. Subramanian, Intelligent Heart Disease Prediction using CANFIS and Genetic Algorithm, International Journal of Biological and Life Sciences, Vol.3, issue 3, 2007.

[10] Coronary heart disease- causes,incidence, risk factors and symptoms, http :// www. ncbi.nlm.nih. gov / pubmed health /PMH0004449/.

[11] Aneeshkumar A.S and Jothi Venkateswaran C, An Integrated Approach for Predicting the Contribution of Sovereign Dynamics, Engineering Sciences International Research Journal, Volume 1, Issue 1, pp. 24-27, February 2013.

[12] Heart disease burden in the next two years, http://www.medicalnewstoday.com / articles / 105302.php.

[13] Causes of increase in the number of heart patients in India, http://health.indiatimes.com/ articleshow/479575.cms.

[14] Heart disease symptoms, http://heart-disease-symptoms.com/.

[15] India's no.1 killer: Heart disease, http://indiatoday.intoday.in/story/ India's+no.1+killer:+ Heart=disease/1/92422.html.

[16] Heart Disease Risk Factors, From Texas Heart institute, http://chinese-school.netfirms. com/heart-disease-causes.html.

Dr.Aneeshkumar A.S. is presently working as an Assistant Professor in the Department of Computer Science at Alpha Arts and Science College, Porur, Chennai. He completed his Ph.D. in Computer Science from University of Madras, Chennai. He finished M.Phil in Computer Science from Vinayaka Mission University. He holds a Master Degree in Information Technology from University of Madras and a Master of Computer Applications degree from Bharathiar University, Coimbatore. In addition to that he completed separate Master degree in Psychology, and in Criminology and Criminal Justice Administration. He presented numerous papers at various National and International conferences and has a handful of publications in SCI, Scopus, National and International level Journals with good impact factor.

Dr.A.K. is the Chief Editor for three International Journals. He is the Editorial Board Member and Indian Reviewer for various SCI and Scopus Indexed Journals. He is having teaching experience of more than nine years and seven years of Research activities. He organized various International and National Conferences, Seminars, Symposiums, Guest Lectures for the Research Scholars and Faculty Members. He invented numerous algorithmic models for complex analyses and acting as an Executive Member, Resource Person and Advisory Member for various International Organizations and Programs. He is expertise in the field of Data Mining, Soft Computing, Software Engineering, Fuzzy Logic, Cloud Computing, Swarm Intelligence and Artificial Intelligence.