Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets

Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar, Anis Suhailis Abdul Kadir
Data Mining and Optimization Research Group, Center for Artificial Intelligence Technology,
Faculty of Information Science and Technology, University Kebangsaan Malaysia, 43600 Bangi,
Selangor Darul Ehsan, MALAYSIA

dhubaswesi@yahoo.com, aab@ftsm.ukm.my, anisu@yahoo.com

***Abstract*—Association rule mining is one of the most important tasks in data mining. The basic concept of association rules is to mine the interesting (positive) frequent patterns from a transaction database. However, mining the negative patterns has also attracted the attention of researchers in this area. The aim of this study is to develop a new model for mining interesting negative and positive association rules out of a transactional data set. The proposed model is an integration between two algorithms, the Positive Negative Association Rule (PNAR) algorithm and the Interesting Multiple Level Minimum Supports (IMLMS) algorithm, to propose a new approach (PNAR_IMLMS) for mining both negative and positive association rules from the interesting frequent and infrequent itemsets mined by the IMLMS model. The experimental results show that the PNAR_IMLMS model provides significantly better results than the previous model.**

Keywords- Negative association rule, frequent itemset, infrequent itemset.

I. INTRODUCTION

The mining association rule is a data mining task that aims to discover relationships among items in a transactional database [1]. This task has been studied widely in the literature for its benefit in many application domains, such as Web usage mining, recommender systems, and intrusion detection. It deals with the market basket analysis to find frequent patterns and generate association rules of the form A⇒B, which can be used to predict that "If a customer buys itemset A, he/she will most likely buy itemset B as well", where A and B are frequent itemsets in a transaction database and $A \cap B=\emptyset$. The rule of the form A⇒B is considered a strong rule if its support (s) and confidence (c) satisfy minimum support (mins) and minimum confidence (minc) thresholds; these are called positive association rules (PARs). In addition to PARs, the relationship represented as "customers that buy itemset A do not buy itemset B" refers to a negative relation between two itemsets. The forms of rules A⇒¬B, ¬A⇒B and ¬A⇒¬B are called negative association rules (NARs) [10]. Previous research has shown that NARs can be discovered from both frequent (FIS) and infrequent itemsets (inFIS). The infrequent itemsets are significant because there are many useful NARs in them; however, few scholars have discussed how to discover negative association rules among the infrequent items. Moreover, some of the discovered itemsets are not interesting and some of the association rules might be redundant. Therefore, a pruning strategy is essential to reduce the search space and guarantee the efficiency of the association rule algorithms. In this paper, we develop an integrated model to discover both interesting FISs and inFISs, which we call the PNAR_IMLMS model. The PNAR_IMLMS is based on the IMLMS model introduced in [11]. The positive and negative association rules are discovered from the FISs and inFISs, respectively. In addition, a measure discussed in [8] that combines the correlation coefficient and minimum confidence called Valid Association Rules based on Correlation, coefficient and Confidence (VARCC) is incorporated into the proposed algorithm.

The main contribution of this paper is to integrate Dong's (2007) [8] algorithm for mining the Positive and Negative Association Rules (PNAR) from frequent and infrequent itemsets with the modified pruning strategy Interesting Multiple Level Minimum Supports (IMLMS) [9]. We call our algorithm PNAR_IMLMS.

The rest of the paper is organized as follows: Section 2 discusses related work on mining positive and negative association rules and frequent and infrequent item sets. The preliminaries of the Interesting Multiple Level Minimum Supports (IMLMS) and Positive and Negative Association Rules (PNAR) are explained in Section 3. The integration of PNAR and IMLMS is proposed in Section 4. Section 5 presents the experimental results of the work, and concluding remarks are presented in the final section.

II. RELATED WORK

The purpose of association rule mining is to find certain associations between a set of items in a database. The task is first introduced in the literature by [1], who notes that it is concerned with patterns that occur frequently in a given database, while sometimes the decision makers concern themselves with items that occur infrequently but are strongly correlated. Therefore, negative association rules are very important to many applications, particularly in competitive analysis. Several studies have been performed on mining association rules. For instance, a statistical test called chi-squared can be used on the negative associations

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

between two items to verify the relationships between the two variables. Since then, there have been several attempts to solve the problem of NAR mining [2]. Negative association rules can be treated in the forms A⇒B, A⇒¬B, ¬A⇒B and ¬A⇒¬B.

Savasere et al. (1998) describe the strong negative association rules [3]. They incorporate frequent itemsets with domain knowledge in the form of taxonomy to mine negative associations. However, their approach is hard to generalize because it is domain dependent and needs a predefined taxonomy. Negative association rules can detect some problems that positive rules cannot, but contradictory information occurs when both types of rules are mined simultaneously. Luo and Bo (2010) solve this problem by designing a new mining algorithm for positive and negative association rules. In addition, their algorithm deletes those contradictory rules based on the correlation between itemsets [5].

Another term is derived from the concept of association rules, particularly quantitative association rules, called linguistic association rules. The work in [13] assigned a set of standards consisting of two heuristic rules to analyze association rules, including quantitative linguistic terms.

Several studies have been conducted on mining positive and negative association rules [5-8]. Dong et al. (2006) presents an approach for mining positive and negative association rules based on multiple confidences and the chi-squared test [6]. The method solves the problem caused with a single confidence threshold and the chi-squared problem of mining uninteresting rules. Although the chi-squared test is effective, it has some limitations: it does not provide enough information about the strength of the relationship, and it is sensitive to small expected values in one or more of the cells in the contingency table.

The work presented by [5-6] did not consider the negative rules extracted from the infrequent item sets. This work can be improved by considering the generation of positive association rules (PARs) from frequent itemsets and NARs from infrequent itemsets [7]. An additional measure called the correlation coefficient is considered on top of the original support-confidence framework. However, it suffers from its inability to generate a complete set of valid NARs. Dong et al. (2006) discusses mining association rules using multiple confidences, and several studies have addressed the issue of mining association rules using Multiple Level Minimum Supports (MLMS) [9-11].

Dong et al. (2007) propose an algorithm called PNAR_MLMS to generate PARs and NARs from frequent and infrequent itemsets using the Multiple Level Minimum Supports (MLMS) model [8]. A modified measure of VARCC combining the correlation coefficient and minimum confidence is presented. Although the measure VARCC can eliminate some of the misleading rules mined by the algorithm, the number of generated rules is still too large to be chosen readily. However, some of the FISs and inFISs obtained by the MLMS are not interesting and should be pruned. The Apriori_IMLMS is proposed to discover both interesting frequent and infrequent itemsets based on the interesting MLMS (IMLMS) model. This model modifies the pruning strategy used in [10] for pruning the uninteresting itemsets. Further, the authors [8-11] use different methods with the modified pruning strategy, which makes their algorithm more efficient, and fewer FISs and inFISs are generated.

III. MATERIALS AND METHODS

In this section, we review the concept of Positive and Negative Association Rules (PNARs) [8] and the Interesting Multiple Level Minimum Supports (IMLMS) [9,11] that will be integrated and implemented in this study.

A. Interesting Multiple Level Minimum Supports (IMLMS) Model

The Interesting Multiple Level Minimum Supports (IMLMS) approach aims to find interesting FISs and inFISs [9]. The approach was designed by adopting the original MLMS model with a pruning strategy, which is modified to prune uninteresting itemsets in a manner that is suitable to the model.

Let $I= \{i_1, i_2, \ldots, i_n\}$ be a set of n distinct attributes called items and $TD = \{t_1, t_2, \ldots, t_m\}$ a set of transactions over I. The number of transactions in TD is indicated as $|TD|$ where each transaction T has a unique identifier called TD, and each transaction contains a set of items such that $T \subseteq I$. A set of items from I is called an itemset. Given itemsets A and B, the number of items contained in itemset A is called its length, identified by len(A). Each itemset has two basic measures, support and confidence, denoted by s and c, respectively. For an itemset $A \subseteq I$, the support is $A.count$ / $|TD|$, where $A.count$ is the number of transactions in TD that contain the itemset A. The support of a rule $A \Rightarrow B$ is denoted as $s(A \cup B)$, where $A, B \subseteq I$, and $A \cap B = \Phi$ while the confidence of the rule $A \Rightarrow B$ is defined as the proportion of $s(A \cup B)$ above $s(A)$, i.e., $c(A \Rightarrow B) = s(A \cup B) / s(A)$.

The MLMS model assigns various minimum supports to itemsets with different lengths. Assume $ms(k)$ is the minimum support of k-itemsets $(k=1,2,\cdots,n)$, which are the thresholds for FIS, $ms(0)$ is a threshold for inFIS, $ms(1) \geqslant ms(2) \geqslant, \cdots, \geqslant ms(n) \geqslant ms(0) > 0$. For any itemset A, if $s(A) \geqslant ms(k)$, then A is a frequent itemset. If $s(A) < ms(k)$ and $s(A) \geqslant ms(0)$, then A is an infrequent itemset. However, the number of FISs and inFISs can be easily constrained in the MLMS model because it allows the user to assign appropriate values for $ms(k)$ and $ms(0)$, which can be specified by users or experts.

Pruning Strategy

Because too many FISs related to PARs and too many inFISs related to NARs are not interesting, pruning is important in the search for interesting itemsets. The pruning strategy introduced in [10] aims to reduce the number of FISs and inFISs. The pruning strategy discussed in [10] prunes the uninteresting itemsets that consider the support,

confidence, and interest of a rule. However, it is limited to a single minimum support. The pruning strategy in [10] is modified to be suitable to the MLMS model. Therefore, the MLMS model with the modified pruning strategy is called IMLMS [11]. Various methods have been used with the modified pruning strategy to reduce the number of FISs and inFISs by pruning the uninteresting ones. The following formula is utilized to prune uninteresting frequent itemsets [10]:

M is considered a potentially frequent itemset of interest(*int*) if $s(M) \geq ms(len(M))$ and $f(A,B, ms(len(A \cup B)), m_i) = 1$ where $len(A)$ is the number of items in an itemset A and $f(\cdot)$ is a constraint function concerning the support and interest of the rule $A \Rightarrow B$, where $int(A,B) = |supp(A \cup B) - supp(A)*supp(B)|$ and m_i is the minimum interest threshold determined by users or experts. Eq(1) is used to control and reduce the number of FISs and Eq(2) is used to reduce the number of inFISs [11].

$$f(A,B,ms(len(A\cup B)),m_i)=\frac{s(A\cup B)+\text{int}(A,B)-(ms(len(A\cup B))+m_i)+1}{|s(A\cup B)-ms(len(A\cup B))|+|\text{int}(A\cup B)-m_i)|+1} \quad (1)$$

$$f(A,B,ms(0),m_i i=\frac{s(A\cup B)+\text{int}(A,B)-(ms(0)+mi)+1}{|s(A\cup B)-ms(0)|+|\text{int}(A\cup B)-mi)|+1} \quad (2)$$

In general, an itemset N is considered a potentially infrequent itemset of interest if $s(N) < ms(length(N)), s(N) \geq ms(0)$ and $f(A,B, ms(0), mi) = 1$.

The generation of FISs and inFISs utilizes the general Apriori algorithm. Besides using the algorithm of Dong et al. (2007) in which PNARs are generated with the Apriori_MLMS model, the Apriori_IMLMS performs the steps of the following algorithm:

Step 1. Initialize both sets of FISs and inFISs with an empty set.
Step 2. Generate all the sets of and with length *1*-itemset that meet the support thresholds *ms(1)* and *ms(0),* respectively.
Step 3. Generate all the sets of and with length *k*-itemset that meet the minimum support thresholds *ms(k)* and *ms(0),* respectively, by a loop, where *k* is greater than or equal to 2-itemset.
Step 4. Prune all uninteresting *k*-itemsets in frequent itemsets (FIS_k) using Eq(1).
Step 5. Prune all uninteresting k-itemsets in infrequent itemsets ($inFIS_k$) by Eq(2).
Step 6. Mine both interesting FISs and interesting inFISs with different lengths.
Step 7. Output the results, which include the sets of interesting FISs and inFISs.

The algorithms Apriori_IMLMS and Apriori_MLMS [8] are the same except for the steps (4) and (5) for pruning uninteresting FISs and inFISs in the IMLMS model, which makes the IMLMS model more efficient than the MLMS model.

B. Positive and Negative Association Rules (PNAR)

PNARs may cause some problems, such as how to find a moderate degree of infrequent itemsets, how to discover PNARs properly, how to deal with the problem caused by single minimum support and so forth. We employ the PNAR and IMLMS for mining PARs from FISs and NARs from both FISs and inFISs discovered by the IMLMS model. NARs of forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$, and $\neg A \Rightarrow \neg B$ are referred to as negative associations between itemsets when their support and confidence are lower than the user-specified minimum support (ms) and minimum confidence (mc) thresholds, respectively. The support and confidence of such rules are difficult to calculate in a direct way. However, we can compute them using positive association rules, as discussed in [6]. Despite these factors, the support-confidence framework is the most popular approach used in association rules mining for pruning associations among items in a database. Many uninteresting rules may be generated, especially when mining PNARs instantaneously. Several important measures discussed in [8] are adopted, such as the Correlation Coefficient (Corr) and the Valid Association Rule based on Correlation coefficient and Confidence (VARCC).

Correlation Coefficient

The correlation coefficient measures the strength and direction of the linear relationship between a pair of two variables. It is also known as the covariance between two variables, divided by their standard deviation (σ), as follows:

$$corr_{AB} = \frac{\text{cov}(A,B)}{\sigma_A \sigma_B} \quad (3)$$

where *cov(A,B)* represents the covariance of two variables and σ represents the standard deviation. The range of values for the correlation coefficient is between -1 and +1. When $corr_{AB} = +1$, then variables A and B are independent. If $corr_{AB} > 1$, then variables *A* and *B* are positive correlated. Similarly, when $corr_{AB} = -1$, then variables *A* and *B* are negative correlated.

Statistically, the strength of the correlation is expressed by the variable $\alpha(0 \leq \alpha \leq 1)$, according to [12]. Cohen (1998) determines that if α=0.5, the strength is large, 0.3 is moderate, and 0.1 is small. This can be interpreted as a rule in which a correlation smaller than 0.1 is trivial. Therefore, we use $corr_{AB} \geq \alpha$ $(0 \leq \alpha \leq 1)$ as a constraint to prune needless rules. The correlation coefficient can be associated with the association rules, and the $corr_{AB}$ between *A* and *B* is defined as follows [8]:

$$corr_{AB} = \frac{s(A \cup B) - s(A)s(B)}{\sqrt{s(A)(1-s(A))(s(B)(1-s(B))}} \quad (4)$$

The equation (4) considers the correlations among the itemsets A and B; however, theorem [8] mentions that if $corr_{AB} \geq \alpha$ ($0 \leq \alpha \leq 1$), then $corr_{\neg AB} \leq -\alpha$, $corr_{A\neg B} \leq -\alpha$, $corr_{\neg A\neg B} \geq \alpha$. The reverse is also true. Eq(4) can be utilized to measure the strength of correlation for the four rules A $\Rightarrow$B, A $\Rightarrow\neg$B, $\neg$A $\Rightarrow$B, and $\neg$A $\Rightarrow\neg$B; we need only to calculate $corr_{AB}$ in the PNAR_IMLMS algorithm using Eq(5). Two conclusions can be made: (i) if $corr_{AB} \geq \alpha$, then the rules of forms $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ will be mined, and (ii) if $corr_{AB} \leq \alpha$, then the rules of forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ will be mined [8].

$$corr_{\neg AB} = \frac{s(\neg A \cup B) - s(\neg A)s(B)}{\sqrt{s(\neg A)(1-s(\neg A))(s(B)(1-s(B))}}$$

$$= \frac{-(s(A \cup B) - s(A)s(B))}{\sqrt{s(A)(1-s(A))(s(B)(1-s(B))}} = -corr_{AB} \leq -\alpha \quad (5)$$

Measure VARCC

VARCC (Valid Association Rule based on Correlation coefficient and Confidence) combines the correlation coefficient *corr* and the minimum confidence *mc*. The rule $A \Rightarrow B$ is considered a valid association rule if it satisfies the condition *VARCC (A, B, α, mc) =1*, where

$$VARCC(A, B, \alpha, mc) = \frac{corr_{AB} - \alpha + c(A \Rightarrow B) - mc + 1}{|corr_{AB} - \alpha| + |c(A \Rightarrow B) - mc| + 1} \quad (5)$$

Eq(5) can be improvised to compute the VARCC for $\neg$A, $\neg$B or $\neg$A$\neg$B cases [8].

IV. PNAR_IMLMS ALGORITHM

We propose the integration of PNAR_IMLMS for mining PARs from FISs and NARs from both FISs and inFISs discovered by the IMLMS model. The PNAR_IMLMS takes advantage of both PNAR rules generation [8] and the IMLMS model [11]. It extracts both PNARs of interest with a correlation coefficient and measure VARCC, using the following steps.

Step 1. Call the IMLMS.

Step 2. Initialize both sets of PARs and NARs with an empty set.

Step 3. Generate all PARs and NARs from frequent itemsets. Within this step, the correlation coefficient is calculated using Eq(4).

3.1. Compare the correlation coefficient with correlation strength (α).

3.2. Generate the rules of forms $A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ if the correlation coefficient is greater than or equal to α and if they meet the conditions *VARCC(A,B,α,mc)=1* and *VARCC($\neg$A,$\neg$B,α,mc)=1* using equations (5) and (8), respectively.

3.3. Generate the rules of forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ if the correlation coefficient is lower than or equal to $-\alpha$ and if they meet the conditions *VARCC(A,$\neg$B,α,mc)=1* and *VARCC($\neg$A,B,α,mc)=1* using equations (6) and (7), respectively.

Step 4. Generate all NARs from infrequent itemsets. Similarly, in this step, the correlation coefficient is calculated using Eq (4).

4.1 Compare the correlation coefficient with correlation strength (α).

4.2 Generate the rules of form $\neg A \Rightarrow \neg B$ if the correlation coefficient is greater than or equal to α and if it meets the condition *VARCC($\neg$A,$\neg$B,α,mc)=1*.

4.3 Generate the rules of forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ if the correlation coefficient is lower than or equal to $-\alpha$ and if they meet the conditions *VARCC(A,$\neg$B,α,mc)=1* and *VARCC($\neg$A,B,α,mc)=1*.

Step 5. Output the results, which include all valid PARs and NARs.

V. EXPERIMENTAL RESULTS

This section will discuss the experimental results to demonstrate the performance of the PNAR_IMLMS for mining both PARs and NARs. All the experiments were performed on a PC Intel Pentium dual core with 1.73 GHZ of CPU, running on a Windows XP operating system and 2 GB of memory. All programs are implemented under the Java compiler, version 1.6. We test and verify the usability of our approach on five datasets from UCI, which involve heart disease, iris, breast cancer, hearing loss and wine. A summary of the datasets' statistical information is depicted in Table 1.

TABLE 1: Characteristics of Datasets

Dataset	No. of Attributes	No. of Instances	Classes	Missing values
Heart disease	14	303	2	6
Hearing loss	9	500	4	0
Breast cancer	11	699	2	16
Iris	5	150	3	0
Wine	13	178	3	0

Because PNAR_IMLMS is employed to mine PARs and NARs among interesting FISs and inFISs with different lengths, it will be compared with the algorithm PNAR_MLMS for mining PARs and NARs from FISs and inFISs. The results are shown in Table 2 through Table 6, where the number of positive association rules is expressed as P while the number of negative association rules is

expressed in three variables: N1 is expressed as $A \Rightarrow \neg B$, N2 is expressed as $\neg A \Rightarrow B$ and N3 is expressed as $\neg A \Rightarrow \neg B$. T is referred to as the total number of rules. In all the following experiments, the length is $k=4$ and Avg Conf refers to the average confidence value for the rules in N1, N2 and N3.

TABLE 2: The numbers of PNARs for the Iris Dataset

Itemset length	PNAR_MLMS P	N1	N2	N3	T	Avg Conf	PNAR_IMLMS P	N1	N2	N3	T	Avg Conf
K=2 FIS	8	0	0	8	16	0.48	8	0	0	8	16	0.55
inFIS	-	21	21	28	70		-	12	12	18	42	
K=3 FIS	18	0	0	18	36	0.52	18	0	0	18	36	0.56
inFIS	-	24	24	24	72		-	12	12	6	30	
K=4 FIS	11	0	0	11	22	0.53	11	0	0	11	22	0.59
inFIS	-	12	12	9	33		-	5	5	0	10	
T	37	57	57	98	249	0.51	37	29	29	50	156	0.56

TABLE 3: The numbers of PNARs for the Hearing Loss Dataset

Itemset length	PNAR_MLMS P	N1	N2	N3	T	Avg Conf	PNAR_IMLMS P	N1	N2	N3	T	Avg Conf
K=2 FIS	32	5	8	26	71	0.55	6	0	0	6	12	0.71
inFIS	-	51	43	31	125		-	5	5	6	16	
K=3 FIS	80	13	23	54	170	0.642	0	0	0	20	46	0.79
inFIS	-	103	105	81	289		-	7	7	9	23	
K=4 FIS	216	35	50	86	387	0.752	0	0	0	20	40	0.93
inFIS	-	72	78	63	213		-	1	1	4	6	
T	328	279	307	341	1225	0.684	6	13	13	65	137	0.82

TABLE 4: The numbers of PNARs for the Breast Cancer Dataset

Itemset length	PNAR_MLMS P	N1	N2	N3	T	Avg Conf	PNAR_IMLMS P	N1	N2	N3	T	Avg Conf
K=2 FIS	39	0	0	44	83	0.52	33	0	0	44	77	0.64
inFIS	-	58	56	84	198		-	17	17	24	58	
K=3 FIS	160	3	2	166	331	0.69	104	0	0	117	221	0.87
inFIS	-	74	73	184	331		-	6	9	37	52	
K=4 FIS	474	45	65	530	1114	0.81	265	0	0	294	559	0.97
inFIS	-	22	0	90	112		-	0	0	11	11	
T	673	202	196	1098	2169	0.74	402	23	26	527	978	0.90

TABLE 5: The numbers of PNARs for the Heart Disease Dataset

Itemset length	PNAR_MLMS P	N1	N2	N3	T	Avg Conf	PNAR_IMLMS P	N1	N2	N3	T	Avg Conf
K=2 FIS	32	14	15	31	92	0.48	8	3	1	9	21	0.49
inFIS	-	88	96	152	336		-	16	21	15	52	
K=3 FIS	96	31	35	103	265	0.56	49	1	1	53	104	0.59
inFIS	-	151	157	414	722		-	21	22	45	88	
K=4 FIS	571	109	123	583	1386	0.65	142	3	3	146	294	0.71
inFIS	-	0	0	0	0		-	0	0	0	0	
T	699	393	426	1283	2801	0.59	199	44	48	268	559	0.64

TABLE 6: The numbers of PNARs for the Wine Dataset

Itemset length	PNAR_MLMS P	N1	N2	N3	T	Avg Conf	PNAR_IMLMS P	N1	N2	N3	T	Avg Conf
K=2 FIS	37	12	9	31	89	0.53	20	5	4	17	46	0.60
inFIS	-	47	58	105	210		-	12	14	44	70	
K=3 FIS	205	20	14	192	431	0.68	149	6	4	144	303	0.72
inFIS	-	61	78	282	421		-	25	28	148	201	
K=4 FIS	915	58	56	940	1969	0.79	657	18	15	668	358	0.82
inFIS	-	0	0	0	0		-	0	0	0	0	
T	1157	198	215	1550	3120	0.74	826	66	65	021	978	0.78

The experimental results expressed in Table 2 through Table 6 show the number of PNARs generated from frequent and infrequent itemsets with different lengths. These rules are mined with two association rule algorithms, the PNAR_MLMS algorithm [8] and the PNAR_IMLMS algorithm, at different multiple minimum supports, when $ms(1)=0.4$, $ms(2)=0.3$, $ms(3)=0.2$, $ms(4)=0.1$, $ms(0)=0.05$, $\alpha=0$, $mi=0.05$ and $mc=0.2$. It can be observed that the algorithm PNAR_IMLMS can successfully generate fewer rules than PNAR_MLMS over the five datasets. For example, in Table 3 the numbers of PARs and NARs mined by PNAR_MLMS are 328 and 927, respectively, whereas the numbers of PARs and NARs generated by PNAR_IMLMS are 46 and 91, respectively. This reveals that there are 282 PARs and 836 NARs detected and removed by PNAR_IMLMS. This important reduction is due to the effective pruning strategies in the pruning method proposed in [10] and used with the IMLMS model for generating fewer FISs and inFISs, as well as the measure VARCC for generating fewer PNARs. This makes the algorithm PNAR_IMLMS an efficient approach for mining PARs and NARs in a suitable degree. In addition, it is worth noting that most of the NARs are mined from infrequent itemsets, which reflects the importance of the inFISs for mining NARs.

Moreover, all the results show that the average confidence of the rules mined by the proposed algorithm is higher than those mined by the compared algorithm for all datasets. This is because our proposed approach is based on the IMLMS model, which generates only interesting FISs and inFISs, which are, in turn, fewer in number due to the pruning strategy used in this model. Therefore, a few itemsets will give us fewer rules with high confidence and high average confidence as well.

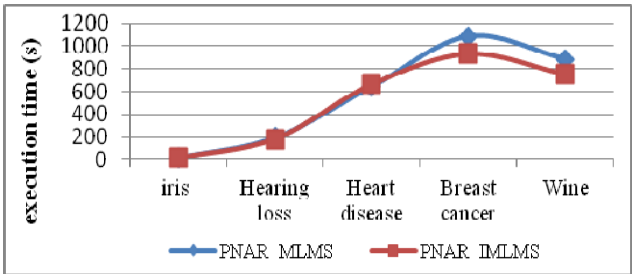

Fig. 1. Execution time of PNAR_MLMS and PNAR_IMLMS

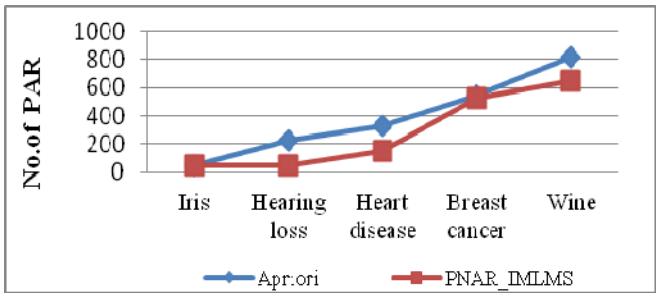

Fig. 2. PARs generated by PNAR_IMLMS and Apriori

From the above experiments, we can see that the algorithm PNAR_IMLMS is more efficient than PNAR_MLMS in mining fewer rules; it also has a low execution time, as shown in Figure 1. Furthermore, Figure 2 shows that the number of PARs mined with the PNAR_IMLMS algorithm is smaller than the number of PARs mined with Apriori when *ms(k)=ms(0)*. This is due to the use of the measure VARCC, which can delete meaningless rules mined by the Apriori algorithm.

VI. Conclusions and Future Work

As infrequent itemsets become more significant for mining the negative association rules that play an important role in decision making, this study proposes a new algorithm for efficiently mining positive and negative association rules in a transaction database. The algorithm is called PNAR_IMLMS and is appropriate for mining positive association rules from frequent itemsets and negative association rules from both frequent and infrequent itemsets discovered by the IMLMS model. The IMLMS model adopted an effective pruning method to prune uninteresting itemsets. An interesting measure VARCC is applied that avoids generating uninteresting rules that may be discovered when mining positive and negative association rules.

References

[1] Agrawal, R., Imielinski, T., Swami A .: Mining association rules between sets of items in large databases. Proceeding of the ACM SIGMOD Intl. Conf. on Management of Data 22(2): 207-216 (1993)

[2] Brin, S., Motwani, R., Silverstein, C.: Beyond market basket: generalizing association rules to correlations. Proceeding of the ACM SIGMOD Conference 26(2) 265-276(1997)

[3] Savasere, A., Omiecinski, E., Navathe, S.: Mining for Strong Negative Associations in a Large Database of Customer Transaction. Proceedings of the Fourteenth International Conference on Data Engineering. 494–502 (1998)

[4] Dong, X., Wang, S., Song, H.: Study of Negative Association Rules. Beijing Institute of Technology Journal 24(11): 978-981(2004)

[5] Luo, J., Bo, Z .: Research on mining positive and negative association rules. International Conference on Computer and Communication Technologies in Agriculture Engineering CCTAE .302-304(2010)

[6] Dong, X., Sun, F., Han, X., Hou, R .: Study of Positive and Negative Association Rules Based on Multi-confidence and Chi-Squared Test. Lecture notes in computer science 4093(LNCS) 100–109(2006)

[7] Antonie, M.-L., Zaiane, O.: Mining Positive and Negative Association Rules: An Approach for Confined Rules. Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases 3202. 27–38(2004)

[8] Dong, X., Niu, Z., Shi, X., Zhang, X., Zhu, D.: Mining both Positive and Negative Association Rules from Frequent and Infrequent Itemsets. ADMA 2007, LNAI 4632, Springer-Verlag Berlin Heidelberg 122–133(2007)

[9] Dong, X., Zheng, Z., Niu, Z., Jia, Q.: Mining Infrequent Itemsets based on Multiple Level Minimum Supports. Proceedings of the Second International Conference on Innovative Computing, Information and Control (ICICIC 2007)(2007).

[10] Wu, X., Zhang, C., Zhang, S.: Efficient Mining of both Positive and Negative Association Rules. ACM Transactions on Information Systems 22(3):381–405(2004)

[11] Dong, X., Niu, Z., Zhu, D., Zheng, Z., Jia,Q.: Mining Interesting Infrequent and Frequent Itemsets Based on MLMS Model. The Fourth International Conference on Advanced Data Mining and Applications, ADMA 5139.444-451(2008)

[12] Cohen, J.: Statistical Power Analysis for the Behavioral Sciences, 2nd edn. Lawrence Erlbaum, New Jersey(1998)

[13] Choo,Y.H., Bakar, A. A.., Hamdan, A. R .: Linguistic Association Rules Mining. Proceedings of the International Conference on Electrical Engineering and Informatics Institut Teknologi Bandung, Indonesia.(2007)

[14] Yuan, X., Buckles, B.P., Yuan, Z., Zhang, J.: Mining Negative Association Rules. In: Proceedings of The Seventh IEEE Symposium on Computers and Communications, 623–629(2002)

[15] Wu, X., Zhang, C., Zhang, S.: Mining both Positive and Negative Association Rules. In: Proceedings of the Nineteenth International Conference on Machine Learning. 658–665(2002)

[16] Liu, B., Hsu, W., Ma, Y.: Mining Association Rules with Multiple Minimum Supports. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 337–341(1999).

[17] Yun-Huoy Choo, Bakar A.A., Hamdan A.R. 2008. The Fitness-rough: A New Attribute Reduction Method Based on Statistical and Rough Set Theory, Intelligent Data Analysis. Vol 12(1). ISI Index. IOS Press. IMPACT FACTOR 0.929.