International Journal of Modern Physics: Conference Series
Vol. 1, No. 1 (2010) 1–5
© World Scientific Publishing Company
DOI: 10.1142/insert DOI here

[redacted]

DETECTING CRITICAL LEAST ASSOCIATION RULES IN MEDICAL DATABASES

ZAILANI ABDULLAH

Department of Computer Science
Universiti Malaysia Terengganu
Mengabang Telipot, Kuala Terengganu 21030, Terengganu, Malaysia
zailania@umt.edu.my

TUTUT HERAWAN*

Database and Knowledge Management Research Group
Faculty of Computer System and Software Engineering
Universiti Malaysia Pahang
Lebuhaya Tun Razak, Gambang 26300, Kuantan, Pahang, Malaysia
tutut@ump.edu.my

MUSTAFA MAT DERIS
Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia
Parit Raja 86400, Batu Pahat, Johor, Malaysia
mmustafa@uthm.edu.my

Received (Day Month Year)
Revised (Day Month Year)

Least association rules are corresponded to the rarity or irregularity relationship among itemset in database. Mining these rules is very difficult and rarely focused since it always involves with infrequent and exceptional cases. In certain medical data, detecting these rules is very critical and most valuable. However, mathematical formulation and evaluation of the new proposed measurement are not really impressive. Therefore, in this paper we applied our novel measurement called Critical Relative Support (CRS) to mine the critical least association rules from medical dataset. We also employed our scalable algorithm called Significant Least Pattern Growth algorithm (SLP-Growth) to mine the respective association rules. Experiment with two benchmarked medical datasets, Breast Cancer and Cardiac Single Proton Emission Computed Tomography (SPECT) Images proves that CRS can be used to detect to the pertinent rules and thus verify its scalability.

Keywords: Critical; least association rules; medial data.

1. Introduction

Mining association rules (ARs) can be classified as one of the most popular and prominent areas in data mining. It aims at discovering the interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. This term was coined by Agrawal *et al.*[1] and amazingly it stills become an active research in knowledge database discovery. For the past decades, ARs have been widely used in various types of applications such as retail transaction, stock market analysis, etc. In brevity, an item is said to be frequent if it appears more than a minimum support threshold. These frequent items are then used to produce the ARs. Besides that, confidence is another measure that always used in pair with the minimum support threshold.

By definition, least item is an itemset whose rarely found in the database but still can produce interesting and potentially valuable ARs. These rules are very important in discovering rarely occurring but significantly important, such as air pollution detection, critical fault detections, network intrusions, and etc. At the moment, many series of ARs mining algorithms are using the minimum supports-confidence framework to limit the number of ARs. As a result, by increasing or decreasing the minimum support or confidence values, the interesting rules might be missing or untraceable. Since the complexity of study, difficulties in algorithms[2] and it may require excessive computational cost, there are very limited attentions have been paid to discover the highly correlated least ARs.

For both frequent and least ARs, it may have a different degree of correlation. Highly correlated least ARs are referred to the itemsets that its frequency does not satisfy a minimum support but are very highly correlated. ARs are classified as highly correlated if it is positive correlation and in the same time fulfils a minimum degree of predefined correlation. Recently, statistical correlation technique has been widely applied in the transaction databases[3], which to find relationship among pairs of items whether they are highly positive or negative correlated. In reality, it is not absolute true that the frequent items have a positive correlation as compared to the least items.

The low minimum support can be set to capture the least items. However, the trade off is it may generate the huge number of ARs. As a result, it is enormously difficult to identify which ARs are most interesting and really significant. Furthermore, the low minimum support will also proportionally increase the computational performance and its complexity. Since the complexity of study, difficulties in algorithms[2] and it may require excessive computational cost, there are very limited attentions have been paid to discover least ARs.

Therefore, this paper is an attempt to mitigate the mentioned above problems based on three contributions. First, a novel measurement called Critical Relative Support (CRS)[4] is employed to discover the desired critical least ARs. A range of CRS is always in 0 and 1. The more CRS value reaches to 1, the more significant and critical those particular rules. Second, SLP-Growth[5] algorithm and enhanced version of tree data structure called LP-Tree are employed. In order to ensure only certain least items are captured, Interval Least Support (ISupp) is suggested and embedded in the SLP-Growth algorithm. Third, experiments on two UCI[6] medical datasets have been conducted to evaluate the CRS measurement. Resulting from the experiments is very important to measure its effectiveness and scalability.

In this paper, we address the problem of mining least ARs with the objectives of discovering significant least ARs but surprisingly are highly correlated. A new CRS measurement[4] and SLP-Growth algorithm[5] are employed to extract these ARs. The proposed algorithm imposes interval support to capture all least itemsets family first before continuing to construct a significant least pattern tree (SLP-Tree). The correlation technique for finding relationship between itemset is also embedded to this algorithm. Two benchmarked medical datasets called Breast Cancer[7] and SPECT Heart[8] are employed in the experiment.

The reminder of this paper is organized as follows. Section 2 describes the related work. Section 3 explains the basic concepts and terminology of ARs mining. Section 4 discusses the proposed method. This is followed by performance analysis thorugh two esperiment tests in section 5. Finally, conclusion and future direction are reported in section 6.

2. Related Work

Until now, several works have been done in proposing the scalable and efficient methods of frequent ARs. However, only few attentions have been paid for mining least ARs. As a result, ARs that are rarely found in the database are always ignored by the minimum support-confidence threshold. In the real world, the rarely ARs are also providing significant and useful information for experts, particularly in detecting the highly critical and exceptional situations.

Zhou *et al.* in Ref. 9 suggested an approach to mine the ARs by considering only infrequent itemset. The limitation is, Matrix-based Scheme (MBS) and Hash-based scheme (HBS) algorithms are facing the expensive cost of hash collision. Ding in Ref. 10 proposed Transactional Co-occurrence Matrix (TCOM for mining association rule among rare items. However, the implementation of this algorithm is too costly. Yun *et al.* in Ref. 2 proposed the Relative Support Apriori Algorithm (RSAA) to generate rare itemsets. The challenge is if the minimum allowable relative support is set close to zero, it takes similar time taken as performed by Apriori. Koh *et al.* in Ref. 11 introduced Apriori-Inverse algorithm to mine infrequent itemsets without generating any frequent rules. The main constraints are it suffers from too many candidate generations and time consumptions during generating the rare ARs. Liu *et al.* in Ref. 12 proposed Multiple Support Apriori (MSApriori) algorithm to extract the rare ARs. In actual implementation, this algorithm is still suffered from the “rare item problem”. Most of the proposed approaches in Ref. 2, 9-12 are using the percentage-based approach in order to improve the performance of existing single minimum support based approaches.

Brin *et al.* in Ref. 13 presented objective measures called lift and chi-square to measure the correlation of ARs. Lift compares the frequency of pattern against a baseline frequency computed under statistical independence assumption. Instead of lift, there are quite a number interesting measures have been proposed for ARs. Omiecinski in Ref. 14 introduces two interesting measures based on downward closure property called all confidence and bond. Lee *et al.* in Ref. 15 proposes two algorithms for mining all confidence and bond correlation patterns by extending the frequent pattern-growth methodology. Han *et al.* in Ref. 16 proposed FP-Growth algorithm which break the two bottlenecks of Apriori series algorithms. Currently, FP-Growth is one of the fastest approach and most popular algorithms for frequent itemsets mining. This algorithm is based on a prefix tree representation of database transactions (called FP-tree).

3. Preliminaries

The following part will discuss in detail all the basic terms and terminology used in this paper.

3.1. *Association Rules (ARs)*

ARs were first proposed for market basket analysis to study customer purchasing patterns in retail stores[1]. Recently, it has been applied in various disciplines such as customer relationship management[17], image processing[18]. In general, association rule mining is the process of discovering associations or correlation among itemsets in transaction databases, relational databases and data warehouses. There are two subtasks involved in ARs mining: generate frequent itemsets that satisfy the minimum support threshold and generate strong rules from the frequent itemsets. Let I is a non-empty set such that $I = \{i_1, i_2, \cdots, i_n\}$, and D is a database of transactions where each T is a set of items such that $T \subset I$. An association rule is a form of $A \Rightarrow B$, where $A, B \subset I$ such that $A \neq \phi$, $B \neq \phi$ and $A \cap B = \phi$. The set A is called antecedent of the rule and the set B is called consequent of the rule. An itemset is a set of items. A k-itemset is an itemset that contains k items. An itemset is said to be frequent if the support count satisfies a minimum support count (minsupp). The set of frequent itemsets is denoted as L_k. The support of the ARs is the ratio of transaction in D that contain both A and B (or $A \cup B$). The support is also can be considered as probability $P(A \cup B)$. The confidence of the ARs is the ratio of transactions in D contains A that also contains B. The confidence also can be considered as conditional probability $P(B|A)$. ARs that satisfy the minimum support and confidence thresholds are said to be strong.

3.2. *Correlation Analysis*

A few years after the introduction of ARs, Aggrawal *et al.* in Ref. 19 and Brin *et al.* in Ref. 13 realized the limitation of the confidence-support framework. Many studies have shown that the confidence-support framework alone is insufficient at discovering the interesting ARs. Therefore, the correlation can be used as complimentary measure of this framework. This leads to correlation rules as

$$A \Rightarrow B \quad (\text{supp}, \text{conf}, \text{corr}) \tag{1}$$

The correlation rule is measure based on the minimum support, minimum confidence and correlation between itemsets A and B. There are many correlation measures applicable for ARs. One of the simplest correlation measures is Lift. The occurrence of itemset A is independence of the occurrence of itemset B if $P(A \cup B) = P(A)P(B)$; otherwise itemset A and B are dependence and correlated. The lift between occurrence of itemset A and B can be defined as:

$$\mathrm{lift}(A,B)=\frac{P(A\cap B)}{P(A)P(B)} \tag{2}$$

The equation of (2) can be derived to produce the following definition:

$$\mathrm{lift}(A,B)=\frac{P(B\mid A)}{P(B)} \tag{3}$$

or

$$\mathrm{lift}(A,B)=\frac{\mathrm{conf}\,(A\Rightarrow B)}{\mathrm{supp}\,(B)} \tag{4}$$

The strength of correlation is measure from the lift value. If $\mathrm{lift}(A,B)=1$ or $P(B\mid A)=P(B)$ $\left(\text{or } P(A\mid B)=P(B)\right)$ then B and A are independent and there is no correlation between them. If $\mathrm{lift}(A,B)>1$ or $P(B\mid A)>P(B)$ $\left(\text{or } P(A\mid B)>P(B)\right)$, then A and B are positively correlated, meaning the occurrence of one implies the occurrence of the other. If $\mathrm{lift}(A,B)<1$ or $P(B\mid A)<P(B)$ $\left(\text{or } P(A\mid B)<P(B)\right)$, then A and B are negatively correlated, meaning the occurrence of one discourage the occurrence of the other. Since lift measure is not down-ward closed, it definitely will not suffer from the least item problem. Thus, least itemsets with low counts which per chance occur a few times (or only once) together can produce enormous lift values.

3.3. *Jaccard Similarity Coefficient (Jaccard)*

Jaccard Similarity Coefficient (Jaccard) in Ref. 21 is a statistical index for measuring the similarity and variety of sample sets. It measures the similarity between sample sets, and the size of the intersection divided by the size of the union of the sample sets:

$$\mathrm{Jaccard}(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{5}$$

The idea is to measure the proportion of the union itemsets, compare to availability of antecedent and consequence in the union of the itemsets. In other words, it is the cardinality of intersection of itemsets divided by the cardinality of their union. Mathematically, it can be denoted as:

$$\mathrm{Jaccard}(A,B)=\frac{P(A\cup B)}{P(A)+P(B)-P(A\cup B)} \tag{6}$$

The Jaccard is focused more on binary similarity between the sets by both sides of the rules. The range of the measurement is between 0 and 1. The more values mean the similarity between itemsets is very much closer.

3.4. *IS Measure (IS)*

IS Measure [22] is an alternative statistical index for measuring the asymmetric variables. The measure is defined as follows:

$$IS(A,B) = \sqrt{lift(A,B) \times P(A \cup B)} = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} \tag{7}$$

Mathematically, IS measure is equivalent to the cosine measure for binary variables. Thus, A and B can be defined as a pair of bit vectors,

$A \bullet B = P(A \cup B)$, the dot product of vectors,

$|A| = \sqrt{P(A)}$, the magnitude of vector A, therefore

$$\mathrm{IS}(A,B) = \frac{P(A \cup B)}{\sqrt{P(A) \times P(B)}} = \frac{A \bullet B}{|A| \times |B|} = \mathrm{cosine}(A,B) \tag{8}$$

The IS value of itemset is low whenever one of its rules has low confidence value. For this measure, the value is in the range of 0 and 1.

4. Methodology

4.1. *Critical Relative Support (CRS)*

Throughout this section the set $I = \{i_1, i_2, \cdots, i_{|A|}\}$, for $|A| > 0$ refers to the set of literals called set of items and the set $D = \{t_1, t_2, \cdots, t_{|U|}\}$, for $|U| > 0$ refers to the data set of transactions, where each transaction $t \in D$ is a list of distinct items $t = \{i_1, i_2, \cdots, i_{|M|}\}$, $1 \le |M| \le |A|$ and each transaction can be identified by a distinct identifier TID.

4.1.1. *Definition*

Definition 1. (Least Items). *An itemset X is called least item if* $\alpha \le \mathrm{supp}(X) \le \beta$, *where* α *and* β *is the lowest and highest support, respectively.*
The set of least item will be denoted as Least Items and

$$\text{Least Items} = \{X \subset I \mid \alpha \le \mathrm{supp}(X) \le \beta\}$$

Definition 2. (Frequent Items). *An itemset X is called frequent item if* $\mathrm{supp}(X) > \beta$, *where* β *is the highest support.*

The set of frequent item will be denoted as Frequent Items and

$$\text{Frequent Items} = \{X \subset I \mid \mathrm{supp}(X) > \beta\}$$

Definition 3. (Merge Least and Frequent Items). *An itemset X is called least frequent items if* $\mathrm{supp}(X) \geq \alpha$ *, where* α *is the lowest support.*

The set of merging least and frequent item will be denoted as LeastFrequent Items and

$$\text{LeastFrequent Items} = \{X \subset I \mid \mathrm{supp}(X) \geq \alpha\}$$

LeastFrequent Items will be sorted in descending order and it is denoted as

$$\text{LeastFrequent Items}^{\text{desc}} = \left\{ \begin{array}{l} X_i \left|\mathrm{supp}(X_i) \geq \mathrm{supp}(X_j)\right), \ 1 \leq i, j \leq k, \ i \neq j, \\ k = \left|\text{LeastFrequent Items}\right|, \ x_i, x_j \subset \text{LeastFrequent Items} \end{array} \right\}$$

Definition 4. (Ordered Items Transaction). *An ordered items transaction is a transaction which the items are sorted in descending order of its support and denoted as* t_i^{desc} *, where*

$$t_i^{\text{desc}} = \text{LeastFrequentItems}^{\text{desc}} \bigcap t_i, 1 \leq i \leq n, \left|t_i^{least}\right| > 0, \left|t_i^{frequent}\right| > 0 .$$

An ordered items transaction will be used in constructing the proposed model, so-called LP-Tree.

Definition 5. (Significant Least Data). *Significant least data is one which its occurrence less than the standard minimum support but appears together in high proportion with the certain data.*

Definition 6. (Critical Relative Support). *A Critical Relative Support (CRS) is a formulation of maximizing relative frequency between itemset and their Jaccard similarity coefficient.*

The value of Critical Relative Support denoted as CRS and

$$\mathrm{CRS}(I) = \max\left(\left(\frac{\mathrm{supp}(\mathrm{A})}{\mathrm{supp}(\mathrm{B})}\right), \left(\frac{\mathrm{supp}(\mathrm{B})}{\mathrm{supp}(\mathrm{A})}\right)\right) \times \left(\frac{\mathrm{supp}(A \Rightarrow B)}{\mathrm{supp}(\mathrm{A}) + \mathrm{supp}(\mathrm{B}) - \mathrm{supp}(A \Rightarrow B)}\right)$$

CRS value is between 0 and 1, and is determined by multiplying the highest value either supports of antecedent divide by consequence or in another way around with their Jaccard similarity coefficient. It is a measurement to show the level of CRS between combination of the both Least Items and Frequent Items either as antecedent or consequence, respectively.

4.2. *Algorithm Development*

4.2.1. *Determine Interval Support for Least Itemset*

Let I is a non-empty set such that $I = \{i_1, i_2, \cdots, i_n\}$, and D is a database of transactions where each T is a set of items such that $T \subset I$. An item is a set of items. A k-itemset is an itemset that contains k items. An itemset is said to be least if the support count satisfies in a range of threshold values called Interval Support (ISupp). The Interval Support is a form of ISupp (ISMin, ISMax) where ISMin is a minimum and ISMax is a maximum values respectively, such that $\mathrm{ISMin} \geq \phi$, $\mathrm{ISMax} > \phi$ and $\mathrm{ISMin} \leq \mathrm{ISMax}$. The set is denoted as R_k. Itemsets are said to be significant least if they satisfy two conditions. First, support counts for all items in the itemset must greater ISMin. Second, those itemset must consist at least one of the least items. In brevity, the significant least itemset is a union between least items and frequent items, and the existence of intersection between them.

4.2.2. *Construct Significant Least Pattern Tree*

A Significant Least Pattern Tree (SLP-Tree) is a compressed representation of significant least itemsets. This trie data structure is constructed by scanning the dataset of single transaction at a time and then mapping onto path in the SLP-Tree. In the SLP-Tree construction, the algorithm constructs a SLP-Tree from the database. The SLP-Tree is built only with the items that satisfy the ISupp. In the first step, the algorithm scans all transactions to determine a list of least items, LItems and frequent items, FItems (least frequent item, LFItems). In the second step, all transactions are sorted in descending order and mapping against the LFItems. It is a must in the transactions to consist at least one of the least items. Otherwise, the transactions are disregard. In the final step, a transaction is transformed into a new path or mapped into the existing path. This final step is continuing until end of the transactions. The problem of existing FP-Tree are it may not fit into the memory and expensive to build. FP-Tree must be built completely from the entire transactions before calculating the support of each item. Therefore, SLP-Tree is an alternative and more practical to overcome these limitations.

4.2.3. *Generate Significant Least Pattern Growth (SLP-Growth)*

SLP-Growth is an algorithm that generates significant least itemsets from the SLP-Tree by exploring the tree based on a bottom-up strategy. ‘Divide and conquer’ method is used to decompose task into a smaller unit for mining desired patterns in conditional databases, which can optimize the searching space. The algorithm will extract the prefix path sub-trees ending with any least item. In each of prefix path sub-tree, the algorithm will recursively execute to extract all frequent itemsets and finally built a conditional SLP-Tree. A list of least itemsets is then produced based on the suffix sequence and also sequence in which they are found. The pruning processes in SLP-Growth are faster than FP-Growth since most of the unwanted patterns are already cutting-off during constructing the SLP-Tree data structure. The complete SLP-Growth algorithm is shown in Fig. 1.

```
Read dataset, D
Set Interval Support (ISMin, ISMax)
for items, I in transaction, T do
     Determine support count, ItemSupp
end for loop
Sort ItemSupp in descending order, ItemSuppDesc
for ItemSuppDesc do
     Generate List of frequent items, FItems > ISMax
end for loop
for ItemSuppDesc do
     Generate List of least items, ISMin <= LItems < ISMax
end for loop
Construct Frequent and Least Items, FLItems = FItems U LItems
for all transactions,T do
     if (LItems ∩ I in T > 0) then
           if (Items in T = FLItems) then
                Construct items in transaction
                   in descending order, TItemsDesc
           end if
      end if
end for loop
for TItemsDesc do
     Construct SLP-Tree
end for loop
for all prefix SLP-Tree do
     Construct Conditional Items, CondItems
end for loop
for all CondItems do
     Construct Conditional SLP-Tree
end for loop
for all Conditional SLP-Tree do
     Construct Association Rules, AR
end for loop
for all AR do
     Calculate Support and Confidence
     Apply Correlation
end for loop
```

Fig. 1. SLP-Growth Algorithm

4.3. *Weight Assignment*

4.3.1. *Apply Correlation.*

The weighted ARs (ARs value) are derived from the formula (4). This correlation formula is also known by lift. The processes of generating weighted ARs are taken place after all patterns and ARs are completely produced.

4.3.2. *Discovery Highly Correlated Least ARs*

From the list of weighted ARs, the algorithm will begin to scan all of them. However, only those weighted ARs with correlation value that more than one are captured and considered as highly correlated. For ARs with the correlation less than one will be pruned and classified as low correlation.

5. Experiment Tests

5.1. *Breast Cancer Wisconsin dataset from Ref.* 7

The first experiment was conducted on Breast-Cancer-Wisconsin dataset. The aim of the dataset is to diagnose the breast cancer according to Fine- Needle Aspirates (FNA) test. The dataset was obtained from a repository of a machine-learning database University of California, Irvin. It was compiled by Dr. William H. Wolberg from University of Wisconsin Hospitals, Madison, Madison, WI, United States. It has 11 attributes and 699 records (as of 15 July 1992) with 158 benign and 241 malignant classes, respectively.

Table 2 displays the mapped of original attributes with new attributes id. Item is constructed based on the combination of attribute id and its domain. For simplicity, let consider an attribute "Clump Thickness" with domain "1". Here, an item "101" will be constructed by means of a combination of an attribute id (first two characters) and its domain (third character). Jaccard similarity coefficient and IS Measure (IS) are employed in the experiment for comparison.

To ensure the least ARs are extracted, ISupp is set in a range of 0.00% to 5.00%. By embedding SFP-Growth algorithm with ISupp feature, only 4,082 ARs are produced. ARs are formed by applying the relationship of an item or many items to an item (cardinality: many-to-one). Here, the maximum number of items appears in each ARs is set to 6. Fig. 4 depicted the correlation's classification of least ARs. For this dataset, the rule is categorized as significant if it has positive correlation and CRS should be at least 0.5.

Table 2. The mapped breast cancer data

Attributes	Domain	Attributes Id
Sample code number	Id number	Ignored
Clump Thickness	1 - 10	10
Uniformity of Cell Size	1 – 10	20
Uniformity of Cell Shape	1 – 10	30
Marginal Adhesion	1 – 10	40
Single Epithelial Cell Size	1 – 10	50
Bare Nuclei	1 – 10	60
Bland Chromatin	1 – 10	70
Normal Nucleoli	1 – 10	80
Mitoses	1 – 10	90
Class	2,4	111

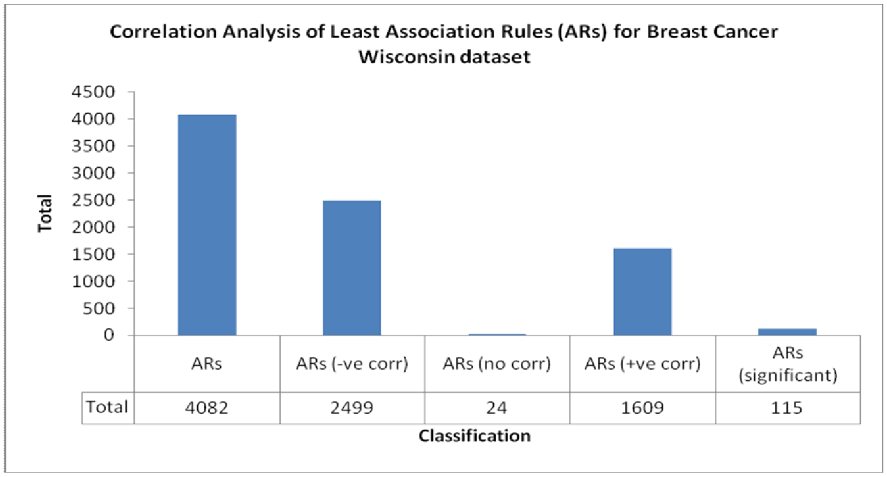
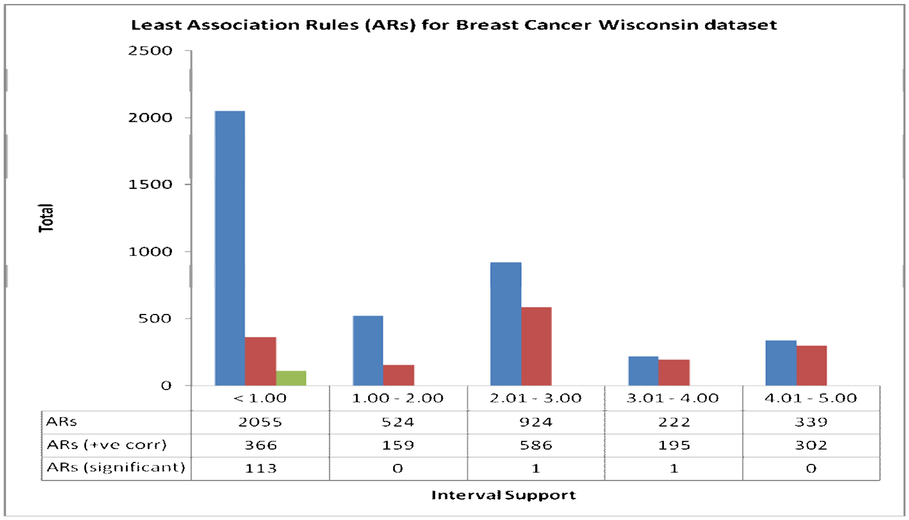


Fig. 4. Classification of ARs using correlation analysis. Only 2.82% from the total of 4,082 ARs are classified as significant least ARs.

Table 3 shows top 10 least ARs with numerous types of measurements. The highest correlation value from the selected ARs is 12.05 (6010 707 2010 → 3010 and 1114 6010 707 2010 → 3010). From these ARs, there are only two dominant of consequence items, item 3010 and 501, respectively. The first item appears more frequent as compared to second item. From the analysis, item 3010 appears 8.29% from the entire dataset. For the second item, it occurs less 1.57% from the first one. Therefore, based on this finding, details analysis and study by physician are recommended to discover the level of significant of these least ARs. It may reveal something interesting and great contribution in the domain knowledge of medicine. Fig. 5 illustrates the summarization of correlation analysis with different ISupp.

Table 3. Top 10 of highest correlation of least association rules with different type of measurements (interval support: 0% – 5.00%)

Association Rules	Supp	Conf	Corr	Interest	Jaccard	CRS
6010 707 2010 --> 3010	1.29	100.00	12.05	12.05	0.16	1.00
1114 6010 707 2010 --> 3010	1.29	100.00	12.05	12.05	0.16	1.00
901 --> 501	6.58	7.94	1.18	1.18	0.08	0.98
1112 --> 501	6.58	10.04	1.49	1.49	0.10	0.98
801 --> 501	6.58	10.38	1.54	1.54	0.10	0.98
901 --> 102	6.87	8.29	1.16	1.16	0.08	0.96
901 1112 --> 501	6.44	10.11	1.50	1.50	0.10	0.95
901 801 --> 501	6.44	10.56	1.57	1.57	0.11	0.95
1112 801 --> 501	6.44	11.19	1.66	1.66	0.11	0.95
1114 --> 506	5.58	16.18	2.76	2.76	0.16	0.94



	< 1.00	1.00 - 2.00	2.01 - 3.00	3.01 - 4.00	4.01 - 5.00
ARs	2055	524	924	222	339
ARs (+ve corr)	366	159	586	195	302
ARs (significant)	113	0	1	1	0

Fig. 5. Correlation analysis of least ARs using variety ISupp. The total numbers of overall ARs are decreased when the predefined ISupp thresholds are increased.

5.2. *Cardiac Single Proton Emission Computed Tomography (SPECT) from Ref.* 8

The second experiment was conducted on Cardiac Single Proton Emission Computed Tomography (SPECT) dataset or also known as SPECT Heart dataset. SPECT is a nuclear medicine technique that uses radiopharmaceuticals to produce images representing slices through the body in different planes. By this technique, the images are functional in nature rather than being purely anatomical such as ultrasound, CT, and MRI. The dataset was obtained from a repository of a machine-learning database University of Colorado, Denver, USA. There are two categories of patient; normal and abnormal. 267 SPECT images were processed in order to extract 44 continuous feature patterns. Further processed was carried out and finally 22 binary feature patterns were extracted. In summary, the dataset had 22 attributes and 267 records. From the total record, 206 patients were classified as abnormal (positive heart diseases) and employed.

Table 4 displays the mapped of original attributes with new attributes id. Item is constructed based on direct mapping between their respective domains. For example, an item "11" represents the partial diagnosis for attribute F1 is "1". If the partial diagnosis for attribute F1 is equal to "0", thus item will not be created. For any attribute that has a value of "0" for their partial diagnosis, it will not be appeared in the record. Here, ISupp is fixed in a range of 3.00% to 8.00%.

By embedding SFP-Growth algorithm with ISupp, 31,710 ARs are produced. As similar to first experiment, ARs are formed by applying the relationship of an item or many items to an item. Here, the maximum number of items appears in each ARs is also set to 6. Fig. 5 depicted the correlation's classification of least ARs. For this experiment, the rule is categorized as significant if it has positive correlation and CRS should be at least 0.5.

Table 7 shows top 10 least ARs with numerous types of measurements. The dominant of consequence items, item 25. From the analysis, item 25 appears 16.85% from the entire dataset. From this finding, details analysis and study by physician are highly recommended to find out the level of significant of these least ARs. It may reveal a new knowledge and very useful in medicine perspective. Fig. 6 illustrates the summarization of correlation analysis with different ISupp.

Table 4. The mapped breast cancer data

Attributes	Domain	Attributes Id
OVERALL_DIAGNOSIS	0,1 (1)	101
F1: (the partial diagnosis 1, binary)	0,1 (1)	11
F2: (the partial diagnosis 2, binary)	0,1 (1)	12
F3: (the partial diagnosis 3, binary)	0,1 (1)	13
F4: (the partial diagnosis 4, binary)	0,1 (1)	14
F5: (the partial diagnosis 5, binary)	0,1 (1)	15
F6: (the partial diagnosis 6, binary)	0,1 (1)	16
F7: (the partial diagnosis 7, binary)	0,1 (1)	17
F8: (the partial diagnosis 8, binary)	0,1 (1)	18
F9: (the partial diagnosis 9, binary)	0,1 (1)	19
F10: (the partial diagnosis 10, binary)	0,1 (1)	20
F11: (the partial diagnosis 11, binary)	0,1 (1)	21
F12: (the partial diagnosis 12, binary)	0,1 (1)	22
F13: (the partial diagnosis 13, binary)	0,1 (1)	23
F14: (the partial diagnosis 14, binary)	0,1 (1)	24
F15: (the partial diagnosis 15, binary)	0,1 (1)	25
F16: (the partial diagnosis 16, binary)	0,1 (1)	26
F17: (the partial diagnosis 17, binary)	0,1 (1)	27
F18: (the partial diagnosis 18, binary)	0,1 (1)	28
F19: (the partial diagnosis 19, binary)	0,1 (1)	29
F20: 0,1 (the partial diagnosis 20, binary)	0,1 (1)	30
F21: 0,1 (the partial diagnosis 21, binary)	0,1 (1)	31
F22: 0,1 (the partial diagnosis 22, binary)	0,1 (1)	32

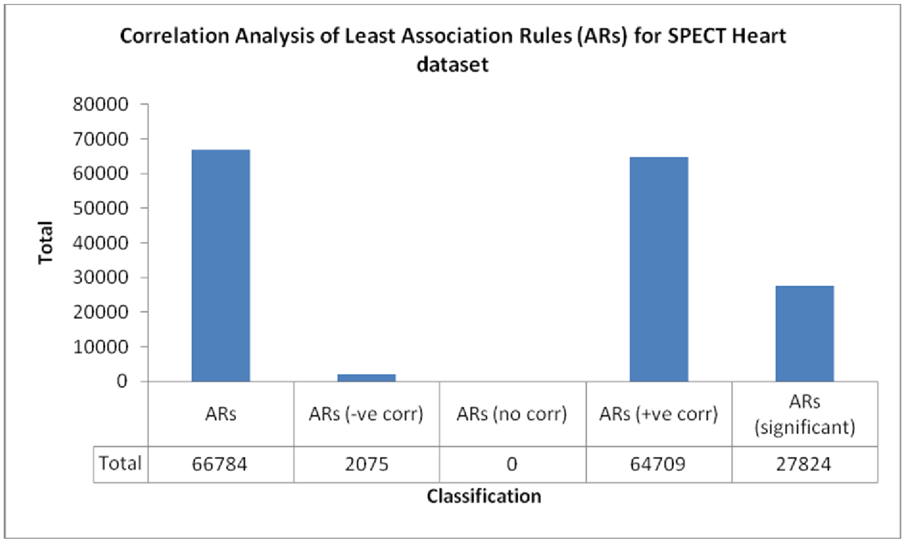
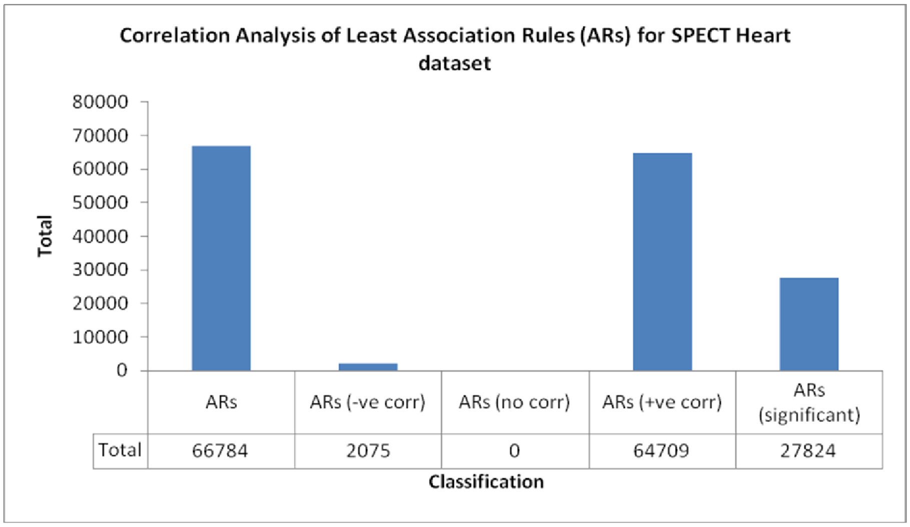


Fig. 6. Classification of ARs using correlation analysis. Only 41.66% from the total of 66,784 ARs are classified as significant least ARs.

Table 5. Top 10 of highest correlation of least association rules with different type of measurements (interval support: 1.00% – 3.50%)

Association Rules	Supp	Conf	Corr	Interest	Jaccard	CRS
32 15 19 12 29 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
11 26 17 14 29 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
23 18 15 19 16 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
15 31 24 21 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
32 15 31 29 16 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
32 13 26 17 29 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
32 20 24 14 29 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
31 20 14 16 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
23 18 11 24 21 --> 25	3.40	100.00	15.53	15.53	0.16	1.00
23 15 19 17 14 --> 25	3.40	100.00	15.53	15.53	0.16	1.00



	ARs	ARs (-ve corr)	ARs (no corr)	ARs (+ve corr)	ARs (significant)
Total	66784	2075	0	64709	27824

Fig. 6. Correlation analysis of least ARs using variety ISupp. The total numbers of overall ARs are decreased when the predefined ISupp thresholds are increased.

6. Conclusion

Mining least ARs is undeniable a very crucial in discovering the rarity or irregularity relationship among itemset in database. It is quite complicated, computationally expensive and absolutely required special measurement in order to capture the least rules. In medical context, detecting the irregularity ARs is very useful and sometime can help save human's life. However, formulating an appropriate measurement and finding a scalable mining algorithm to extract the respective rules are very challenging. Therefore, Critical Relative Support (CRS) measurement and SLP-Growth algorithm are employed in the experiment of medical datasets. Here, Breast Cancer and SPECT Heart datasets have been used for evaluations. The result shows that CRS and SLP-Growth algorithm can be used in detecting the critical least ARs with highly correlated, and thus verify it scalabilities.

Acknowledgement

This work was supported by Research and Innovations Center, Universiti Malaysia Pahang.

References

1. R. Agrawal, T. Imielinski and A. Swami, Database Mining: A Performance Perspective, *IEEE Transactions on Knowledge and Data Engineering,* **5** (**6**), 914 (1993).

2. H. Yun, D. Ha, B. Hwang and K.H. Ryu, Mining Association Rules on Significant Rare Data Using Relative Support*, The Journal of Systems and Software* **67 (3)**, 181-191 (2003).
3. H. Xiong, S.Shekhar, P-N. Tan and V.Kumar, Exploiting A Support-Based Upper Bond Pearson's Correlation Coefficient For Efficiently Identifying Strongly Correlated Pairs, *Proceeding of ACM SIGKDD'04*, 334 (2004).
4. Z. Abdullah, T. Herawan and M.M. Deris, Scalable Model for Critical Least Association Rules, *Lecture Notes in Computer Science* **6377**, (Springer-Verlag, New York 2010).
5. Z. Abdullah, T. Herawan and M.M. Deris, Mining Significant Least Association Rules using fast SLP-Growth Algorithm, *Lecture Notes in Computer Science* **6059** (Springer-Verlag, New York 2010).
6. http://archive.ics.uci.edu/ml/
7. http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
8. http://archive.ics.uci.edu/ml/datasets/SPECT+Heart
9. L. Zhou and S. Yau, Association Rule and Quantitative Association Rule Mining Among Infrequent Items, *The Proceeding of ACM SIGKDD'07*, 156 (2007).
10. J. Ding, Efficient Association Rule Mining Among Infrequent Items, *Ph.D Thesis*, University of Illinois at Chicago (2005).
11. Y.S. Koh and N. Rountree, Finding sporadic rules using apriori-inverse, *Lecture Notes in Computer Science* **3518** (Springer-Verlag, 2005).
12. B. Liu, W. Hsu and Y. Ma, Mining Association Rules With Multiple Minimum Supports, *Proceedings of SIGKDD Explorations*, 337 (1999).
13. S. Brin, R. Motwani, and C. Silverstein, Beyond Market Basket: Generalizing ARs to Correlations, *Proceedings of Special Interest Group on Management of Data (SIGMOD)*, 265 (1997).
14. E. Omniecinski, Alternative Interest Measures For Mining Associations, *IEEE Trans. Knowledge and Data Engineering*, **15**, 57 (2003).
15. Y.-K. Lee, W.-Y. Kim, Y.D. Cai and J. Han, CoMine: Efficient Mining Of Correlated Patterns, *Proceeding of ICDM'03*, 569 (2003).
16. J. Han, H. Pei, and Y. Yin, Mining Frequent Patterns Without Candidate Generation: A Frequent Pattern Tree Approach, *Data Mining and Knowledge Discovery*, **8**, 53 (2004).
17. J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Series in Data Management Systems, 2nd edn. (Morgan Kaufmann, San Francisco, 2006).
18. W. H. Au and K.C.C. Chan, Mining Fuzzy ARs In A Bank-Account Database, *IEEE Transactions on Fuzzy Systems* **11 (2)**, 238 (2003).
19. C.C. Aggarwal and P.S. Yu, A New Framework For Item Set Generation, *Proceedings of the ACMPODS Symposium on Principles of Database Systems*, Seattle, Washington, 18 (1998).
20. H. Li, Y. Wang, D. Zhang, M. Zhang and E.Y. Chang, Pfp: Parallel Fp-Growth For Query Recommendation, *Proceedings of RecSys'2008*, 107 (2008).
21. D. Roussinov, and J.L. Zhao, Automatic Discovery of Similarity Relationships Through Web Mining, *Decision Support Systems*, **25**, 149 (2003).
22. T. Brijs, K. Vanhoof, and G. Wets, Defining Interestingness for Association Rules, *International Journal of Information Theories & Applications*, **10 (4)**, 370 (2003).