Assessment of Risk Factors of Coronary Heart Disease Based On Weighted Association Rule Mining

Sathya Rajendran, Gomathy Ramesh
Bannari Amman Institute Of Technology,Sathyamangalam
sathurajendran@gmail.com

Abstract

One of the major reason for disability in adults an death in the developed countries is Coronary heart disease (CHD). Although significant progress has been made in the diagnosis and treatment of CHD, further investigation is still needed. The objective of this study was to develop a data-mining system for the assessment of heart event-related risk factors targeting in the reduction of CHD events. The risk factors investigated were: 1) before the event, and 2) after the event. The events investigated were: myocardial infarction (MI), percutaneous coronary intervention (PCI), and coronary artery bypass graft surgery (CABG). Data-mining analysis was carried out using WAR algorithm. It is anticipated that data mining could help in the identification of high and low risk subgroups of subjects, a decisive factor for the selection of therapy, i.e., medical or surgical. However, further investigation with larger datasets is still needed. One of the challenging problems in the weighted association rules mining is to assign weights to items. For practice, self-assigned weights technique is more useful. In this paper, we proposed a self-assigned weights method to discover positive and negative association rules, instead of assigning the weights by users. To avoid mining misleading and uninteresting rules, a new type parameter, called sawinterest, is proposed to eliminate the redundant rules. The rational results are presented.

Index Terms—Coronary Heart Disease (CHD), data mining, WAR.

1. Introduction

Coronary Heart Disease (CHD) is the single most common cause of death in Europe, responsible for nearly two million deaths a year. Advances in the field of medicine over the past few decades enabled the identification of risk factors that may contribute toward the development of CHD. However, this knowledge has not yet helped in the significant reduction of CHD incidence. There are several factors that contribute to the development of a coronary heart event. These risk factors may be classified into two categories, not modifiable and modifiable. The first category includes factors that cannot be altered by intervention such as age, gender, operations, family, history, and genetic attributes. Modifiable risk factors are those for which either treatment is available or in which alternations in behavior can reduce the proportion of the population exposed. Established, modifiable risk factors for CHD currently include smoking, hypertension, diabetes, cholesterol, high-density lipoprotein, low-density lipoprotein, triglycerides. The objective of this study was to develop a data mining system based on weighted association rule mining for the assessment of CHD related risk factors targeting in the reduction of CHD events. Data-mining analysis was carried out using weighted association rule mining using five different splitting criteria for extracting rules based on the aforementioned risk factors. Data mining facilitates data exploration using data analysis methods with sophisticated algorithms in order to discover unknown patterns. Such algorithms include association rule mining that have been used extensively in medicine. Several studies have been carried out that investigated the usefulness of weighted association rule mining in CHD-related problems. In this study, we investigate how data mining based on weighted association rule can help for the evaluation of the risk of CHD. The aim is to identify the most important risk factors based on the classification rules to be extracted. These rules will enable the better management of the patient targeting in the reduction of events, as well as, reduction of the cost of therapy, due to the expected restriction of interventions in necessary cases only.

2. Weighted association rules overview

Association rules mining was primarily proposed for market basket to understand consumer purchasing patterns in retailing industry. It proposed the support-confidence measurement framework and reduced association rule mining to discover frequent itemsets. The following year, a mining algorithm, Apriori, was proposed [9]. Much effort has been dedicated to the classical (boolean) association rule mining since then. A lot of algorithms have been proposed to extract the rules more efficiently. WARM generalizes the traditional model to the case where items have weights. Some researchers introduced weighted support and weighted confidence of association rules based on the costs assigned to items or transactions. The
definition broke the downward closure property. As a result, the weighted algorithm became more complicated. However, another problem came following the traditional algorithms: the weights need to be assigned by users, but users may not know how to confirm the weights correctly. Those algorithms follow the classical measurement framework. But if the minimum support and minimum confidence are given, the results are the same. All the classical algorithms produce positive associations between items existing in transactions. What about association of the type: “one medical record contains disease A but not disease B”. Recently, mining association rules has received some attention and is proved to be useful. Unfortunately, that algorithm was not easy to mine negative rules.

3. Data Collection, Cleaning, and Coding

Data from 1500 consecutive CHD subjects were collected between the years 2003–2006 and 2009 (300 subjects each year) according to a prespecified protocol, under the supervision of the participating cardiologist (Dr. J. Moutiris, second author of this paper) at the Department of Cardiology, at the Paphos General Hospital in Cyprus. Subjects had at least one of the following criteria on enrollment, history of MI, or percutaneous coronary intervention (PCI), or coronary artery bypass graft surgery (CABG). Data for each subject were collected as given in Table I: 1) risk factors before the event, a) nonmodifiable—age, sex, and family history (FH); 2) modifiable—smoking before the event (SMBEF), history of hypertension (HxHTN), and history of diabetes (HxDM); and 2) risk factors after the event, modifiable—smoking after the event (SMAFT), systolic blood pressure (SBP) in mmHg, diastolic blood pressure (DBP) in mmHg, total cholesterol (TC) in mg/dL, high-density lipoprotein (HDL) in mg/dL, low-density lipoprotein (LDL) in mg/dL, triglycercides (TG) in mg/dL and glucose (GLU) in mg/dL. To clean the data, the fields were identified, duplications were extracted, missing values were filled, and the data were coded. After data cleaning, the number of cases was reduced, mainly due to unavailability of biochemical results.

The problem of mining association rules that satisfy some minimum weighted support and weighted confidence can be decomposed into three subproblems:
1. Rank an item and assign its weight.
2. Find significant itemsets whose weighted supports meeting the given threshold.
3. Build rules within the itemsets found in Step 2.
In this section, we provide a method for solving them step by step.

3.1 Ranking Items with HITS

Let \( I = \{I_1, I_2, ..., I_m\} \) be a set of items and let \( T = \{T_1, T_2, ..., T_n\} \) be a set of transactions in database. Clearly, \( D \) is equivalent to the bipartite graph \( G = \{D, I, E\} \), where \( E = \{(T , I_i): I_i \in T, T \in D, I_i \in I\} \)

**Example 1:** Consider the database shown. It can be equivalently represented as a bipartite graph. The database contains medical records. The items \{Nausea,Lumbar,Urine,Micturition,Urethra,UrineyBladder,Nephritis\} are different records.

The graph representation of the transaction database is inspiring. It gives us the idea of applying link-based ranking models to the evaluation of transactions. In this bipartite graph, the support of an item \( I_i \) is proportional to its degree, which shows again that the classical support does not consider the difference between transactions. Intuitively, an important transaction, which is highly weighted, should contain many important items; at the same time, an important item should be contained by many important transactions. The reinforcing relationship of transactions and items is just like the relationship between hubs and authorities in the HITS model.

Regarding the relationship between the hubs and the authorities, we can apply HITS to this bipartite graph. The following equations are used to each iteration:

\[
\text{auth}(I) = \sum_{T \in T} \text{hub}(T) \cdot \text{hub}(I) - \sum_{T \in T} \text{auth}(I)
\]

when the HITS model eventually converges, the auth weights of all items are obtained. These weights represent the potential of transactions to contain high-value items. An item contained by few transactions may still be a good auth if all transactions are top ranked. Conversely, the item contained by many ordinary transactions may have a low hub weight.

3.2. Weighted Metric

After the iteration process, we obtain every \( \text{auth}(I) \). The auth of the item can reflect the importance of the item. For this reason, we pick up the maximal auth to normalize all auths. The normalized \( \text{auth}(I) \) is the weight of the \( i \) denoted as \( w_i \).

\[
W = \{w_1,w_2,...,w_m\}
\]

\[
\text{authMax} = \max \{\text{auth}(I_i) | I_i \in I\}
\]

\[
\text{wi} = \frac{\text{auth}(I_i)}{\text{authMax}}
\]
The items are always arranged by their weights from the largest one to the smallest one in an item set.

**Definition 1.** The weighted support of an itemset \( X \) is defined as
\[
sawsup(X) = sup(X) \times w(X)
\]
(4)
\[
sup(X) = \text{count}(X)/|T|
\]
(5)
\[
w(X) = \max\{w_1, w_2, \ldots, w_p\}
\]
(6)
Where \( w(X) \) is a weight of \( X \), and \( \text{count}(T) \) is the number of transactions in \( T \) containing \( X \). Actually, \( p \) is the items’ count of the itemset \( X \). An itemset is called weighted-frequent (large) if its support is larger than a user-specified value (also called minimum weighted support (minsawsup)). For example, if \( sawsup(X) \geq \text{minsawsup} \), the itemset \( X \) is weighted-frequent.

**Definition 2.** The sawsupport of an association rule \( X \rightarrow Y \) is defined as
\[
sawsup(X \rightarrow Y) = sawsup(X \cup Y) / sawsup(X)
\]
(7)
and the weighted confidence is
\[
sawconf(X \rightarrow Y) = sawsup(X \rightarrow Y) / sawsup(X)
\]
(8)
Basically, saw-support measures how significantly \( X \) and \( Y \) appear together; saw-confidence measures how strong the rule is. If its weighted support \( sawsup(X \rightarrow Y) \) and weighted confidence \( sawconf(X \rightarrow Y) \) meet minimum weighted support \( \text{minsawsup} \) and minimum weighted confidence \( \text{minsawconf} \), \( X \rightarrow Y \) is a strong rule. The task of mining association rules is to discover the strong rules.

**Definition 3.** The interest metric between itemset \( X \) and \( Y \) is defined as
\[
sawinterest(X,Y) = sawsup(X \cup Y) / (sawsup(X) \times sawsup(Y))
\]
(9)
If the \( \text{sawinterest-1} < \text{mini} \), \( X \) and \( Y \) don’t influence each other obviously. The rule between \( X \) and \( Y \) isn’t interesting. If \( \text{sawinterest-1} \geq \text{mini} \), the rule between them may be interesting.

**Definition 4.** The positive and negative rules between item set \( X \) and \( Y \) are defined as \( X \rightarrow Y \) is positive rule and it has three corresponding negative rules: \( X \rightarrow \neg Y, \neg X \rightarrow Y, \neg X \rightarrow \neg Y \). Mining negative association rules raises a lot of critical issues. The number of infrequent itemsets increases by exponential. For instance, if the number of item is 100 and the number of frequent itemsets is 210, the number of the infrequent itemsets is 2100-210. Therefore, the support and confident of the negative association rules is difficult and disinteresting directly. But we can deduce the negative association rules from the corresponding positive itemsets’ support in the frequent itemsets.

### 3.3. Generate Rules

The third step is more time-consuming than the classical algorithm. The classical algorithm (Apriori-ap-entes) generates rules with the theorem that if \( \text{confidence}(X \rightarrow Y, X) < \text{minsconf}, \text{confidence}(X' \rightarrow Y - X') < \text{minsconf} \), \( X' \subseteq X \) is satisfied, \( \text{sup}(Y)/\text{sup}(X) = \text{support}(Y)/\text{support}(X) \). Without weights, we can easily find the result. Because \( \text{support}(X') > \text{support}(X) \). But when we assign the weight, the result can’t be got. Indeed, when \( X' \) is weighted frequent itemset, \( X \) may not be. We don’t know the exact relationship between \( sawsup(X) \) and \( sawsupport(X') \). When we get one rule, we should check the weighted support and weighted confidence.

**Proof.** Let \( X \) be an itemset, and the sets \( X' \) are the subset of \( X \). \( X \) and \( X' \) have the same first item. \( sawsup(X) \leq sawsup(X') \). The items are always rearranged by descending in itemsets. \( w(X) = \{w_1, w_2, \ldots, w_p\} = \max\{w_1, w_2, \ldots, w_p\} \), so \( w(X) = w_1.w(X') = w_1.X' \) has the same first item with \( X \).

\[
sawsup(X) = supsup(X) \times w(X) = supsup(X) \times w_1.
\]
\[
sawsup(X') = supsup(X') \times w(X') = supsup(X') \times w_1
\]
\[
sup(X) \leq sup(X'), X' \text{ is subitem of } X.
\]

Hence, \( sawsup(X) \leq sawsup(X') \). If \( X' \) don’t satisfy \( sawsup(X') \geq \text{minsawsup} \), we can prove that \( X \) also don’t satisfy \( sawsup(X') \geq \text{minsawsup} \). In the function new-genrules, we can check the itemset \( X \) whether it is the frequent weighted itemset. If the result is true, then we check \( sawinterest \) and \( sawconf \); if the result is false, we delete the item. That will reduce the cost of the running time of the algorithm.

The method generates the association rules layer by layer, the number of the layer corresponds with the number of the rule consequent. Firstly, we try the rules that the number of the rule consequent is one. If the rules satisfy the limit, we use them to produce new candidate. For example, if \{acd\} => \{a\} and \{acd\} => \{b\} are the rules satisfied the limit, we can get the new candidate \{ad\} => \{bc\}. If there is a node that can’t meet the given threshold, we can cut the subgraph formed by the node.

**Proposed algorithm:**

1. Initialize \( \text{hub}(t) \) to 1 for each transaction \( t \) in \( T \)
2. for(l=0; l<\text{num_it}; l++) do begin
3. \( \text{hub}(t)=0 \) for each transaction \( t \) in \( T \)
4. for all item \( i \), I do begin
5. \( \text{auth}(I)=\Sigma \text{hub}(t) \)
6. \( \text{hub}^{\prime}(t)=\text{auth}(I) \) for each transaction obtain \( I \)
7. \( \text{hub}(t)=\text{hub}^{\prime}(t) \) for each transaction in \( T \)
8. authMax = \max\{\text{auth}(I_i) | I_i \in 1\}, W(I_i) = \text{auth}(I_i) / \text{authMax} \\
9. WARDM generate frequent item sets \( L \)
10. for each \( f_k \) of \( K \)-frequent item sets, \( K \geq 2 \)
    do
11. for each \( h_1 \in H_1 \) do
12. if \( \text{sawsup}(h_1) \geq \text{minsawsup} \) then
13. if \( |\text{sawinterest}-1| > \text{mini} \) then
14. if \( \text{sawconf}(f_k - h_1) \geq \text{minsawconf} \) then
15. \( \text{PARS} = \text{PARS} \cup \{f_k - h_1 \} \)
16. \( \text{sawconf}(\neg(f_k - h_1)) \geq \text{minsawconf} \) then
17. \( \text{NARS} = \text{NARS} \cup \{\neg(f_k - h_1) \} \)
18. \( \text{sawconf}(\neg(h_1)) \geq \text{minsawconf} \) then
19. \( \text{PARS} = \text{PARS} \cup \{\neg(h_1) \} \)
20. \( \text{NARS} = \text{NARS} \cup \{\neg(h_1) \} \)
21. else return
22. Call new-genrules \( (f_k, H_1) \)

4. Conclusion

In this paper, we have presented a novel framework in association rule mining. First, the HITS model and algorithm are used to derive the weights of items from a database with only boolean attributes. Second, we assign the weights to the items. Third, we use those weights to mine positive and negative association rules under the \( \text{sawsup} \) and \( \text{sawconf} \) framework. Experimental results show that the algorithm is efficient to delete many uninteresting rules and mine the negative rules also. By comparing our findings with other studies: 1) the rules extracted facilitated the grouping of risk factors into high and low risk factors, and 2) the rules extracted are associated with an event risk, however, this needs further investigation.

5. References