J Med Syst (2013) 37:9896
DOI 10.1007/s10916-012-9896-1

ORIGINAL PAPER

Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction

Dong Gyu Lee · Kwang Sun Ryu · Mohamed Bashir · Jang-Whan Bae · Keun Ho Ryu

Received: 15 June 2012 / Accepted: 4 September 2012

Abstract The knowledge discovery has been widely applied to mine significant knowledge from medical data. Nevertheless, previous studies have produced large numbers of imprecise patterns. To reduce the number of imprecise patterns, we need an approach that can discover interesting patterns that connote causality between antecedent and consequence in a pattern. In this paper, we propose association rule mining method that can discover interesting patterns that include medical knowledge in Korean acute myocardial infarction registry that consists of 1,247 young adults collected by 51 participating hospitals since 2005. Proposed method can remove imprecise patterns and discover target patterns that include associations between blood factors and disease history. The association that blood factors affect to disease history is defined as target pattern. In our experiments, the interestingness of a target pattern is evaluated in terms of statistical measures such as lift, leverage, and conviction. We discover medical knowledge that glucose, smoking, triglyceride total cholesterol, and creatinine are associated with diabetes and hypertension in Korean young adults with acute myocardial infarction.

Keywords Knowledge discovery · Association rule mining · Medical database · Acute myocardial infarction · Diabetes · Hypertension

Introduction

The knowledge discovery in databases (KDD) has been studied to mine interesting knowledge in medical community [1–8]. One of popular methods for KDD is association rule mining (ARM) that can mine frequent itemsets and generate association rules. ARM has been widely applied to analyze medical dataset. In particular, the knowledge discovered from medical dataset has significant meanings and provides useful information for the disease management and prevention. Our study is to mine interesting knowledge through ARM to acute myocardial infarction (AMI).

AMI is one of the diseases that impact middle-aged and elder people. Most knowledge of AMI has been obtained from elder AMI patients. Nevertheless, AMI has recently developed in young adults [9–12]. Every year 1.5 million Americans are stricken by AMI disease, about 80,000 of them are 40 years old or less [13]. Framingham study has reported AMI morbidity for 10 years as 12.9 men in 30 to 34 years old and 5.2 women in 35 to 44 years old [14, 15]. The proportion of AMI patients who are 40 years old or less is 2 to 8 % of all heart diseases, while patients who are 46 years old or less have increased to 10 % [16–18].

In this paper, to reduce the number of ambiguous patterns, we propose improved association rule mining method that can generate target patterns with association between blood factors and disease history. Proposed method can mine interesting patterns that include medical knowledge and satisfy statistical measures in the knowledge discovery process as shown in Fig. 1. Korean Acute Myocardial Infarction Registry (KAMIR) stores 14,885 AMI patients with 141 risk factors. Some missing values and errors of these dataset are eliminated in data preprocessing step. We select 1,247 young adults and 12 risk factors for discovering associations between blood factors and disease history.

Proposed method consists of Complete Target Pattern tree (CTP-tree) which can mine all frequent itemsets, and Pattern Generation algorithm which can generate

D. G. Lee · K. S. Ryu · M. Bashir · J.-W. Bae · K. H. Ryu (✉)
Chungbuk National University, Cheongju-si, Chungcheongbuk-do, South Korea
e-mail: khryu@dblab.chungbuk.ac.kr

Published online: 15 January 2013

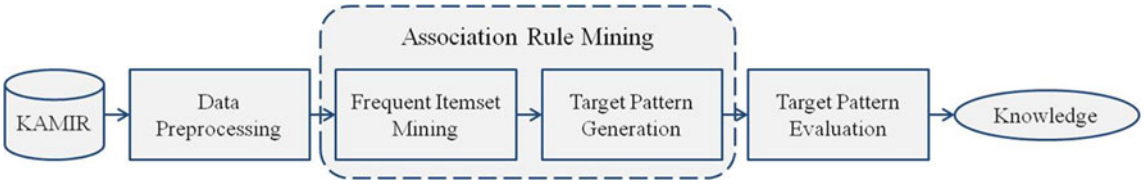

Fig. 1 A knowledge discovery process for mining significant medical knowledge using association rule mining method

target patterns. A large number of patterns are generated from frequent itemsets that are a subset of 12 risk factors greater than or equal to minimum support threshold. However, since almost patterns have imprecise associations in relationship between antecedent and consequence, we define a target pattern in that blood factors affect specific disease history. Those target patterns are extracted in all patterns generated from frequent itemsets. I.e., {smoking, glucose}→{diabetes}, {smoking, total cholesterol}→{diabetes, hypertension}.

In our experiments, we mine the target patterns associated with blood factors and disease history in 1,247 young AMI patients who 45 years old or less. The target patterns are equal or greater than the minimum support and confidence thresholds. However, the support and confidence framework cannot ensure the interestingness of patterns. We need statistical measures to evaluate interestingness such as lift, leverage, and conviction. Typical target patterns that include medical knowledge between blood factors and disease history are well-evaluated on various statistical measures.

Related work

Existing studies have been applied statistical methods and data mining approaches into medical datasets for finding important risk factors that affect relevant diseases. Association rule mining is one of the data mining techniques that can find associations between antecedent and consequence. Representative algorithms that can mine association rules are Apriori [19] and FP-Growth [20]. This section reviews knowledge discovery using association rule mining in medical studies based on real datasets of patients. Moreover, evaluating the interestingness of a pattern is significantly important and statistical measures are required for objective evaluation. We see what measures are used to evaluate the interestingness of a pattern.

STULONG is an epidemiologic study that elaborates the risk factors of atherosclerosis which is one of the diseases occurring more frequently among middle aged men [21]. It develops slowly and involves multiple causes such as high blood pressure, dyslipidemia, alcohol consumption, and tobacco smoking. In STULONG study, 4 ft-Miners [22] are used to automatically extract a set of conditional association rule using 4 ft-quantifiers. There is a study that presents a new approach, which employs an effective data mining method to find the association rules related to hyperlipidemia [23]. In general, association rule mining generates a large number of rules and most rules are medically irrelevant. To mitigate this limitation, in [24], the authors introduce an algorithm that uses search constraints to reduce the number of rules in 655 patients with heart disease. A study that generates association rules with high confidence elaborates general decision model for medical diagnosis based on the so-called confirmation rules that are generated separately for each diagnostic class so that selected rules cover many numbers of target classes [25]. In the French biomedical database as the STANISLAS cohort, Apriori-Rare [26] computes the set of minimal rare itemsets [27]. In [28], frequent patterns are mined by MAFIA (Maximal Frequent Itemset Algorithm) from a heart disease dataset. MAFIA [29] efficiently extracts the association rules when the dataset specifically consists of very long itemsets. After mining the frequent patterns using MAFIA, the significance weightage of each pattern is calculated. The significant patterns are determined based on the significance weightage greater than the pre-defined threshold. Apriori algorithm has been applied to coronary heart disease datasets to assess heart event related to risk factors [30]. HASARD (Hybrid Adaptative Sequential Association Rules Discovery) extracts the adaptative temporal association rules for the analysis of long term medical observations of atherosclerosis risk factors. The HASARD approach combines the CLOSE algorithm [31] and heuristic approach relying on a genetic algorithm for searching closed itemsets [32].

HASARD also evaluates the quality of individual association rule based on the multiplication of support, confidence, and lift measures, defined as Fitness function. The strongest patterns are mined by RSD (Relational Subgroup Discovery) [33] in medical sequential data of STULONG Study [34]. In order to evaluate the quality of a rule confidence, coverage and lift are applied to sequential data and non-sequential data, which include cardiovascular risk factors. The AKAMAS algorithm is a variant of the Apriori algorithm, which does not use the iterative technique of k-itemset to build the (k+1)-itemsets [35]. It offers various measures to evaluate the quality of a rule. AKAMAS can find the most important risk factors, which are gender, family history and smoking. The association rules with improved semantics are found by means of fuzzy sets from large medical databases [36]. Also, to evaluate the accuracy of a rule, a new measure is used as called the certainty factor (CF). Contrary to confidence, it detects both

statistical negative dependence and independence between antecedent and consequent.

AMI dataset

KAMIR has stored the medical data of Korean AMI patients. It is an online Korean prospective multicenter registry that has been investigating the risk factors of AMI patients since November 2005. KAMIR consists of 51 community and teaching hospitals with facilities for primary percutaneous coronary intervention (PCI) and on-site cardiac surgery. Medical data were collected by trained research coordinators using a standardized case-report and protocol. The research population was enrolled as 14,885 AMI patients with 141 risk factors in a nationwide prospective KAMIR [37, 38].

We extract 1,247 AMI patients who are 45 years old or less from total 14,855 patients. 12 of 141 risk factors are selected to find associations between blood factors and disease history in young AMI adults. Selected 12 risk factors are shown in Table 1. Blood factors indicate pack/years (PY), glucose (Gl), creatinine (Cr), total cholesterol (TC), triglyceride (TG), high density lipoprotein cholesterol (HDL-C), low density lipoprotein cholesterol (LDL-C), and high sensitive C reactive protein (hs-CRP). Ischemia (IC), hypertension (HT), diabetes mellitus (DM) and hyperlipidemia (HL) are represented in disease history that has already been known or diagnosed in the past.

Pack/years (PY) The risk of smoking is represented by pack/years. It is a way to calculate the amount that a person has smoked over a long period of time. Given CPD is the number of cigarettes smoked per day, YS is the number of years smoked; it is calculated via the following formula:

$$\mathrm{PY} = (\mathrm{CPD}/20) \times \mathrm{YS} = (\mathrm{CPD} \times \mathrm{YS})/20 \quad (1)$$

where one pack is 20 cigarettes. For example, a patient who has smoked 10 cigarettes every day for 30 years has (10×30)/20=15 pack/years. However, determining an objective risk range of the calculation is still a difficult problem to resolve despite PY is an important factor in AMI. It also has caused a controversy in the medical community.

Glucose (Gl) After ingesting food, since the blood sugar level is quickly elevated to 140 mg/dL, Gl may be raised a bit. The American Diabetic Association (ADA) recommends that the glucose level is less than 180 mg/dL after meals, and 90 to 130 mg/dL before meals.

Creatinine (Cr) The normal range for men is 0.7 to 1.2 mg/dL, and 0.5 to 1.0 mg/dL for women. 2.0 mg/dL for bodybuilders indicates the normal range, and 1.2 mg/dL of old women may mean critical kidney disease. If the filtering of the kidney is deficient, then the Cr blood levels increase.

Cholesterol (TC, LDL-C, HDL-C) Total Cholesterol produced by the liver is more than the amount in ingested food. More than two thirds of the cholesterol in blood is in LDL-C, which is more useful than the total cholesterol as the predictive factor of AMI. LDL-C is often contrasted with HDL-C, which is sometimes called "good cholesterol". AMI patients may have fewer problems of high HDL-C level, while AMI includes high disease rates on the low HDL-C level. The LDL-C level is desirable to be managed at less than 100 mg/dL in AMI patients.

Triglyceride (TG) As it is one kind of fat, unused energy is stored to subcutaneous fat, most of which is TG. After ingesting food, it is absorbed in the small intestine, combined with lipoprotein, and it then flows in a blood. While the TG level is increased in the blood, it increases the risk of AMI as cholesterol. The cholesterol level does not cause very high risk in Korean AMI patients and the TG level may cause high risk.

High sensitive C reactive protein (hs-CRP) The normal range of total AMI patients is equal to or less than 10 mg/dL. 10 to 40 mg/dL is usually measured in late pregnant women and inflammation and viral infections. hs-CRP has been investigated extensively as an important factor predicting the risk of cardiovascular disease [39, 40]. There is also a study that hs-CRP is associated with hypertension and diabetes [41].

The blood factors as continuous data type are discretized to the categorical data type on risk ranges as shown in Table 2. The data discretization is required to mine patterns in dataset.

Table 1 Representing blood factors and disease history with risk factors, abbreviation, and data type

Class	Risk Factor	Abbr.	Data type
Blood Factors	Glucose	Gl	Numeric
	Pack/years	PY	Numeric
	Creatinine	Cr	Numeric
	Total Cholesterol	TC	Numeric
	Triglyceride	TG	Numeric
	High density lipoprotein cholesterol	HDL-C	Numeric
	Low density lipoprotein cholesterol	LDL-C	Numeric
	High sensitive C reactive protein	hs-CRP	Numeric
Disease History	Ischemic	IC	Yes or No
	Hypertension	HT	Yes or No
	Diabetes	DM	Yes or No
	Hyperlipidemia	HL	Yes or No

Table 2 Representing risk ranges such as Normal (=N), High (=H) and Higher (=HH) and the unit to measure the levels in each risk factor

Factor	Unit	Normal(N)	High(H)	Higher(HH)
PY	Pack/years	N=0	0<H≤30	HH>30
Gl	mg/dL	N<80	80≤H<140	HH≥140
Cr	mg/dL	N<0.5	0.5≤H<1.3	HH≥1.3
TC	mg/dL	N<200	200≤H<240	HH≥240
TG	mg/dL	N<150	150≤H<500	HH≥500
HDL-C	mg/dL	N≥60	50≤H<60	HH<50
LDL-C	mg/dL	N<100	100≤H<200	HH≥200
hs-CRP	mg/dL	N<10	10≤H<40	HH≥40

The categorical value is divided into Normal, High, and Higher. 'Normal' means no risk. 'High' has high risk, and 'Higher' contains significant risk. When PY, Gl, Cr, TC, TG, LDL-C, and hs-CRP indicate small level, the risk factors mean normal status. When HDL-C has small level, it presents high risk.

The risk factors mentioned above are deeply associated with the disease history in AMI patients. The associations between disease history and AMI patients have been studied in [42–44]. We discover the associations between blood factors and disease history in young adults with AMI. Though existing studies have already investigated most risk factors, most studies have been worked on elder AMI patients. Therefore, we focus on discovering interesting medical knowledge that includes association between blood factor and disease history in young adults with AMI.

Association rule mining

The ARM consists of frequent itemset mining and pattern generation. Mining frequent itemsets finds all frequent itemsets that are greater than or equal to the minimum support threshold predefined. Pattern generation produces combinations of frequent itemsets on more than the minimum confidence threshold, called an association rule or a pattern. In this section, we describe the association rule mining method which can mine complete target patterns based on the support and confidence framework.

Preliminaries

Let $I=\{i_1, i_2, \ldots, i_m\}$ be a set of items, and $TD=\{T_1, T_2, \ldots, T_n\}$ be a set of transactions, called a *transaction database*. Each transaction T is a set of items, where $T \subseteq I$. Let X be a subset of items. For $X \subseteq I$, if $X \subseteq T$, a transaction T contains X. The set X is called an *itemset* or *k-itemset* if it contains k items. The *count* of an itemset X is the number of transactions that contain X in TD. The *support* of an itemset X is the proportion of transactions in TD that contain X. Thus, if the total number of transactions in TD is n, then the *support* of X is the *count* of X in TD divided by n as formula (2).

$$support(X, TD) = count(X, TD)/n \tag{2}$$

An itemset X is called *frequent* if the *support* of X is greater than or equal to the predefined *minimum support threshold*, called *minsup*. The *minimum support threshold* is predefined. All frequent itemsets, where $support(X) \geq minsup$, are defined as the *complete set*. The problem of finding the *complete set* of *frequent itemsets* is called the *frequent itemsets mining*.

The pattern generation step discovers all association rules that have *confidence* greater than or equal to *minimum confidence threshold,* called *minconf,* from complete frequent itemsets. In association rules, if X and Y are *itemsets*, and $X \cap Y=\{\}$, then a rule is expressed with $X \rightarrow Y$, where $X \rightarrow Y$ is called the *association rule*. If a transaction T contains all items of X, then T also contains all items of Y; X is called *antecedent*, and Y is called *consequence*. The *support* of *association rule* $X \rightarrow Y$ is the *support* of $X \cup Y$ in TD. Namely, *support* of *association rule* means the frequency of $X \cup Y$. If the *support* of $X \rightarrow Y$ is greater than or equal to given *minsup*, then the *association rule* is called *frequent*. The *confidence* of an association rule $X \rightarrow Y$ is expressed as the conditional probability, in which Y is contained in T and X is also included in T as formula (3).

$$\begin{aligned} confidence(X \rightarrow Y, TD) &:= \mathrm{P}(\mathrm{Y}|\mathrm{X}) \\ &= \mathrm{support}(\mathrm{X} \cup \mathrm{Y}, \mathrm{TD})/\mathrm{support}(\mathrm{X}, \mathrm{TD}). \end{aligned} \tag{3}$$

If $P(Y|X)$ is greater than or equal to *minconf*, then the association rule $X \rightarrow Y$ is called *reliable*, where1 $0 \leq c \leq 1$. The importance of rules discovered in ARM is decided based on *support* and *confidence* being greater than or equal to the predefined *minsup* and *minconf* thresholds.

Generating complete target pattern

The association rule mining technique consists of CTP-tree structure, frequent itemset mining, and target pattern generation. In the building process of CTP-tree, the header table is constructed by counting the items in the transaction database, and CTP-tree is then constructed using the header table. The CTP-tree is used to mine frequent itemsets greater than or equal to *minsup* in the transaction database. From frequent itemsets, we can generate candidate patterns. If the candidate patterns are greater than or equal to *minconf*, then we can obtain the association rules. The association rules can be removed, except for target patterns with disease history such as ischemic, hypertension, diabetes, and hyperlipidemia in the consequence of a pattern in pruning step.

```
Algorithm BuildHeaderTable (D, minsup)
Input: TD is the transaction database;
       minsup is the minimum support threshold;
Output: a header table;

01: scan transaction t in TD;
02:     for each item i in t
03:         if i is in the header table
04:             item count++;
05:         else
06:             i is added in the header table;
07:         end if
08:     end for
09: delete items if infrequent;
10: sort frequent items in support-descending order;
```

Fig. 2 Building the header table from a transaction database

Building the Teader Table The header table consists of items and a list of frequent count. An item is an element in transactions. Frequent count is the number of each item. To compute the frequency count of each item, we need to scan the transaction database. While scanning all transactions, we need to add new items, or count the items existing in the header table. If the *support* of each item is less than *minsup*, then it is removed from the header table. Frequent items are sorted in descending order according to support in the header table. An algorithm of constructing the header table is described in pseudo code as shown in Fig. 2.

Constructing CTP-tree A node includes the link information such as items, level, and count. The level points are the location of the first item in a transaction. Count means the number of an item in the transaction database. The initial tree is constructed via combination of the header table and single link, whose nodes have an item, level = 0, and count = 0. The initial tree easily builds the CTP-tree based on the transaction database.

In building the CTP-tree, node insertion and split can be occurred. In the case of node split, the link information of nodes is reconstructed. The reconstruction time is large. However, since the header table has the link information and knows the locations of individual node, the building time of the tree can be reduced.

```
Algorithm BuildCTPtree (TD, minsup)
Input: TD is the transaction database;
       minsup is the minimum support threshold;
Output: CTP-tree;

01: initialize the header table;
02: scan transaction t in TD;
03: if current node is split
04:     level count++;
05: else
06:     add i in current node;
07:end if
```

Fig. 3 An algorithm of constructing the CTP-tree structure from items with the support equal or greater than the minimum support threshold (*minsup*)

```
Algorithm MineFrequentItemset (CTP-tree, minsup)
Input: CTP-tree;
       minsup;
Output: Frequent Item Set List (FISL);

01: if i is linked by single path in localTree
02:     add frequent itemsets into FISL;
03: else
04:     create localTree;
05:     call Mine(localTree);
06:end if
```

Fig. 4 Mining all frequent itemsets by recursive call

The building step initializes CTP-tree through the header table as shown in Fig. 3. While scanning the transaction database, the building step constructs the CTP-tree and decides to split a node for assigning the items. The current node is divided into new nodes whose level count increases by 1. If the node split is unnecessary, then the items in a transaction are inserted into the current node. We can thus build the CTP-tree for mining frequent itemsets using the initial tree with header table.

Mining Frequent Itemsets An algorithm, which can mine frequent itemsets is shown in Fig. 4. The complete CTP-tree is applied to the Mine function as a parameter that can extract frequent itemsets. The mine function determines whether or not the nodes are connected through a single path. If a frequent itemset is connected through a single path, then it is inserted in FISL. If not, then the local tree of each item is constructed and also FIM is performed via recursive call.

Generating Target Pattern We mine a large number of patterns from 1,247 AMI patients using the association rule mining technique. Most mined patterns are irrelevant to medical meanings and much time is wasted for finding meaningful patterns. To reduce the number of patterns mined and time cost for finding meaningful patterns, we set the so-called 'target pattern' which has specific associations between

```
Algorithm GenPattern(FISL, minconf)
Input: FISL;
       minconf is minimum confidence threshold;
Output: Patterns;
01: Patterns(X→ Y) =genCandidate (FISL);
02: confidence(X→ Y, TD) = support(X∪ Y, TD)/support(X, TD);
03: if conf≥ minconf and Patterns(X→ Y)=TP
04:     output Patterns(X→ Y);
05:else
06:     delete Patterns(X→ Y);
07: end if
08: call GenPattern(FISL, minconf);
```

Fig. 5 An algorithm of generating all patterns that are greater than or equal to the minimum confidence threshold (*minconf*)

Table 3 Representing the target patterns associated with cigarette smoking, obesity, and diabetes in AMI patients who are 45 years old or less based on conf. (≥30 %), lift (>1), leverage (>0) and conviction (>1)

ID	Antecedent	Consequence	Conf.(%)	Lift	Leverage	Conviction
TP_1	$PY_{HH} \wedge Gl_{HH} \wedge TG_H$	DM_Y	42.25	2.72	0.02	1.43
TP_2	$PY_{HH} \wedge Gl_{HH}$	DM_Y	39.13	2.52	0.03	1.37
TP_3	$Gl_{HH} \wedge TG_H$	DM_Y	35.17	2.26	0.04	1.29
TP_4	$Gl_{HH} \wedge TC_H$	DM_Y	29.61	1.90	0.02	1.19
TP_5	Gl_{HH}	DM_Y	30.45	1.96	0.07	1.21

antecedent and consequence as defined in Def. 1. However, all target patterns may not have medical meanings. While a target pattern exhibits desirable relationships, it sometimes indicates irrelevant information or trivial value on the medical meaning.

Definition 1 (*TP*: Target Pattern) let $BF=\{B_1, B_2, \ldots, B_i\}$ be a set of blood factors, and $DH=\{D_1, D_2, \ldots, D_i\}$ be a set of disease history. Let *bf* be a subset of *BF*, and *dh* be a subset of *DH*. Target Pattern *TP* is,

$$TP : bf \rightarrow dh$$

where $bf \subseteq BF$, $dh \subseteq DH$ and $BF \cap DH = \{\}$. *BF* is in the antecedent of a target pattern and *DH* occurs in the consequence of a target pattern. All target patterns are satisfied on the minimum support and minimum confidence thresholds.

In order to generate the association rules, we generate candidate patterns as a combination of frequent items from FISL as shown in Fig. 5. The *confidence* of each candidate pattern is determined by the *support* computed during frequent itemset mining. If the *confidence* of a pattern is equal to or greater than *minconf* and the pattern is concurrently the target pattern as defined in Def. 1, then it is decided as an association rule or a target pattern. If not, then it is eliminated. Such process is performed recursively until the confidence of all patterns is determined.

Experimental evaluation

Proposed method can mine target patterns with associations between 8 risk factors and 4 disease histories in young adults with AMI. Used dataset is 1,247 AMI patients who are 45 years old or less. The target patterns are associated with diabetes and hypertension in young adults with AMI. The *minsup* and *minconf* are 1 % and 10 %, respectively. We find all target patterns satisfied on *minsup* and *minconf*. However, the support and confidence thresholds may be assured the interestingness of target patterns. Although the target patterns satisfy on *minconf* and casual relationship between antecedent and consequence, almost patterns may be also ambiguous.

The ambiguous patterns are evaluated on an objective measure. The ARM technique has the potentiality that generates a large number of the patterns corresponding with medical knowledge on the support and confidence thresholds. Since we can easily find thousands or even millions of uninterested patterns, objective measures that can guarantee the interestingness of the target patterns are important to be established. The pattern evaluation step measures the interestingness of a target pattern based on the statistical interest factors such as Lift [45], Leverage [46], and Conviction [47].

An association rule is presented as $X \rightarrow Y$, where *X* and *Y* are itemsets, and $X \cap Y = \{\}$. *X* is called as antecedent and *Y* is the consequence of a pattern. Since the confidence (*c*) ignores the support (*s*) of the itemset presented in the consequence, the rules with high confidence can sometimes be misleading. One way to address this problem is by applying a metric known as the following lift:

$$\mathrm{Lift}(X \rightarrow Y) = c(X \rightarrow Y)/s(Y) \quad (4)$$

Lift measures the ratio between the rule's confidence and the support of the itemset in *Y* of a pattern. Leverage measures the difference between *X* of *Y* appearing together in the dataset. The minimum leverage threshold at the same time incorporates an implicit frequency constraint. A second way based on statistical arguments is by applying as the following

Table 4 Target patterns associated with glucose, cigarette smoking, obesity, and kidney function and hypertension in AMI patients who are 45 years old or less based on conf. (≥30 %), lift (>1), leverage (≥0) and conviction (>1)

ID	Antecedent	Consequence	Conf.(%)	Lift	Leverage	Conviction
TP_6	$Gl_{HH} \wedge CR_{HH}$	HT_Y	44.45	1.52	0.01	1.25
TP_7	$Gl_{HH} \wedge TC_H \wedge TG_H$	HT_Y	40.79	1.39	0.01	1.17
TP_8	$Gl_{HH} \wedge TG_H$	HT_Y	37.71	1.29	0.02	1.13
TP_9	$PY_{HH} \wedge TC_H$	HT_Y	32.58	1.11	0.00	1.04

Table 5 Target patterns with two disease histories associated with glucose, cigarette smoking, and obesity in AMI patients who are 45 years old or less based on conf. (≥15 %), lift (>1), leverage (>0) and conviction (>1)

ID	Antecedent	Consequence	Conf.(%)	Lift	Leverage	Conviction
TP_{10}	$Gl_{HH} \wedge TC_H \wedge TG_H$	$HT_Y \wedge DM_Y$	17.11	2.60	0.01	1.11
TP_{11}	$Gl_{HH} \wedge TG_H$	$HT_Y \wedge DM_Y$	16.95	2.58	0.02	1.12
TP_{12}	$PY_{HH} \wedge Gl_{HH}$	$HT_Y \wedge DM_Y$	16.67	2.53	0.01	1.11

leverage:

$$\text{Leverage}(X \rightarrow Y) = c(X \rightarrow Y) - s(Y) \quad (5)$$

As a third way for measuring the objective interestingness, conviction compares the probability that X appears without Y if they were dependent from the actual frequency of the appearance of X without Y, which is defined as follows:

$$\text{Conviction}(X \rightarrow Y) = (1 - s(Y))/(1 - c(X \rightarrow Y)) \quad (6)$$

Using Eqs. (4) and (6), we can evaluate the interestingness factor (IF) as the lift and conviction measures as follows:

- IF(X, Y)=1, if X and Y are independent
- IF(X, Y)>1, if X and Y are positively correlated
- IF(X, Y)<1, if X and Y are negatively correlated

Using Eq. (5), we can interpret the leverage measure as follows:

- IF(X, Y)=0, if X and Y are independent
- IF(X, Y)>0, if X and Y are positively correlated
- IF(X, Y)<0, if X and Y are negatively correlated

Young adults with AMI are associated with diabetes as shown in Table 3. Namely, diabetes is occurred by cigarette smoking, glucose, triglyceride, and total cholesterol. We consider that the confidence is more than 30 % and 5 target patterns are then extracted from a large number of target patterns based on lift, leverage, and conviction. We evaluate the interestingness of the target patterns based on lift, leverage, and conviction measure. The lift and the conviction signify high interestingness when the measures are greater than one. The leverage measure is desirable when it is greater than zero.

TP_1 represents 3 risk factors such as cigarette smoking, glucose, and triglyceride, which contain relatively higher interestingness than other target patterns. TP_5 has a single factor as glucose that influences diabetes. TP_2 and TP_3 show that glucose is combined with triglyceride and pack/years, nevertheless they indicate higher interestingness than TP_5. In contrast, the combination of glucose and total cholesterol reduces the interestingness as shown in TP_4. Ultimately, in young adults with AMI, glucose is a significant factor that develops diabetes. The target patterns that contain a combination of pack/years, triglyceride, and glucose indicate higher interestingness in diabetes. We evaluate that TP_1 is reliable in terms of lift, leverage, and conviction measure. TP_1 represents the association between 3 risk factors and diabetes that has high interestingness in the statistical measures. 30 of 71 young adults with pack/years, glucose, and triglyceride have diabetes whose proportion is 42.3 %.

The target patterns associated with hypertension are shown in Table 4. Glucose factor is significantly important for hypertension like the target patterns related to diabetes. Creatinine is found newly in target patterns related to hypertension. Pack/years, triglyceride, and total cholesterol still influence hypertension.

TP_6 with glucose and creatinine have greater reliability than other patterns in confidence, lift, and conviction. TP_7 and TP_8 contain also total cholesterol and triglyceride related to obesity, which have high interestingness in the statistical measures. TP_9 has reliable interestingness in confidence, lift, and conviction measure. However, TP_9 represented as 0 on leverage measure. It means that the probability of risk factors occurred concurrently is the same with the probability occurred respectively in antecedent and consequence. Namely, antecedent and consequence are likely to be independent. The association between pack/years and hypertension is significant for young adults with AMI.

As addressed above, the target patterns with single disease history are only evaluated. AMI patients may be suffered from multiple diseases as well as single disease. Hypertension and diabetes are occurred together as shown in Table 5. The target patterns with both hypertension and diabetes have significant influence on AMI. Also similarly as shown in Tables 3 and 4, risk factors are composed of

Table 6 Summarizing significant risk factors discovered from young adults with AMI in 6 countries and our study

	Risk Factors
US	Smoking, Family history, Hypertension, Hyperlipidemia
Canada	Smoking, Family history
Thai	Smoking, Family history
Korea	Smoking, Hypercholesterolemia
Italia	Smoking, Family history, Dabetes, Hypertension, Hpercholesterolemia
Taiwan	Smoking, Diabetes, Hypertension, Obesity
Our Study	Smoking, Diabetes, Hypertension

glucose, total cholesterol, triglyceride, and pack/years in Table 5. Confidence of TP_{10}, TP_{11}, and TP_{12} is lower than other target patterns shown in Tables 3 and 4. Although TP_{10}, TP_{11}, and TP_{12} indicate low confidence, lift, leverage, and conviction are acceptable. Lift and conviction measure have positive correlation when the measures are greater than one. Leverage has positive correlation when it is greater than zero. Table 5 shows that three measures satisfy positive correlation in TP_{10}, TP_{11}, and TP_{12}. Glucose and triglyceride are significant risk factors that can affect both hypertension and diabetes.

Discussion

Several countries have been studied the risk factors related to young adults with AMI. The Coronary Artery Surgery Study (CASS) included a registry of 24,958 consecutive patients undergoing diagnostic cardiac catheterization for suspected coronary artery disease at 15 institutions in the United States and Canada [15]. The smoking factor in CASS is significant for young adults with AMI. Family history is also important in 55 years old or younger people. The University of Michigan assesses the frequency and risk factors of young patients with AMI. Their young patients are represented over 10 % of all AMI patients and the risk factors are family history and smoking [18]. In the Thai Acute Coronary Syndrome (ACS) Registry, young patients are 5.8 % whose risk factors are smoking and family history, whereas diabetes and hypertension are infrequent [48]. A retrospective cross-sectional study has been performed for more than 7 years to define the medical explanation and risk factors for myocardial infarction patients 40 years old or less, whose risk factors are followed by smoking, family history, hypertension, and hyperlipidemia [49]. Yonsei University Severance Hospital in Korea investigates the risk factors in 631 AMI patients, and young adults are 10.3 % whose risk factors are smoking and hypercholesterolemia [50]. The evaluation of haemorheological pattern for young AMI patients is performed in the initial stage, 3 months, and 12 months. The risk factors are studied in 96 young adults who are 46 years old or less. The risk factors include family history with coronary artery disease, smoking or past smoking, hypercholesterolemia, diabetes, and hypertension [51]. In Taiwan, 70 young adults are analyzed whose risk factors are smoking, diabetes, hypertension, and obesity [52].

In the United States, Canada, and Thai, smoking and family history were imp ortant factors as shown in Table 6. Hyperlipidemia is important risk factor for the Americans. Korean included smoking, and hypercholesterolemia. Taiwanese was associated with smoking, diabetes, hypertension, and obesity. Italian is related with family history, smoking, hypercholesterolemia, diabetes and hypertension. We discovered glucose, triglyceride, cigarette smoking, total cholesterol, and creatinine which are associated with hypertension and diabetes in young AMI patients. Thus, the common significant risk factor is smoking. We discovered that Koreans have high risk in glucose, smoking, and creatinine which are associated with diabetes, and hypertension.

Conclusion

We proposed association rule mining method that can discover target patterns associated with hypertension and diabetes from young AMI patients who are 45 years old or less. The discovered risk factors are glucose, triglyceride, cigarette smoking, total cholesterol, and creatinine which are associated with hypertension and diabetes in young adults with AMI.

The existing association rule mining has been computed to eliminate uninteresting patterns on the support and confidence thresholds. The support measure may eliminate many potentially interesting patterns with low support. The confidence threshold ignores the support of the itemset in consequence of patterns. Therefore, we used lift, leverage, and conviction measure to evaluate the interestingness of patterns. It was sufficient to focus on a combination of support, confidence, lift, leverage, and conviction to evaluate quantitatively the interestingness of the target patterns. However, the reliability and the reasonability of a target pattern are subjective in the medical domain.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST)(No. 2012-0000478).

References

1. Gu, D., Liang, C., and Li, X., Intelligent technique for knowledge reuse of dental medical records based on case-based reasoning. *J. Med. Syst.* 34:213–222, 2010.
2. Du, G., Jiang, Z., Diao, X., and Yao, Y., Knowledge extraction algorithm for variances handling of CP using integrated hybrid genetic double multi-group cooperative PSO and DPSO. *J. Med. Syst.* 36:979–994, 2012.
3. Arif, M., Malagore, I. A., and Afsar, F. A., Detection and localization of myocardial infarction using K-nearest neighbor classifier. *J. Med. Syst.* 36:279–289, 2012.
4. Shon, H. S., Ryu, K. S., Park, S.H., Bae, J.W., Cha, E. J. and Ryu, K. H., Risk factors of major adverse cardiac events after percutaneous coronary intervention in non ST elevation myocardial infarction. *Int. Conf. Ubiquit. Healthc*. 58–60, 2011.
5. Li, P., Pok, G., Jung, K. S., Shon, H. S., and Ryu, K. H., QSE: A new 3-D solvent exposure measure for the analysis of protein structure. *Proteomics* 11(19):3794–3801, 2011.
6. Bashir, M. E., Lee, D. G., Akasha, M., Yi, G. M., Cha, E. J., Bae, J. W., Cho, M. C., and Ryu, K. H., Highlighting the current issues with pride suggestions for improving the performance of real time cardiac health monitoring. *Inf. Technol. Bio- Med. Inform* 6266:226–33, 2010.
7. Bashir, M. E., Ryu, K. S., Park, S. H., Lee, D. G., Shon, H. S., and Ryu, K. H., Superiority real-time cardiac arrhythmias detection

using trigger learning method. *Inf. Technol. Bio- Med. Informa.* 6865:53–65, 2011.
8. Shon, H. S., Ryu, K. H., Yang, K. S., and Yoo, C. W., Feature selection method using WF-LASSO for gene expression data analysis. *ACM Conf. Bioinforma, Comput. Biol. Biomed.* 522–24, 2011.
9. Towbin, J. A., Bricker, J. T., and Garson, A., Electrocardiographic criteria for diagnosis of acute myocardial infarction in childhood. *Am. J. Cardiol.* 69(19):1545–1548, 1992.
10. Weinberger, I., Rotenberg, Z., Fuchs, J., Sagy, A., Friedmann, J., and Agmon, J., Myocardial infarction in young adults under 30 years: Risk factors and clinical course. *Clin. Cardiol.* 10(1):9–15, 1987.
11. Chouhan, L., Hajar, H. A., and Pomposiello, J. C., Comparison of thrombolytic therapy for acute myocardial infarction in patients aged <35 and >55 years. *Am. J. Cardiol.* 71(2):157–159, 1993.
12. Perski, A., Olsson, G., Landou, C., de Faire, U., Theorell, T., and Hamsten, A., Minimum heart rate and coronary atherosclerosis: Independent relations to global severity and rate of progression of angiographic lesions in men with myocardial infarction at a young age. *Am. J. Cardiol.* 123(3):609–616, 1992.
13. AHA (American Heart Association), *Heart and Stroke Facts Statistics*. American Heart Association, Dallas, 1993.
14. Kannel, W. B., and Abbott, R. D., Incidence and prognosis of unrecognized myocardial infarction. An update on the Framingham study. *N. Engl. J. Med.* 311(18):1144–1147, 1984.
15. Zimmerman, F. H., Cameron, A., Fisher, L. D., and Ng, G., Myocardial infarction in young adults: Angiographic characterization, risk factors and prognosis (Coronary Artery Surgery Study Registry). *J. Am. Coll. Cardiol.* 26(3):654–661, 1995.
16. Füllhaas, J. U., Rickenbacher, P., Pfisterer, M., and Ritz, R., Long-term prognosis of young patients after myocardial infarction in the thrombolytic era. *Clin. Cardiol.* 20(12):993–998, 1997.
17. Imazio, M., Bobbio, M., Bergerone, S., Barlera, S., and Maggioni, A. P., Clinical and epidemiological characteristics of juvenile myocardial infarction in Italy: The GISSI experience. *G. Ital. Cardiol.* 28(5):505–512, 1998.
18. Doughty, M., Mehta, R., Bruckman, D., Das, S., Karavite, D., Tsai, T., and Eagle, K., Acute myocardial infarction in the young-The University of Michingan experience. *Am. Heart J.* 143(1):56–62, 2002.
19. Agrawal, R., and Srikant, R., *Fast algorithms for mining association rules in large databases. Int. Conf. Very Large Data Bases*. 487–99, 1994.
20. Han, J., Pei, J., and Yin, Y., Mining frequent patterns without candidate generation. *ACM SIGMOD Int. Conf. Manag. Data* 29 (2):1–12, 2000.
21. STULONG study website, Available: http://euromise.vse.cz/challenge/. 2002.
22. Rauch, J., and Šimůnek, M., Alternative approach to mining association rules. *Found. Data Min. Knowl. Disc.* 6:211–31, 2005.
23. Dogan, S., and Turkoglu, I., Diagnosing hyperlipidemia using association rules. *Math. Comput. Appl.* 13(3):193–202, 2008.
24. Ordonez, C., Association rule discovery with the train and test approach for heart disease prediction. *IEEE Trans. Inf. Technol. Biomed.* 10(2):334–343, 2006.
25. Gamberger, D., Lavrač, N. and Jovanoski, V., High confidence association rules for medical diagnosis. *Intell. Data Anal. Med. Pharmacol.* 42–51, 1999.
26. Szathmary, L., Napoli, A., and Valtchev, P., Towards rare itemset mining. *Int. Conf. Tools with Artificial Interlligence.* 1:305–312, 2007.
27. Szathmary, L., Valtchev, P., and Napoli, A., Finding minimal rare itemsets and rare association rules. *Knowl. Sci. Eng. Manag.* 6291:16–27, 2010.
28. Patil, S. B., and Kumaraswamy, Y. S., Extraction of significant patterns from heart disease warehouses for heart attack prediction. *Int. J. Comput. Sci. Netw. Secur.* 9(2):228–235, 2009.
29. Burdick, D., Calimlim, M., Flannick, J., Gehrke, J., and Yiu, T., MAFIA: A Maximal Frequent Itemset Algorithm. *IEEE Trans. Knowl. Data Eng.* 17(11):1490–1504, 2005.
30. Karaolis, M., Moutiris, J. A., Papaconstantinou, L. and Pattichis, C. S., Association rule analysis for the assessment of the risk of coronary heart events. *IEEE Eng. Med. Biol. Soc.* 6238–41, 2009.
31. Pasquier, N., Taouil, R., Bastide, Y., Stumme, G., and Lakhal, L., Generating a condensed representation for association rules. *J. Intell. Inform. Syst.* 24(1):29–60, 2005.
32. Brisson, L., Pasquier, N., Hebert, C., and Collard, M., HASARD: Mining sequential association rules for atherosclerosis risk factor analysis. *Eur. Conf. Princ. Pract. Knowl. Discov. Databases.* 14–25, 2004.
33. Lavrač, N., Železný, F., and Flach, P. A., *RSD: Relational subgroup discovery through first-order feature construction*, Lecture Notes in Computer Science, vol. 2583. Springer, Berlin Heidelberg New York, pp. 149–165, 2003.
34. Kléma, J., Holas, T., Železný, F., and Karel, F., Mining the strongest patterns in medical sequential data. *Eur. Med. Biol. Eng. Conf.* 2005.
35. Karaolis, M., Moutiris, J. A., Papaconstantinou, L. and Pattichis, C. S., AKAMAS: Mining association rules using a new algorithm for the assessment of the risk of coronary heart events. *Inf. Technol. Appl. Biomed.* 1–6, 2009.
36. Delgado, M., Sánchez, D., Martín-Bautista, M. J., and Vila, M., Mining association rules with improved semantics in medical databases. *Artif. Intell. Med.* 21:241–245, 2001.
37. Kim, H. K., Jeong, M. H., Ahn, Y., Kim, J. H., Chae, S. C., Kim, Y. J., Hur, S. H., Seong, I. W., Hong, T. J., Choi, D. H., Cho, M. C., Kim, C. J., Seung, K. B., Chung, W. S., Jang, Y. S., Rha, S. W., Bae, J. H., Cho, J. G., and Park, S. J., Other Korea Acute Myocardial Infarction Registry Investigators: Hospital discharge risk score system for the assessment of clinical outcomes in patients with acute myocardial infarction (Korea Acute Myocardial Infarction Registry [KAMIR] score). *Am. J. Cardiol.* 107(7):965–971, 2011.
38. Sim, D. S., Jeong, M. H., and Kang, J. C., Current management of acute myocardial infarction: Experience from the Korea Acute Myocardial Infarction Registry. *J. Cardiol.* 56(1):1–7, 2010.
39. Ridker, P. M., Hennekens, C. H., Buring, J. E., and Rifai, N., C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. *N. Engl. J. Med.* 342 (12):836–843, 2000.
40. Ridker, P. M., Cushman, M., Stampfer, M. J., Tracy, R. P., and Hennekens, C. H., Inflammation, aspirin, and the risk of cardiovascular disease in apparently healthy men. *N. Engl. J. Med.* 336 (14):973–979, 1997.
41. Anand, A. V., Muneeb, M., Divya, N., Senthil, R., Kapoor, M., Gowri, J., and Begum, T. N., Clinical significance of hypertension, diabetes and inflammation, as predictor of cardiovascular disease. *Int. J. Biol. Med. Res.* 2(1):369–373, 2011.
42. Oviagele, B., Markovic, D., and Fonarow, G. C., Recent US patterns and predictors of prevalent diabetes among acute myocardial infarction patients. *Cardiol. Res. Pract.* 2011(145615):1–8, 2011.
43. Lee, M. G., Jeong, M. H., Ahn, Y., Chae, S. C., Hur, S. H., Hong, T. J., Kim, Y. J., Seong, I. W., Chae, J. K., Rhew, J. Y., Chae, I. H., Cho, M. C., Bae, J. H., Rha, S. W., Kim, C. J., Choi, D., Jang, Y. S., Yoon, J., Chung, W. S., Cho, J. G., Seung, K. B., and Park, S. J., Comparison of clinical outcomes following acute myocardial infarctions in hypertensive patients with or without Diabetes. *Korean Circ. J.* 39(6):243–250, 2009.
44. Kang, D. G., Jeong, M. H., Ahn, Y., Chae, S. C., Hur, S. H., Hong, T. J., Kim, Y. J., Seong, I. W., Chae, J. K., Rhew, J. Y., Chae, I. H., Cho, M. C., Bae, J. H., Rha, S. W., Kim, C. J., Jang, Y. S., Yoon, J.,

Seung, K. B., and Park, S. J., Clinical effect of hypertension on the mortality of patients with acute myocardial infarction. *J. Korean Sci.* 24(5):800–806, 2009.
45. Quinlan, J. R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann. 1993.
46. Piatetsky-Shapiro, G., Discovery, analysis, and presentation of strong rules. *Knowl. Disc. Databases* 229:229–248, 1991.
47. Brin, S., Motwani, R., Ullman, J. D., and Tsur, S., Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Int. Conf. Manag. Data* 26(2):255–264, 1997.
48. Tungsubutra, W., Tresukosol, D., Buddhari, W., Boonsom, W., Sanguanwang, S., and Srichaiveth, B., Acute Coronary Syndrome in Young Adults: The Thai ACS Registry. *J. Med. Assoc. Thai.* 1:81–90, 2007.
49. Kanitz, M. G., Giovannucci, S. J., Jones, J. S., and Mott, M., Myocardial Infarction in Young Adults: Risk Factors and Clinical Features. *J. Emerg. Med.* 14(2):139–145, 1996.
50. Hong, M. K., Cho, S. Y., Hong, B. K., Chang, K. J., Chung, M. I., Lee, H. M., Lim, W. S., Kwon, H. M., Jang, Y. S., and Chung, N. S., Acute myocardial infarction in the young adults. *Yonsei Med. J.* 35(2):184–189, 1994.
51. Caimi, G., Valenti, A., and Lo Presti, R., Acute myocardial infarction in young adults: Evaluation of the haemorheological pattern at the initial stage, after 3 and 12 months. *Ann. Ist Super Sanita.* 43 (2):139–143, 2007.
52. Lin, Y., Hsu, L., Ko, Y., Kuo, C., Chen, W., Lin, C., Pan, W., and Chang, C., Impact of conventional cardiovascular risk factors on acute myocardial infarction in young adult Taiwanese. *Acta Cardiol Sin.* 26:228–234, 2010.