# Early Heart Disease Prediction using Frequent Pattern Mining Techniques

M.Revathy Meenal
*Anna Adarsh College for Women, Chennai,Tamilnadu,India..*

***Abstract*** — The successful application of data mining in highly visible fields like e-business, marketing and retail has led to its application in other industries and sectors. Among these sectors just discovering is healthcare. The Healthcare industry is generally "information rich", but unfortunately not all the data are mined which is required for discovering hidden frequent patterns & effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced data mining modeling techniques can help remedy this situation.Data mining is a process which finds useful patterns from large amount of data. Data items are frequent in itemset is to be organized in multilevel and multi dimensional way. Data mining is the process of discovering interesting knowledge such as Patterns and Associations. The process of looking for patterns to document is called pattern mining. Pattern mining is a data mining method that involves finding existing patterns in the data. Mining frequent patterns is probably one of the most important concepts in data mining. Graph transformation method is used for mining of patterns in frequent itemset. An itemset is closed if none of its immediate supersets has the same support as the itemset. Frequent itemsets are so important . This paper intends to use data mining Classification Modeling Techniques, namely, Decision Trees, Naïve Bayes and Neural Network, along with weighted association Apriori algorithm in Heart Disease Prediction.

***Keywords*** — *Data Mining; Graph; Frequent Itemsets; Patterns; Association Rule*

## 1. Introduction

Data mining is the process of finding previously unknown Patterns and trends in databases and using that information to build predictive models. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases.

The World Health Statistics 2012 report enlightens the fact that one in three adults worldwide has raised blood pressure – a condition that causes around half of all deaths from stroke and heart disease. Heart disease, also known as cardiovascular disease (CVD), encloses a number of conditions that influence the heart – not just heart attacks. Heart disease was the major cause of casualties in the different countries including India. Heart disease kills one person every 34 seconds in the United States. Coronary heart disease, Cardiomyopathy and Cardiovascular disease are some categories of heart diseases. The term "cardiovascular disease" includes a wide range of conditions that affect the heart and the blood vessels and the manner in which blood is pumped and circulated through the body. Diagnosis is complicated and important task that needs to be executed accurately and efficiently. The diagnosis is often made, based on doctor's experience & knowledge. This leads to unwanted results & excessive medical costs of treatments provided to patients. Therefore, an automatic medical diagnosis system would be exceedingly beneficial. Pattern mining is a data mining method that involves finding existing patterns in the data. Mining frequent patterns is probably one of the most important concepts in data mining.

## 2. Methodology

This paper exhibits the analysis of various data mining techniques which can be helpful for medical analysts or practitioners for accurate heart disease diagnosis. Due to resource constraints and the nature of the paper itself, the main methodology used for this paper was to use data mining Classification Modeling Techniques.

### *2.1 Collecting the data*

This process is concerned with the collection of data from different sources and locations.The current methods used to collect data are:

- Internal Data: data are usually collected from existing databases, data warehouses, and OLAP. Actualb transactions recorded by individuals are the richest source of information, and at the same time, the most challenging to be useful.
- External Data: data items can be collected from web graphics. In addition to data shared within a company.

### *2.1 Association Rule*

Association and correlation is usually to find frequent item set findings among large data sets. shopping behavior analysis. Association Rule algorithms need to be able to

generate rules with confidence values less than one.
Types of association rule

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

### *2.2 Graph Transformation Method*

It is a technique of creating a new graph out of an original graph algorithmically. It has numerous applications, ranging from software engineering to layout algorithms and picture generation. Graph transformations can be used as a computation abstraction. The basic idea is that the state of a computation can be represented as a graph, further steps in that computation can then be represented as transformation rules on that graph. Rules consist of an original graph, which is to be matched to a subgraph in the complete state, and a replacing graph, which will replace the matched subgraph. Formally, a graph rewriting system usually consists of a set of graph rewrite rules of the form **L→R**, with L being called pattern graph (or left-hand side) and **R** being called replacement graph (or right-hand side of the rule). A graph rewrite rule is applied to the host graph by searching for an occurrence of the pattern graph and by replacing the found occurrence by an instance of the replacement graph.

### *2.3 Frequent itemset mining for graphs*

A graph is a quintuple G= {V, E,}, where V is the set of vertices, E is the set of edges. A graph Gs = {Vs, Es} is said to be subgraph isomorphic to G, which is denoted by Gs∈ G, if there exists a 1–1
mapping f : Vs -> Vsuch that, v ∈ Vs
Further, we say Gs occurs in G if Gs -> G. Let the database D contain a collection of graphs G, then, the support of a subgraph Gs
in D is the number of occurrences of Gs
in D. A graph may be transformed into an edge list L of G = {V, E, } in order to allow
for applying itemset mining algorithms. An edge list is defined as L = (vi, vj , ek | vi,vj)
Within this framework, a graph mining problem can be viewed as a frequent
itemset mining problem such that; let L be the set of itemset. Let the transaction database D be a multiset of subsets of L.

Table1:frequent itemset

| | | | | |
|---|---|---|---|---|
| C1 | S1 | | S3 | |
| C2 | | S2 | | |
| C3 | | | | S4 |
| C4 | | | | S4 |
| C5 | | S2 | S3 | |
| C6 | | S2 | S3 | |
| C7 | S1 | S2 | S3 | S4 |
| C8 | S1 | | S3 | |
| C9 | S1 | S2 | S3 | |
| C10 | S1 | S2 | S3 | |

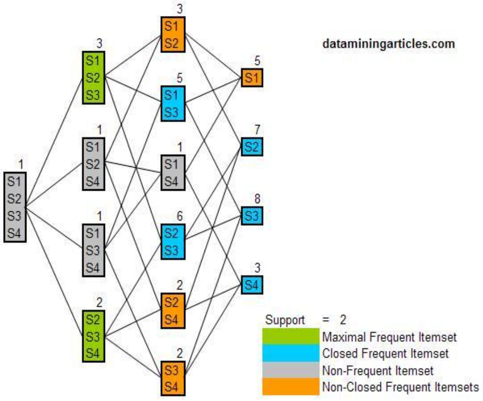

Fig.1:Maximal and Closed frequent itemsets

## 3. Maximal Frequent Itemsets

Setting up a support percentage for an itermset, solved only a part of the problem. We know how frequent an itemset should be to become worth considering it. But the toughest part is still unsolved. In order to find a frequent itemset we have to go through all the sub-itemsets which themselves are frequent due to the Closure Property. So we unavoidably generate an exponential number of subpatterns that we might not really need. Let's go back to our table 1 and say that we want to find all the frequent itemsets that have a support of 30%. We take advantage of our very small dataset and observe that (S1, S2, S3) is present in 3 contracts: C7, C9, C10 - that means the itemset is present in 30% of contracts and hence it is frequent. And the only one superset is (S1, S2, S3, S4) which is not frequent. But what about all the subsets? There are 6 subsets that are obviously present in at least 30% of the contracts so they are frequent. But we have to go through all of them to get to the maximum number of items that form a frequent itemset. This procedure is very time consuming because the search space is huge. The presence of a frequent itemset of length k implies the presence of $2^{K-2}$ additional frequent itemsets. we could consider only the frequent itemset that has the maximum number of items bypassing all the sub-itemsets. This is how the Maximal Frequent Itemset was invented. The definition says that an itemset is maximal frequent if none of its immediate supersets is frequent. In our example (S1, S2, S3) is maximal frequent because the only one superset is not frequent.

### *2.3.1 Closed Frequent Itemsets*

The only one downside of a maximal frequent itemset is that, even though we know that all the sub-itemsets are frequent, we don't know the actual support of those sub-itemsets. And we'll see how important this is when we'll try to find the association rules within the itemsets. An itemset is closed if none of its immediate supersets has the same support as the itemset.

### *2.3.2 Maximal and Closed frequent itemsets*

(Discovering Frequent Closed Itemsets for Association Rules by Nicolas Pasquier, Yves Bastide, Rafik Taouil, and Lotfi Lakhal ) where they proposed to mine only closed frequent itemsets using an algorithm called A-CLOSE. In the following years a lot of other algorithms have been invented (CHARM, CLOSET, etc) that improved the performances of the initial algorithm. Let see some examples of Closed and Maximal Frequent Itemsets. As you see in the graph, all individual items S1, S2, S3, S4 are frequent itemsets because their support in greater than 2. But only 3 of them are closed because S1 has the superset (S1, S3) having the same support. So it contradicts the definition. The itemsets (S1, S2, S3) and (S2, S3, S4) are frequent because they are present in at least 2 of the contracts, and they are maximal as well because their frequent superset - (S1, S2, S3, S4) is the only one superset and it's not frequent.

## 4. Conclusion

Data mining has importance regarding finding the atterns, discovery of knowledge. To conclude, we have to keep in mind the following important concepts: A frequent item set is one that occurs in at least a user-specific percentage of the database. That percentage is called support. An itemset is closed if none of its immediate supersets has the same support as the itemset. An itemset is maximal frequent if none of its immediate supersets is frequent.

## References

[1] Divya Bhatnagar, NeeruAdlakha and A.S Saxena, “Mining Frequent itemsets without candidate generation using optical Neural Network”.

[2] Jiawei Han, Hongcheng, Dongxin, Xifengyan, ”Frequent Pattern mining : Current Status and future directions”

[3] Thashmee Karunaratne, “Is Frequent Pattern mining useful in building Predictive models”

[4] Thashmee Karunaratne and Henrik Bostrom “Use of frequent itemset mining for learning from graphs – What is gained and What is lost”

[5] Jiawei Han and Micheline Kamber (2006), Data Mining Concepts and Techniques, published by Morgan Kauffman,2nd ed.

[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm ; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[7] Erickson, T., Smith, D., Kellogg, W., Laff, M., Richards, J and Bradner, E. Socially translucent conversations: Social proxies, persistent conversation, and the design of “Babble.”Proc. ACM CHI (1999), 72–79.

[8] Hollan, J., Hutchins, E. and Kirsh, D. Distributed cognition: Toward a new foundation for human computer interaction research. ACM TOCHI, 7(2),(2000), 174

[9] 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, ISSN 1450-216X Vol.31 No.4, 2009.

[10] Statlog database: http://archive.ics.uci .edu/ml/machinelearning-databases/statlog/heart