24th International Conference on Business Information Systems (BIS 2021)
Artificial Intelligence
https://doi.org/10.52825/bis.v1i.40

Published: 02 July 2021

Generating a Condensed Representation for Positive and Negative Association Rules

A Condensed Representation for Association Rules

Bemarisika Parfait[1][bemarisika7@yahoo.fr], and Totohasina André[1][andre.totohasina@gmail.com]

[1]ENSET, Université d'Antsiranana, Madagascar

Abstract. Given a large collection of transactions containing items, a basic common association rules problem is the huge size of the extracted rule set. Pruning uninteresting and redundant association rules is a promising approach to solve this problem. In this paper, we propose a Condensed Representation for Positive and Negative Association Rules representing non-redundant rules for both exact and approximate association rules based on the sets of frequent generator itemsets, frequent closed itemsets, maximal frequent itemsets, and minimal infrequent itemsets in database $\mathcal{B}$. Experiments on dense (highly-correlated) databases show a significant reduction of the size of extracted association rule set in database $\mathcal{B}$.

Keywords: Association rules, Generator itemsets, Closed itemsets, Maximal itemsets, Minimal infrequent itemsets.

1 Introduction and Motivations

Positive and negative association rules (PNAR) mining have been studied extensively in Data mining problem. Let X and Y be two disjoints itemsets, an association rule $X \rightarrow Y$ states that a significant proportion in database $\mathcal{B}$ containing items in the premise (or antecedent) X also contain items in the consequent (or conclusion) Y. This rule can indicate the positive relations between different items, is called positive association rule (PAR) in database $\mathcal{B}$. the association rule at other three forms $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$ and $\overline{X} \rightarrow \overline{Y}$, which can indicate the negative relations between items in database $\mathcal{B}$, are called negative association rules (NAR) in database $\mathcal{B}$.

A basic common association rules problem is the huge number of association rules generated many of which are uninteresting (Definition 1) and redundant (Definition 2). Many approaches [13], [14], [16], based on traditional measure confidence [1], has been developed for reducing the size of the extracted rule set. However, no method to prune uninteresting association rules (UAR) has been found in the literature. Indeed, this classic measure confidence is not efficient to prune uninteresting rules. In addition, these approaches are insufficient, because they consider only the positive association rules, and this, with less selective pair support-confidence [1]. Therefore, discovering NAR, which can be interest to several domains [4], [6], [11], [15] such as Artificial Intelligence, Machine Learning, Data Mining, Big Data, Visualization, Marketing, Web mining, etc, is much more less developed than PAR due to the significant problem complexity caused by high computational cost and huge search space in calculating NAR candidates.

In this paper, we propose a Condensed Representation representing non-redundant positive and negative association rules based on *generator itemsets*, *closed itemsets*, *maximal*

itemsets and *minimal infrequent itemsets*. The main contributions are summarized as follows. 1) We propose GC2M algorithm for mining simultaneously all frequent generators, all frequent closeds, all maximal frequent itemsets, and all minimal infrequent itemsets. GC2M is an abbreviation of *Generator itemsets, Closed itemsets, Maximal itemsets, and Minimal infrequent itemsets*. 2) We introduce a formal definition for uninteresting association rules (UAR), then propose an efficient strategy for pruning UAR using M_{GK} measure [7]. 3) We propose an efficient strategy for search space pruning. 4) We propose three new efficient bases based on M_{GK} measure : Concise Basis for Positive Approximate Rules (CBA), Concise Basis for Negative Exact Rules (CBE^{-}), and Concise Basis for Negative Approximate Rules (CBA^{-}). We prove that these concise bases are a lossless representation of non-redundant rules since all valid rules can be derived from these (cf. Theorems 2, 3, 4 and 5). 7) Based on these formalizations, we develop an efficient algorithm, called CONCISE, to discover non-redundant rules.

This paper is organized as follows. Section 2 discusses the related works. Section 3 gives the basic concepts. A Condensed Representation for PNARs is detailed in Section 4. Section 5 presents the experimental results. Conclusion and future work are given in Section 6.

2 Related works

The approaches of association rules mining can be roughly divided into two categories: i Bases of positive association rules, and ii Bases of negative association rules.

In positive basis, we present Duquenne-Guigues basis [10]. Without going into the details of its calculation, this approach is not informative. Bastide's approach [2] adapts Duquenne-Guigues basis. However, it inherits the same flaws as Guigues's approach [10]. In [13], the authors define two bases: *Exat Min-Max Association Rules* and *Approximate Min-Max Association Rules*. Despite their indisputable interests, these two bases contain UARs, and not complete (i.e. they do not generate the negative association rules). In [14], Pasquier defines two bases: *Generic Base for Exact Rules* and *Generic Base for Approximate Rules*. However, this approach is still incomplete and not optimal: it extracts only the positive association rules, many of which are UARs due to the confidence. Xu's approach [16] also extends Pasquier's approach [13], and defines two bases: *Reliable Approximate Basis* and *Reliable Exact Basis*, using CF (Certainty Factor). Similar to Pasquier's approach [14], Xu's approach [16] is also incomplete, it only considers positive rules, don't consider negative association rules.

In negative basis, it is important to mention that the extraction of negative rules is less developed compared to that of positive rules. Note that it emerges from the bibliographic study conducted so far that Feno's approach [7] is the first approach to have studied the problem of bases for negative rules. It extends the Pasquier's approach [13], and defines four bases: *Basis for Exact Positive rules (BPE), Basis for Approximate Positive rules (BPA), Basis for Exact Negative Rules (BNE)* and *Basis for Approximate Negative Rules (BNA)*. However, this approach is not informative, because it selects the premises from a positive borders [12] (or pseudo-closed [13]) which intuitively returns the maximal elements, not in accordance with the notion of minimal premise. It is not very selective due to the use of critical value (cf. Equation (4)) when selecting valid rules. In addition, its formulation of negative exact rules is not appropriate which can present a high memory for searching space. Recently, Dong et al. [5] propose an efficient method for pruning redundant negative and positive rules, using Confidence and Correlation coefficient. Similar to Pasquier's approach, no methods to prune UARs has been found. In particular, Dong's approach does not consider a concept of bases for non-redundant rules, then its configuration semantics is not comparable to our approach.

From this quick literature, mining informative association rules is still a major challenge, for several reasons. On the one hand, the majority of existing approaches are limited on positive association rules which are not sufficient to guarantee the interest of knowledge extraction. On

the other hand, these approaches are also limited on classic pair support-confidence [1] which produces a high number of association rules whose interest is not always guaranteed.

3 Basic concepts

In association rules problem, a Database (cf. Table 1) is a triplet $\mathcal{B} = \mathcal{T}, \mathcal{I}, \mathcal{R}$. $\mathcal{T}$ and $\mathcal{I}$ are finite sets of transactions and items respectively. $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$ is a binary relation between $\mathcal{T}$ and $\mathcal{I}$. A relation $i\mathcal{R}t$ denotes that the item i satisfies the transaction t. Let $X \subseteq \mathcal{I}$, $\overline{X} = \{t \in \mathcal{T} | \exists i \in X : i, t \notin \mathcal{R}\}$ is complementary set of X. A subset $X \subseteq \mathcal{I}$ with $k = |X|$ is called k-itemset, where $|X|$ denotes the cardinality of X. The set $\phi X = X' = \{t \in \mathcal{T} | i\mathcal{R}t, \forall i \in X\}$ is called extension of X. Similarly, the set $\psi Y = Y' = \{i \in \mathcal{I} | i\mathcal{R}t, \forall t \in Y\}$ is intension of Y. Both functions ϕ and ψ form a Galois connection between $\mathcal{PI}$ and $\mathcal{PT}$ [8], where $\mathcal{P}O$ is a power set of O. The composite function $\gamma X = \psi o \phi X$ is called Galois closure operator. Let $X, Y \subseteq \mathcal{I}$, the support of X is defined as $suppX = PX' = \frac{|X'|}{|\mathcal{T}|}$, where P is a discrete probability. The support and confidence [1] of $X \rightarrow Y$ are defined by $suppX \cup Y$ and $confX \rightarrow Y = PY'|X'$ respectively. Let $minsup \in 0, 1$ a minimum support threshold, X is frequent if $suppX \geqslant minsup$. We define $\mathcal{F}$ the set of all frequent in database $\mathcal{B}$ as $\mathcal{F} = \{X \subseteq \mathcal{I} | suppX \geqslant minsup\}$. Let $X, Y \subseteq \mathcal{I}$, X and Y are said to be equivalent, denoted by $X \cong Y$, iff $\gamma X = \gamma Y$. The set of itemsets that are equivalent to X is $X = \{Y \subseteq \mathcal{I} \mid X \cong Y\}$. The item C is closed iff $C = \gamma C$. We define the set $\mathcal{FC}$ of all frequent closed itemsets in database $\mathcal{B}$ as: $\mathcal{FC} = \{C \in \mathcal{I} | C = \gamma C,\ suppC \geqslant minsup\}$. An itemset G is said a minimal generator of a closed C iff $\gamma G = C$ and $g \subseteq \mathcal{I}$ with $g \subseteq G$ such that $\gamma g = C$. We define the set $\mathcal{G}_C$ of all frequent generators as: $\mathcal{G}_C = \{G \in C | C \in \mathcal{FC},\ g \subset G,\ suppG \geqslant minsup\}$. We define $\mathcal{MFC}$ the set of all maximal frequent in database $\mathcal{B}$ as: $\mathcal{MFC} = \{C \in \mathcal{FC} \mid \mathcal{C} \supset C,\ \mathcal{C} \in \mathcal{FC}\}$.

Table 1. Context $\mathcal{B}$

TID	Items
1	ACD
2	BCE
3	ABCE
4	BE
5	ABCE
6	BCE

4 Condensed Representations for PNARs

Our approach is divided into two successive steps: (i) it extracts the set $\mathcal{FC}$, $\mathcal{MFC}$, $\mathcal{G}_{\gamma.}$, and the set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimal infrequent itemsets in $\mathcal{B}$; (ii) it derives from these frequent sets the non-redundant informative rules. An association rule is informative if its premise (resp. conclusion) is minimal (resp. maximal). For lack of space, certain proofs of the Properties are omitted.

4.1 Generating of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$

Our main motivation lies in absence of an autonomous approach for mining $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$. We then propose an efficient algorithm, GC2M, that simultaneously collects these four sets $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$ and $\overline{\mathcal{F}}_{\text{MIN}}$ in database $\mathcal{B}$. Here we briefly describe GC2M algorithm. It's composed of two algorithms (Algo. 1 and Algo. 2). Its main orginality lies in the effective support counting strategy: *Let X be a frequent k-itemset ($k \geqslant 3$) and $\widetilde{X}$ a $k-1$-subsets of X. Then, X is not a generator iff $suppX = min\{supp\widetilde{X} | \widetilde{X} \subset X\}$* [2], i.e. no access to context $\mathcal{B}$ is made if X is non-generator. On search space pruning, it uses the following properties: *(i) All subsets of a frequent are frequent, (ii) All supersets of an infrequent itemset are infrequent, (iii) All subsets of a generator are also generator, (iv) All supersets of a non-generator are also non-generator* [2]. These results will be synthesized in the algorithm 1. The following Figure 1 shows examplary execution of Algorithm 1 with a small context $\mathcal{B}$ from table 1 and fixed $minsup = 26$. From $\mathcal{MFC}$, we can derive the set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimimal infrequent in $\mathcal{B}$.

Definition 1 (Minimal infrequent itemset) *Let $\mathcal{MFC}$ be the set of maximal frequent, and $\mathcal{F}$ the set of frequent in $\mathcal{B}$. The set $\overline{\mathcal{F}}_{\text{MIN}}$ of minimal infrequent itemsets in $\mathcal{B}$ is defined as :*

$$\overline{\mathcal{F}}_{\text{MIN}} = \{X \in 2^{\mathcal{I}} \backslash \mathcal{MFC} \mid Y \subset X, Y \notin \mathcal{F}\}. \quad (1)$$

Require: A database $\mathcal{B}$, A minimum support threshold $minsup \in 0, 1$.
Ensure: List of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$.
1: $\mathcal{FCC}_1$.GENERATORS $\leftarrow$ {1-itemsts}
2: **for all** $(k \leftarrow 1; \mathcal{FCC}_k$.GENERATORS $\neq \emptyset; k$) **do**
3: $\mathcal{FCC}_k.closure \leftarrow \emptyset$; $\mathcal{FCC}_k.support \leftarrow 0$;
4: $\mathcal{FCC}_k \leftarrow$ GENCLOSURES$\mathcal{FCC}_k$
5: **for all** (candidate itemsets $c \in \mathcal{FCC}_k$) **do**
6: Calculate $suppc$;
7: **if** $(suppc \geqslant minsup)$ **then**
8: $\mathcal{FC}_k \leftarrow \mathcal{FC}_k \cup \{c\}$
9: **end if**
10: **end for**
11: $\mathcal{FCC}_{k1} \leftarrow$ GENGENERATORS$\mathcal{FC}_k$
12: $\mathcal{FCC}_{k1} \leftarrow$ GENMAXIMAL$\mathcal{FC}_k$
13: **end for**
14: $\mathcal{FC} \leftarrow \overset{k-1}{\underset{j=1}{}}\{\mathcal{FC}_j$.CLOSURE, $\mathcal{FC}_j$.support$\}$
15: **return** $\mathcal{FC}$

Algorithm 1: GENERATING $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ AND $\mathcal{MFC}$

Scan $\mathcal{B}$ $\longrightarrow$

$\mathcal{FCC}_1$		$supp$
A	AC	3/6
B	BE	5/6
C	C	5/6
~~D~~	~~ACD~~	$min36, 16, 16 = 16 < 26$
E	BE	5/6

Pruning the infrequent itemsets $\longrightarrow$

$\mathcal{G}_{\gamma.}$	$\mathcal{FC}$	$\mathcal{MFC}$	$supp$
A	AC	-	3/6
B	BE	-	5/6
C	C	-	5/6
E	BE	-	5/6

Scan $\mathcal{B}$ $\longrightarrow$

$\mathcal{FCC}_2$		$supp$
AB	ABCE	$min26, 26, 26, 46 = 26$
AE	ABCE	$min26, 26, 26, 46 = 26$
BC	BCE	$min46, 56, 46 = 46$
CE	BCE	$min46, 56, 46 = 46$

Pruning the infreq. itemsets $\longrightarrow$

$\mathcal{G}_{\gamma.}$	$\mathcal{FC}$	$\mathcal{MFC}$	$supp$
AB	ABCE	ABCE	2/6
AE	ABCE	ABCE	2/6
BC	BCE	-	4/6
CE	BCE	-	4/6

Figure 1. List of $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$ using Algorithm 1 with $minsup = 26$

The Algorithm 2 inputs a database $\mathcal{B}$, a $minsup$, and outputs the set $\overline{\mathcal{F}}_{\text{MIN}}$. Let's take our

Require: $\mathcal{B}$, $\mathcal{MFC}$ and $minsup \in 0, 1$.
Ensure: $\overline{\mathcal{F}}_{\text{MIN}}$ the set of minimal infrequent itemsets.
1: $\overline{\mathcal{F}}_{\text{MIN}} \leftarrow \emptyset$
2: **for all** $(X \in 2^{\mathcal{I}} \backslash \mathcal{MFC})$ **do**
3: **if** $(Y \subset X \mid suppY \leq minsup)$ **then**
4: $\overline{\mathcal{F}}_{\text{MIN}} \leftarrow \overline{\mathcal{F}}_{\text{MIN}} \cup \{X\}$
5: **end if**
6: **end for**
7: **return** $\overline{\mathcal{F}}_{\text{MIN}}$

Algorithm 2: GENERATING MINIMAL INFREQUENT ITEMSETS

example in Figure 1 at $minsup = 26$, we have $D \in 2^{\mathcal{I}} \backslash \mathcal{MFC}$. We see that $D \notin \mathcal{F}$ and $\widetilde{D} \subset D$ such taht $supp\widetilde{D} \leq minsup$. This means D is a minimal infrequent itemsets (i.e. $D \in \overline{\mathcal{F}}_{\text{MIN}}$).

4.2 Generating non-redundant PNARs

This Subsection is based essentially on 5 components : Pruning UAR, Modelization of significant rules, Search space pruning, Pruning redundant PNARs, and CONCISE algorithm.

4.2.1 Pruning Uninteresting Association Rules (UAR).

We first formalize the idea of UAR, and then propose a strategy to prune UAR. Note that the classic support-confidence [1] is not able to prune UAR. Table 2 ullustrates these limits. The information given in this Table 2 can be used to evaluate the association $A \rightarrow B$ and tea $\rightarrow$ coffee. For the pair A, B, we have $suppA \cup B = 0.72$ and $confA \rightarrow B = 0.9$. For the pair (tea,coffee), we have $supp$tea $\cup$ coffee $= 0.2$ and $conf$tea $\rightarrow$ coffee $= 0.8$. The support and confidence are considered fairly high for both rules, i.e. $A \rightarrow B$ and tea $\rightarrow$ coffee are interesting rules. How-

Table 2. Contingency table

	A	$\neg A$			coffee	¬coffee	
B	72	18	90	tea	20	5	25
$\neg B$	8	2	10	¬tea	70	5	75
	80	20	100		90	10	100

ever, $PB'|A' = PB' = 0.9$ and $conf$tea $\rightarrow$ coffee = $0.8 < 0.9 = supp$coffee implie A and B are independent (resp. tea disfavors coffee), i.e. $A \rightarrow B$ and tea $\rightarrow$ coffee are UAR.

Definition 2 (Uninteresting Association Rules (UAR)) *Let $X, Y \subseteq \mathcal{I}$ such that $X \cap Y = \emptyset$. An association rule $X \rightarrow Y$ is said to be uninteresting rule if Y is independent on X (i.e. $PY'|X' = PY'$) or Y is negatively dependent on X (i.e. $PY'|X' < PY'$).*

We then propose an UAR pruning strategy by measuring the degree dependency of X and Y, denoted $\Delta_{X,Y} = PY'|X' - PY'$. We then use M_{GK} measure [7], defined as :

$$M_{GK}X \rightarrow Y = \begin{cases} \frac{PY'|X' - PY'}{1 - PY'}, \text{if } \Delta_{X,Y} > 0 \\ \frac{PY'|X' - PY'}{PY'}, \text{if } \Delta_{X,Y} \leqslant 0. \end{cases} \quad (2)$$

The M_{GK} refers to dependencies between the antecedent and consequent of an association rule. Values in $-1, 0$ show that there is a negative dependence between X and Y. Values in $0, 1$ show that there is a positive dependence between X and Y. Value equal 0 show that Y independent on X. We recall, rules with M_{GK} equal to 1 are called Exact Association Rules, and rules with M_{GK} less than 1 are called Approximate Rules. Theorem 1 below states that value of UARs defined by Definition 2 will be statistically **null** or **negative**.

Theorem 1 ([2) *] Let $X, Y \subseteq \mathcal{I}$. (1) If $PY'|X' \leqslant PY'$, we have $-1 \leqslant M_{GK}X \rightarrow Y \leqslant 0$. (2) If $PY'|X' > PY'$, then $0 < M_{GK}X \rightarrow Y \leqslant 1$.*

From the same example of table 2, we have $M_{GK}A \rightarrow B = \frac{0.9-0.9}{1-0.9} = 0$, this verifies that A and B are independent. So, $A \rightarrow B$ is UAR. We also obtain M_{GK}tea $\rightarrow$ coffee = $\frac{0.8-0.9}{1-0.9} = -1 < 0$, this means that `coffee` and `tea` are negatively dependent. In other words `tea` $\rightarrow$ `coffee` is UAR. As result, the UARs are systematically pruned using M_{GK}.

4.2.2 Modelization of significant rules using M_{GK}.

Note that the first component of M_{GK} (Eq. (2)) is implicative but the second is not, only the first will be active in modelization. We introduce the quantities $n = |\mathcal{T}|$, $n_X = |\phi X|$, $n_Y = |\phi Y|$, $n_{X \wedge Y} = |\phi X \cup Y|$ and $n_{X \wedge \overline{Y}} = |\phi X \cup \overline{Y}|$. The quantity $N_{X \wedge \overline{Y}}$ indicates a random variable which generates $n_{X \wedge \overline{Y}}$, and $N_{X \wedge Y}$ generates $n_{X \wedge Y}$. In that case, the Eq. (2) can be rewritten :

$$M_{GK}X \rightarrow Y = 1 - \frac{n n_{X \wedge \overline{Y}}}{n_X n_{\overline{Y}}} \quad (3)$$

The current versions [7] are based on critical value γ_α defined as

$$\gamma_\alpha = \sqrt{\frac{1}{n} \frac{n - n_X}{n_X} \frac{n_Y}{n - n_Y} \chi^2 \alpha}, \quad (4)$$

where α a real in the interval $0, 1$ and $\chi^2\alpha$ is a Chi-square statistic of a single degree of freedom. This means that $X \rightarrow Y$ will be valid if $M_{GK}X \rightarrow Y \geqslant \gamma_\alpha$. However, this critical value can nevertheless present some limits. Indeed, a low α leads to a high critical value which rapidly exceeds the M_{GK} value. This rejects certain robust rules. Conversely, a large value of α leads to a very low critical value. This accepts certain very weak rules (i.e. independent rules).

To overcome these limits, we define a new model based on the test H_0 independence hypothesis of X and Y in the face of a positive dependence hypothesis H_1, of the rule $X \rightarrow Y$. We then model, under H_0 independence hypothesis, the probability between the random variable $N_{X \wedge \overline{Y}}$ and the observed counter-examples $n_{X \wedge \overline{Y}}$ using measure M_{GK}. We notice that

the sensitivity of this measure M_{GK} to variations in the occurrences of the observed counter-examples $n_{X\wedge\overline{Y}}$ reades with the partial derivative given in the following Equation (5):

$$\frac{\partial M_{GK}}{\partial n_{X\wedge\overline{Y}}} = -\frac{1}{\frac{n_X n_{\overline{Y}}}{n}} \tag{5}$$

This shows that M_{GK} decreases when the number $n_{X\wedge\overline{Y}}$ increases and all the more quickly as the quantity $\frac{n_X n_{\overline{Y}}}{n}$ is low. In other words, M_{GK} grows when $n_{X\wedge\overline{Y}}$ decreases, which is semantically acceptable, but the rate of variation is constant, independent of the rate of decrease of this number, variations of n_Y. Consider $\widetilde{M_{GK}}$ as the realization of a variable M_{GK}, defined as:

$$\widetilde{M_{GK}}X \to Y = -\widetilde{M_{GK}}X \to \overline{Y} = -\frac{n_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{\frac{n_X n_{\overline{Y}}}{n}}}\sqrt{\frac{n}{n_X n_{\overline{Y}}}} \tag{6}$$

It is the opposite of the directed contribution of the cell $X \cup \overline{Y}$ to the $\frac{\chi^2}{n}$ except for a constant. In practice, it is quite common to observe a few transactions which contain X and not Y without having the general trend to have Y when X is present contested. Therefore, $n_{X\wedge\overline{Y}}$ must be taken into account to statistically accept to retain or not the rule $X \to Y$. Suppose we draw at random two subsets $U, Z \subseteq \mathcal{I}$ which contain n_X and n_Y respectively, i.e. $N_{X\wedge\overline{Y}} = |\phi U \cup \overline{Z}|$. This variable $N_{X\wedge\overline{Y}}$ follows a Poisson law with parameter $\frac{n_X n_{\overline{Y}}}{n}$ [9]. We then measure the smallness of random variable $N_{X\wedge\overline{Y}}$ expected to the number $n_{X\wedge\overline{Y}}$ under H_0 independence hypothesis between X and Y. Such an association rule $X \to Y$ is then said to be admissible at the threshold $\alpha \in 0, 1$ if the probability that the random variable $N_{X\wedge\overline{Y}}$ is lower than that observed number $n_{X\wedge\overline{Y}}$ under H_0 independence hypothesis on X and Y is relatively low :

$$PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 \leqslant \alpha. \tag{7}$$

We then have :

$$PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 = P\left(\frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}}\frac{n}{\sqrt{n_X n_{\overline{Y}}}} \leqslant \widetilde{M_{GK}}X \to \overline{Y}\right)$$

Noting Φ. the standard normal distribution, we have $\frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}0, 1$, and $\frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{p.s} 1 \Rightarrow \frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathbb{P}} 1 \Rightarrow \mathcal{K}X, \overline{Y} = \frac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{n_X n_{\overline{Y}}}} \times \frac{n}{\sqrt{n_X n_{\overline{Y}}}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}0, 1$. Finally, we have:

$$\begin{aligned} PN_{X\wedge\overline{Y}} \leqslant n_{X\wedge\overline{Y}}|H_0 &= P\left(\mathcal{K}X, \overline{Y} \leqslant \widetilde{M_{GK}}X \to \overline{Y}\right) \\ &\overset{TCL}{\simeq} \int_{-\infty}^{\widetilde{M_{GK}}X\to\overline{Y}} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t = \Phi\widetilde{M_{GK}}X \to \overline{Y}. \end{aligned}$$

Our model of significant association rules is given in the following Definition 3.

Definition 3 (Valid association rule) *Given a minimum threshold $\alpha \in 0, 1$. An association rule $X \to Y$ is said to be valid at level confidence $1 - \alpha$, called $1 - \alpha$-valid, if only if :*

$$p_{XY} = 1 - \Phi - \widetilde{M_{GK}}X \to \overline{Y} \geqslant 1 - \alpha \tag{8}$$

For example, from Table 1, consider a rule $A \to \overline{BCE}$ with $\alpha = 1\%$. Here, $n_A = 2$, $n_B = 5$ and $n_{BCE} = 4$. From $n_A = 2$, $n_B = 5$, $n_{BCE} = 4$ and $\frac{n_A n_{BCE}}{n} = 0 < 3$ (Gaussian hypothesis, cf. [9]), we have $p_{A\overline{BCE}} = 0.5 < 0.99 \Rightarrow A \to \overline{BCE}$ is 99%-invalid (i.e. it's not valid at $\alpha = 1\%$).

4.2.3 Search space pruning.

Pasquier's approach [14] is the most popular approach for generating of non-redundant rules. However, no methods for search space pruning of significant valid rules is used by this approach. While it is possible to restrict the search space by partitioning into 2 the 8 ($X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$, $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$) candidates in database $\mathcal{B}$.

We then explain this restriction. In [2], we demonstrated that if X favors Y (i.e. $PY'|X' > PY'$), then these are the four association rules $X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$ and $\overline{Y} \rightarrow \overline{X}$, which will be studied. If X disfavors Y (i.e. $PY'|X' \leqslant PY'$), then these are the four contrary association rules $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$ which will be studied. We then obtain two classes: Class of rules ($X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$), denoted $\mathscr{C}_1$, and Class of rules ($X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$, $Y \rightarrow \overline{X}$), denoted $\mathscr{C}_2$. We also demonstrated that all rules of $\mathscr{C}_1$ can be derived from $X \rightarrow Y$, and all rules of $\mathscr{C}_2$ can be derived from $X \rightarrow \overline{Y}$. So, we only study two rules such as $X \rightarrow Y$ and $X \rightarrow \overline{Y}$. This gives the reduction space 100(8-2)/8=75%.

4.2.4 Pruning redundant PNARs.

The most popular method to prune redundant rules is the base of rules that is a set of reduced size rules that do not contain any redundant rule. Definition 4 defines a redundant rule.

Definition 4 (Redundant rule) *The rule $r_1 : X_1 \rightarrow Y_1$ is redundant if $\exists r_2 : X_2 \rightarrow Y_2$, where $X_1 \supset X_2$, $Y_1 \subset Y_2$ such that $suppr_1 = suppr_2$ and $M_{GK}r_1 = M_{GK}r_2$.*

Corresponding to the three popular approaches [7], [14], [16], we propose three more efficient bases called Concise Bases as defined in Definitions 6, 7 and 8. In addition, we define a base for Positive Exact Rules using M_{GK}, called CBE (cf. Definition 5). More precisely, CBE basis is similar of *Base for Exact Rules* defined in [14], because an exact rule of Confidence is also exact of M_{GK} (cf. [2]). We prove that these concise bases are a lossless representation of non-redundant rules since all valid rules can be derived from these (cf. Theorems 2, 3, 4, 5).

Definition 5 (CBE Basis) *Let $\mathcal{FC}$ be the set of frequent closed itemsets. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_\mathcal{C}$ be the set of minimal generators of $\mathcal{C}$, we have:*

$$CBE = \{G \rightarrow \mathcal{C} \backslash G \mid G \in \mathcal{G}_\mathcal{C},\ \mathcal{C} \in \mathcal{FC},\ G \neq \mathcal{C}\} \tag{9}$$

Theorem 2 *(i) All valid positive exact rules and their supports can be derived from to CBE basis. (ii) All rules in CBE are non-redudant exact rules.*

Proof 1 *i Let $r_1 : X_1 \rightarrow Y_1 \backslash X_1$ be the exact positive rule between two frequents X_1 and Y_1 such that $X_1 \subset Y_1$. Let $\mathcal{C}$ be a frequent closed itemset in $\mathcal{B}$ (i.e. $\mathcal{C} \in \mathcal{FC}$). Since $M_{GK}r_1 = 1$, we have $suppX_1 = suppY_1$. From $suppX_1 = suppY_1$, we derived that $supp\gamma X_1 = supp\gamma Y_1 \Rightarrow \gamma X_1 = \gamma Y_1 = \mathcal{C}$. Obviously, there exists a rule $r_2 : G \rightarrow \mathcal{C} \backslash G \in CBE$ such that G is a generator of $\mathcal{C}$ for which $G \subseteq X_1$ and $G \subseteq Y_1$. We show that the rule r_1 and its supports can be derived from the rule r_2 and its supports. From $\gamma X_1 = \gamma Y_1 = \mathcal{C}$ and $\gamma G = \mathcal{C}$, we then have $suppr_1 = supp\gamma X_1 = supp\gamma Y_1 = supp\mathcal{C} = suppr_2$, and deduce that $M_{GK}r_1 = M_{GK}r_2$. This explains that r_1 can be derived from r_2, and is a redundant rule of r_2, so it's pruned in CBE base.*

ii Let $r_2 : G \rightarrow \mathcal{C} \backslash G \in CBE$, we then have $G \in \mathcal{G}_\mathcal{C}$ and $\mathcal{C} \in \mathcal{FC}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow Y_3 \backslash X_3 \in CBE$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subseteq \gamma G = \mathcal{C}$. We deduce that $X_3 \notin \mathcal{G}_\mathcal{C} \Rightarrow r_3 \notin CBE$. If $\mathcal{C} \subseteq Y_3$, we then have $\mathcal{C} = \gamma\mathcal{C} = \gamma G \subset Y_3 = \gamma Y_3 \Rightarrow G \notin \mathcal{G}_{Y_3}$. In other words, r_2 is non-redundant. This proves that CBE is a non-redundant base.

Definition 6 (CBA Basis) *Let $\mathcal{FC}$ be the set of frequent closed. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_{\mathcal{C}}$ be the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$CBA = \{G \rightarrow \mathcal{C}\backslash G | G, \mathcal{C} \in \mathcal{G}_{\gamma G} \times \mathcal{FC}, \gamma G \subset \mathcal{C}, P\mathcal{C}' | G' > P\mathcal{C}', p_{G\mathcal{C}} \geqslant 1 - \alpha\} \tag{10}$$

Theorem 3 *(i) All valid positive approximate association rules, their supports and M_{GK}, can be derived from the rules of CBA. (ii) All association rules in the CBA basis are non-redundant approximate association rules.*

Proof 2 *i Let $r_1 : X_1 \rightarrow Y_1\backslash X_1 \in CBA$ such that $X_1 \subset Y_1$. For any X_1 and Y_1, there is a generator G_1 such that $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and a generator G_2 such that $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$. Since $X_1 \subset Y_1$, we have $X_1 \subseteq \gamma G_1 \subset Y_1 \subseteq \gamma G_2$ and the rule $r_2 : G_1 \rightarrow \gamma G_2\backslash G_1 \in CBA$. We show that r_1 can be derived from r_2. Since $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$, we have $suppG_1 = suppX_1$ and $suppG_2 = suppY_1 = supp\gamma G_2$. This gives that $suppr_1 = suppr_2$ and $M_{GK}r_1 = M_{GK}r_2$, in other words, r_1 can be derived from r_2 and therefore, r_1 is a redundant rule of r_2.*

ii Let $r_2 : G \rightarrow \mathcal{C}\backslash G \in CBA$, we then have $\mathcal{C} \in \mathcal{FC}$ and $G \in \mathcal{G}_{\mathcal{C}}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow Y_3\backslash X_3 \in CBA$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subset \gamma G = \mathcal{C} \Rightarrow X_3 \notin \mathcal{G}_C$. If $\mathcal{C} \subseteq Y_3$, we then have $\mathcal{C} = \gamma\mathcal{C} \subset Y_3 = \gamma Y_3$. As result, $G \notin \mathcal{G}_{Y_3} \Rightarrow r_3 \notin CBA$, in other words, r_2 is a non-redundant rule. This proves that CBA is a non-redundant base.

Definition 7 (CBE^- Basis) *Let $\mathcal{MFC}$ be the set of maximal frequent itemsets, $\overline{\mathcal{F}}_{\text{MIN}}$ the set of minimal infrequent on database $\mathcal{B}$. For each $\mathcal{M} \in \mathcal{MFC}$, let $\mathcal{G}_{\mathcal{M}}$ be the set of minimal generators of $\mathcal{M}$, we have:*

$$CBE^- = \{G \rightarrow \overline{y} \mid G \in \mathcal{G}_{\mathcal{M}},\ \mathcal{M} \in \mathcal{MFC},\ y \in \overline{\mathcal{F}}_{\text{MIN}}\} \tag{11}$$

Theorem 4 *(i) All valid negative exact association rules, their supports and M_{GK}, can be derived from the rules of the CBE^- basis. (ii) All association rules in the CBE^- basis are non-redudant negative exact association rules.*

Proof 3 *i Let $r_1 : X_1 \rightarrow \overline{Y_1}\backslash X_1 \in CBE^-$ such that $X_1 \subset \overline{Y_1} \subseteq \mathcal{M}$ where $\mathcal{M} \in \mathcal{MFC}$. Since $M_{GK}r_1 = 1$, we have $X_1 \cong \overline{Y}_1 \Rightarrow suppX_1 = supp\overline{Y_1}$. Since $suppX_1 = supp\overline{Y_1}$, we have $supp\gamma X_1 \cup \overline{Y_1} = supp\gamma X_1 = supp\gamma\overline{Y_1} \Rightarrow \gamma X_1 \cup \overline{Y}_1 = \gamma X_1 = \gamma\overline{Y_1} = \mathcal{M}$ a. Obviously, $\exists r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$ such that $G \in \mathcal{G}_{\mathcal{M}}$ for which $G \subseteq X_1$ and $G \subseteq \overline{Y_1}$, and thus $G \subseteq \overline{y}$ (by Definition 7). We show that the rule r_1 can be derived from r_2. Since $r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$, we have $suppG \cup \overline{y} = suppG$. From $suppG \cup \overline{y} = suppG$, we have $supp\gamma G \cup \overline{y} = supp\gamma G = supp\gamma\overline{y} \Rightarrow \gamma G \cup \overline{y} = \gamma G = \gamma\overline{y} = \mathcal{M}$ a'. From relations a and a', we have $\gamma G \cup \overline{y} = \gamma X_1 \cup \overline{Y_1} \Leftrightarrow suppr_1 = suppr_2$. Since $G \subseteq X_1 \subset \overline{Y_1} \subset \overline{y} \subseteq \gamma G = \mathcal{M}$, we have $suppG = suppX_1 = supp\overline{Y_1} = supp\overline{y} = supp\mathcal{M} \Rightarrow M_{GK}r_1 = M_{GK}r_2$. These results explain that r_1 can be derived from r_2, and is a redundant rule w.r.t r_2.*

ii Let $r_2 : G \rightarrow \overline{y}\backslash G \in CBE^-$, i.e. $G \in \mathcal{G}_{\mathcal{M}}$ and $y \in \overline{\mathcal{F}}_{\text{MIN}}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y}_3\backslash X_3 \in CBE^-$ such as $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subseteq G$ and $\overline{y} \subseteq \overline{Y_3}$. If $X_3 \subseteq G$, we then have $\gamma X_3 \subseteq \gamma G \subset \gamma\overline{y} = \mathcal{M}$. We deduce that $X_3 \notin \mathcal{G}_{\mathcal{M}}$ and conclude that $r_3 \notin CBE^-$. If $\overline{y} \subseteq \overline{Y_3}$, we then have $\gamma G \subset \gamma\overline{y} \subseteq \gamma\overline{Y_3} = \mathcal{M}$. We deduce that $G \notin \mathcal{G}_{\overline{Y}_3}$ and conclude that $r_3 \notin CBE^-$. This implies that r_2 is a non-redundant rule, and proves that CBE^- is a non-redundant base.

Definition 8 (CBA^- Basis) *Let $\mathcal{FC}$ be the set of frequent closeds. For each $\mathcal{C} \in \mathcal{FC}$, let $\mathcal{G}_{\mathcal{C}}$ be the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$CBA^- = \{G \rightarrow \overline{g} | G, g \in \mathcal{G}_{\gamma G} \times \mathcal{G}_{\gamma g}, \gamma G \subsetneq \gamma g, PG' | \overline{g'} > P\overline{g'}, p_{G\overline{g}} \geqslant 1 - \alpha\} \tag{12}$$

Theorem 5 *(i) All valid negative approximate association rules, their supports and M_{GK}, can be derived from the rules of CBA^-. (ii) All association rules in the CBA^- are non-redudant negative approximate association rules.*

Proof 4 *i Let $r_1 : X_1 \rightarrow \overline{Y_1}\backslash X_1 \in CBA^-$ with $X_1 \subset \overline{Y_1}$. For any frequent X_1 and Y_1, there is a generator G_1 such that $G_1 \subset X_1 \subseteq \gamma X_1 = \gamma G_1$ and a generator G_2 such that $G_2 \subset Y_1 \subseteq \gamma Y_1 = \gamma G_2$. Since $X_1 \subset \overline{Y_1}$, we have $X_1 \subseteq \gamma G_1 \subset \overline{Y_1} \subset \overline{G_2} \subseteq \gamma\overline{Y_1} = \gamma\overline{G_2}$. Obviously, $\exists r_2 : G_1 \rightarrow \overline{G_2}\backslash G_1 \in CBA^-$ such that $\gamma G \subsetneq \gamma g$ (by Definition 8). We show that r_1 can be derived from r_2. From $G_1 \subset X_1 \subseteq \gamma G_1$ and $G_2 \subset Y_1 \subseteq \gamma G_2$, we then have $G_1 \cong X_1$ and $\overline{Y_1} \cong \overline{G_2} \Rightarrow suppX_1 \cup \overline{Y_1} = suppG_1 \cup \overline{G_2}$ and $M_{GK}X_1 \rightarrow \overline{Y_1} = M_{GK}G_1 \rightarrow \overline{G_2}$. This explains that r_1 can be derived from r_2, and is a redundant rule w.r.t. r_2. ii Let $r_2 : G \rightarrow \overline{g}\backslash G \in CBA^-$, i.e. $G \in \mathcal{G}_C$ and $g \in \mathcal{G}_\mathcal{C}$ such that $\gamma G \subsetneq \gamma g$ (i.e. $C \subsetneq \mathcal{C}$). We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y_3}\backslash X_3 \in CBA^-$ such that $suppr_3 = suppr_2$, $M_{GK}r_3 = M_{GK}r_2$, $X_3 \subset G$ and $\overline{Y_3} \supset \overline{g}$. If $X_3 \subset G$, we then have $\gamma X_3 \subset \gamma G = C \Rightarrow X_3 \notin \mathcal{G}_C$. Since $X_3 \subset G$, we have $suppX_3 > suppG \Rightarrow M_{GK}r_3 < M_{GK}r_2$. If $\overline{g} \subset \overline{Y_3}$, we then have $supp\overline{g} > supp\overline{Y_3} \Rightarrow M_{GK}r_2 > M_{GK}r_3$. This means that r_2 is a non-redundant rule, and proves that CBE^- is a non-redundant base.*

4.2.5 CONCISE algorithm.

CONCISE is composed of three algorithms (Algo. 1, Algo. 2,Algo. 3). The principal procedure (Algo. 3) takes as input $\mathcal{G}_\mathcal{C}$, $\mathcal{FC}$, $\mathcal{MFC}$, $\overline{\mathcal{F}}_{\text{MIN}}$, $minsup$ and α. It returns all non-redundant PNARs.

Require: $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$, $\mathcal{MFC}$, $\overline{\mathcal{F}}_{\text{MIN}}$, $minsup$ and α.
Ensure: $\mathcal{CB}$, A Concise Base of Non-Redundant PNARs.

$\mathcal{CB} = \emptyset$;
for all ($\mathcal{C} \in \mathcal{FC}$) **do**
  for all ($G \in \mathcal{G}_\mathcal{C}$) **do**
    if ($P\mathcal{C}'|G' > P\mathcal{C}'$) **then**
      if ($\gamma G = \mathcal{C}$) **then**
        if ($G \neq \gamma G$ & $suppG \cup \mathcal{C} \geq minsup$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \mathcal{C}\backslash G\}$; /* CBE Basis */
        end if
      else if ($\gamma G \subset \mathcal{C}$) **then**
        if ($suppG \cup \mathcal{C} \geq minsup$ & $p_{GC} \geqslant 1 - \alpha$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \mathcal{C}\backslash G\}$; /* CBA Basis */
        end if
      end if
    else if ($P\mathcal{C}'|G' < P\mathcal{C}'$) **then**
      for all ($\mathcal{M} \in \mathcal{MFC}$) **do**
        for all ($G \in \mathcal{G}_\mathcal{M}$) **do**
          for all ($y \in \overline{\mathcal{F}}_{\text{MIN}}$) **do**
            if ($suppG \cup \overline{y} \geq minsup$) **then**
              $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \overline{y}\backslash G\}$; /* CBE^- Basis */
            end if
          end for
        end for
      end for
      for all ($g \in \mathcal{G}_{\gamma g}|\gamma G \subsetneq \gamma g$ & $Pg'|G' < Pg'$) **do**
        if ($suppG \cup \overline{g} \geq minsup$ & $p_{G\overline{g}} \geqslant 1 - \alpha$) **then**
          $\mathcal{CB} \leftarrow \mathcal{CB} \cup \{G \rightarrow \overline{g}\backslash G\}$; /* Base CBA^- */
        end if
      end for
    end if
  end for
end for
return $\mathcal{CB}$

Algorithm 3: GENERATING NON-REDUNDANT ASSOCIATION RULES

5 Experimental evaluation

We evaluate CONCISE with two comparable baseline approaches Pasquier's approach [14] and Feno's approach [7]. All algorithms are implemented in R. All the experiments are run on a PC Core i3-2350M with 4CPUs and 4GB memory on the Windows 7. We compare their number of valid rules and computational costs on different databases (cf. Table 3): T10I4D100K[1], T20I6D100K (cf. footnote 1), C20D10K[2] and MUSHROOMS (cf. footnote 2). We make CONCISE and Feno's approach to follow the same constraint $\alpha = 5\%$. For Pasquier's approach, consider a minimal confidence $minconf = 80\%$. The number of extracted rules for the three algorithms, by varying the $minsup$, is shown in Table 4. For this, E (resp. A) indicates the positive exact (resp. approximate) rules. E^- (resp. A^-) indicates the negative exact (resp. approximate) rules. We also denote by "-" a subset which could not generated. We observe that no negative association rules are generated by Pasquier's approach. For each algorithm, no E and A^- are generated on T10I4D100K when $minsup \leq 30\%$. The reason is that all frequent are closed itemsets. On other databases, Feno's approach represents a number smaller than Pasquier's approach and CONCISE. The explanation is that Feno's approach uses the set of pseudo-closed [13] which returns a reduced number of frequent itemsets and thus, it is the same for number of rules generated, but it's not informative. Whereas Pasquier's approach and CONCISE algorithm generate the more informative non-redundant association rules.

Table 3. Data characteristics

Database	$\lvert\mathcal{T}\rvert$	$\lvert\mathcal{I}\rvert$	Avg. size
T10I4D100K	100 000	1 000	10
T20I6D100K	100 000	1 000	20
C20D10K	10 000	386	20
MUSHROOMS	8 416	128	23

Table 4. Number of all valide non-redundant association rules

Dataset	$minsup$	Pasquier's approach				Feno's approach				CONCISE			
		$\lvert E\rvert$	$\lvert A\rvert$	$\lvert E^-\rvert$	$\lvert A^-\rvert$	$\lvert E\rvert$	$\lvert E^-\rvert$	$\lvert A\rvert$	$\lvert A^-\rvert$	$\lvert E\rvert$	$\lvert E^-\rvert$	$\lvert A\rvert$	$\lvert A^-\rvert$
T10I4D100K	10%	0	11625	-	-	0	0	10555	1256	0	0	725	52
	20%	0	8545	-	-	0	0	6656	1058	0	0	545	34
	30%	0	3555	-	-	0	0	2785	954	0	0	355	25
T20I6D100K	10%	115	71324	-	-	95	98	51899	3897	115	103	1804	56
	20%	76	57336	-	-	66	91	35560	2705	76	95	1403	38
	30%	58	45684	-	-	43	63	21784	1887	58	63	1175	27
C20D10K	10%	1125	33950	-	-	975	255	28588	11705	1125	285	1856	182
	20%	997	23821	-	-	657	135	19582	8789	997	185	1453	123
	30%	967	18899	-	-	567	98	11581	4800	967	101	1221	97
MUSHROOMS	10%	958	4465	-	-	758	289	3850	3887	958	304	1540	89
	20%	663	3354	-	-	554	178	2144	2845	663	198	1100	78
	30%	543	2961	-	-	444	109	1140	1987	543	115	998	39

On dense databases (C20D10K and MUSHROOMS), CONCISE algorithm is much more selective than Pasquier's and Feno's approaches for all $minsup$. For example, on C20D10K database and less $minsup = 1\%$, Pasquier's (resp. Feno's) approch contains 33950 (resp. 28588) positive approximate rules as showed in Table 4, while the CONCISE contains 1856 positive approximate rules; this gives the reduction ratio 94.5% and 93.51% respectively. In this case, 32094 (resp. 26732) positive approximate rules can be deduced either from the Pasquier's (resp. Feno's) approach or from the CONCISE algorithm. The main reason is associated to the different techniques to prune both UARs and redundant association rules.

We present in the following the execution times of CONCISE compared to those existing. However, this comparison is still very difficult, for several reasons. First, Feno's approach is not comparable to CONCISE, because it ignores the big phase for generating $\mathcal{G}_{\gamma.}$, $\mathcal{FC}$ and $\mathcal{MFC}$. Pasquier's approach could not generate the negative rules. We partialy compare CONCISE and Pasquier's approach on execution times of E and A. The results will be represented in Fig. 2 by varying the $minsup$ at fixed $\alpha = 0.05$ and $minconf = 0.6$. On sparse databases (T10I4D100K and T20I6D100K), CONCISE and Pasquier's approach are almost identical for positive exact rules E for all $minsup$ (cf. Fig. 2a and 2b). On approximate rules A, it is very obvious that CONCISE is better than Pasquier's approach (cf. Fig. 2a, 2b). The explanation is that all frequent are closed itemsets, that complicates the task of Pasquier's approach who performs more operations than CONCISE for counting frequent closed itemsets.

[1] http://www.almaden.ibm.com/cs/quest/syndata.html

[2] http://kdd.ics.uci.edu/

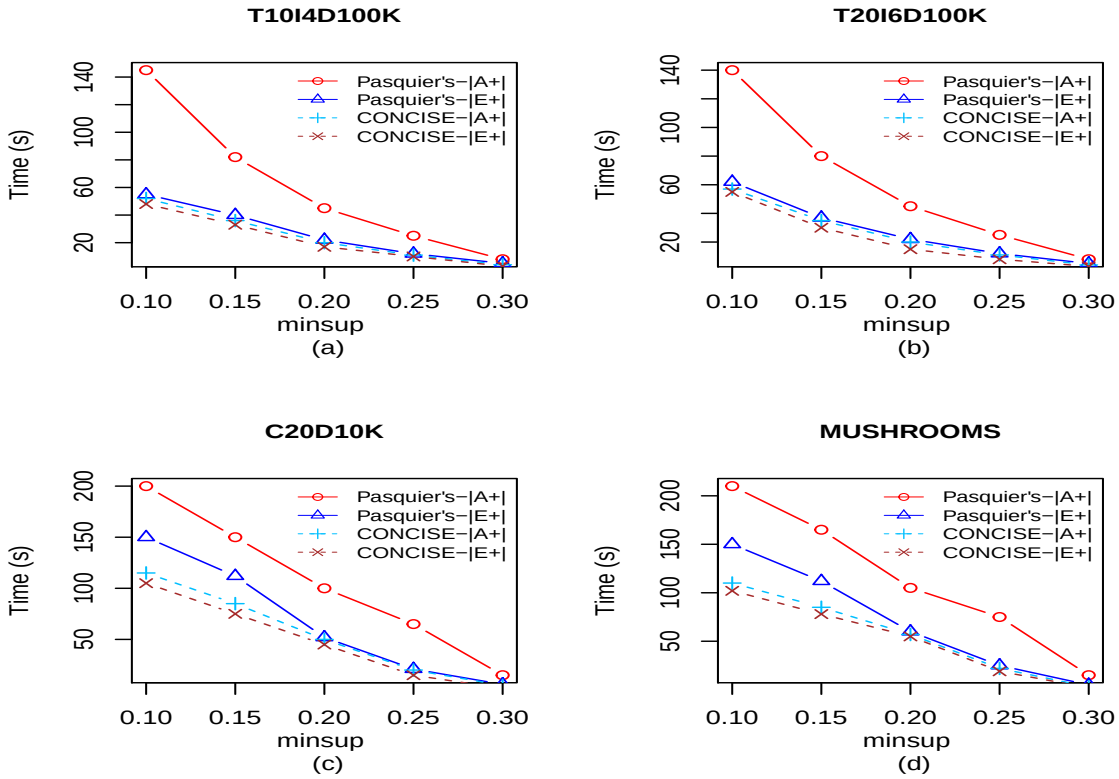

Figure 2. Response times by varying $minsup$ at fixed $\alpha = 0.05$ and $minconf = 0.6$

On dense databases (C20D10K and MUSHROOMS), CONCISE algorithm leads to significant average time compared to Pasquier's approach for all $minsup$ (cf. Figure 2c and Figure 2d). The main reason is associated to the technique for pruning search space of valid positive/negative association rules. Thanks to the different optimizations as defined on Subsection 4.2.3, CONCISE algorithm can reduce considerable amount the execution time for all minimum support threshold $minsup$, it is not the case for Pasquier's approach. The latter obtains the least performance. This is mainly due to the lack of techniques for pruning the search space for valid association rules. This obviously affects its execution time. However, Pasquier's approach joins CONCISE algorithm for the E execution times, when $minsup$ is 20% to 30%.

6 Conclusion

In this paper, we presented and evaluated a condensed representation for association rules. It is an efficient method for representing non-redundant positive and negative rules. We theoretically proved and experimentally confirmed that our approach can eliminate considerable amount of redundancy and uninteresting rules. Compared to the Pasquier's and Feno's approaches, our approach is not only a concise but also a lossless extraction of positive and negative association rules. From this, all informative association rules can be deduced. The perspective would be to extend this proposal in Graphs and Classification problems.

References

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

1. R. Agrawal, and R. Srikant, "Fast Algorithms for Mining Association Rules". In Proceedings of 20th VLDB Conference, pp. 487–499. Santiago, Chile (1994).
2. Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets". In CL'2000 international conference Computational Logic, pp. 972–986 (2000).
3. P. Bemarisika, and A. Totohasina, An Informative Base of Positive and Negative Association Rules on Big Data. In Proc. of BigData, pp. 2428–2437 (2019).
4. L. Cao, X. Dong, and Z. Zheng, "E-NSP: Efficient negative sequential pattern mining". In Artificial Intelligence, pp. 156–182 (2016).
5. X. Dong, H. Hao, L. Zhao, and T. Xu, "An efficient method for pruning redundant negative and positive association rules". In NEUCOM 2018. `https://doi.org/10.1016/j.neucom.2018.09.108` (2018).
6. X. Dong, G. Yongshun, and L. Cao, "F-NSP: A Fast Negative Sequential Patterns Mining Method with Self-adaption Data Storage Strategy". Pattern Rec.(2018).
7. D. Feno, J. Diatta, and A. Totohasina, "Galois Lattices and Based for M_{GK}-valid Association Rules". In Ben Yahia et al. (Eds.), CLA 2006, pp. 186–197, (2006).
8. B. Ganter, and R. Wille, "Formal concept analysis: Mathematical foundations". In Springer Verlag (1999).
9. R. Gras, J-C. Régnier, C. Marinica, and F. Guillet, "L'ASI, Méthode exploratoire et confirmatoire la recherche de causalités". In Cépadus Editions, pp. 11–40 (2013).
10. J.L. Guigues, and V. Duquenne, "Familles minimales d'implications informatives résultant d'un tableau de donnés binaires". Maths et Sci. Humaines, 5–18 (1986).
11. T. Guyet, and R. Quiniou, "NegPSpan: efficient extraction of negative sequential patterns with embedding constraints". `https://hal.inria.fr/hal-01743975v2` (2018).
12. M. Mannila, and H. Toivonen, "Levelwise Search and Borders of Theories in Knowledge Discovery". In Data Mining Knowledge Discovery, pp. 241–258 (1997).
13. N. Pasquier, R. Taouil, and Y. Bastide, G. Stumme, and L. Lakhal, "Generating a condensed representation for association rules". In J. of Intell. Info. Syst., pp. 29–60 (2005).
14. N. Pasquier, "Frequent Closed Itemsets Based Condensed Representations for Association Rules". In Tech. for Eff. Knowl. Extraction, pp. 248–273 (2009).
15. T. Xu, T. Li, and X. Dong, "Efficient High Utility Negative Sequential Patterns Mining in Smart Campusy". In IEEE Access, pp. 23839–23846, (2018).
16. Y. Xu, Y. Li, and G. Shaw, "Reliable representations for association rules". In Data and Knowledge Engineering, pp. 555–575 (2011).