An Efficient Method for Mining Informative Association Rules in Knowledge Extraction

Parfait Bemarisika, André Totohasina

An Efficient Method for Mining Informative Association Rules in Knowledge Extraction

Parfait Bemarisika[1] and André Totohasina[1]

Laboratoire de Mathématiques et Informatique
ENSET, Université d'Antsiranana, Madagascar
bemarisikap7@yahoo.fr, andre.totohasina@gmail.com

Abstract. Minining association rules is an important problem in Knowledge Extraction (KE). This paper proposes an efficient method for mining simultaneously informative positive and negative association rules, using a new selective pair support-M_{GK}. For this, we define four new bases of positive and negative association rules, based on Galois connetion semantics. These bases are characterized by frequent closed itemsets, maximal frequent itemsets, and their generator itemsets; it consists of the non-redundant exact and approximate association rules having minimal premise and maximal conclusion, i.e. the informative association rules. We introduce NONRED algorithm allowing to generate these bases and all valid informative association rules. Results experiments carried out on reference datasets show the usefulness of this approach.

Keywords: Knowledge Extraction · Minimal generators · Frequent closed itemsets · Maximal frequent itemsets · Informative basis of rules

1 Introduction and Motivations

Mining association rules is an important problem in Knowledge Extraction (KE). This paper focuses on informative basis for association rules which is a subset of non-redundant from which we can derive all valid rules. Given $\mathcal{I}$ a set of items from a context, an association rule is an implication of the form $X \rightarrow Y$, where $X, Y \subseteq \mathcal{I}$ and $X \cap Y = \emptyset$. The itemsets X and Y are called premise (or antecedent) and conclusion (or consequent) of this rule, respectively. An association rule is called *exact* if its intensity (M_{GK}) is equal to 1 and *approximate*, otherwise. An *informative* association rule is one that, from the minimal premise, provides the maximal consequent. A rule r_1 is said to be redundant with respect to r_2 if (i) it shares the same information as r_2, (ii) its premise (resp. conclusion) is superset of premise (resp. subset of conclusion) of the r_2.

In KE concept, the concept of bases for association rules has developed by several approaches [8,9,11,12]. However, most of them are not based on negative rules[1] but rather based on positive rules, and this, with less selective pair support-confidence [1]. However, the positive rules exclusively cannot cover all

[1] An association of the form $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$ and $\overline{X} \rightarrow \overline{Y}$, where $\overline{I} = \neg I = \mathcal{I} \backslash I$.

interest of mining association rules in databases, it also needs the negative rules. To tackle these notable limitations, we propose an efficient method for mining simultaneously positive and negative informative rules, based on Galois connetion, and this, with a selective pair support-M_{GK}. For this, we define four new bases: Informative Basis for Positive Exact Association Rules (IBE^+), Informative Basis for Positive Approximate Association Rules (IBA^+), Informative Basis for Negative Exact Association Rules (IBE^-), and Informative Basis for Negative Approximate Association Rules (IBA^-). Different optimizations for searching space are then developed. Based on these optimizations, we introduce NonRed algorithm for mining these bases and all valid informative association rules.

The rest of this paper is organized as follows. Section 2 discusses the related works. Section 3 gives the preliminaries, where we formally introduce the basic concepts relative to Galois connection and association rules, and the main properties of M_{GK} [10,16,18]. Section 4 details our approach. Section 5 describes the experimental evaluation. A conclusion and perspectives are given in Section 6.

2 Related works

In KE concept, minig association rules is structured on two lines of research: (1) Bases of positive association rules, and (2) Bases of negative association rules.

In base of positive assoiation rules, we present Duquenne-Guigues basis [9]. Without going into the details of its calculation, the Duquenne-Guigues basis is not informative because the premises are chosen from to pseudo-closed that is incompatible of minimality. Bastide et al. [2,15] adapts this Duquenne-Guigues basis. In order to derive the redundant approximate association rules, it uses the closure mechanism based on Galois lattices [7]. However, it has been shown in Kryszkiewicz [11] that this concept of closures mechanism is invalid and the Duquenne-Guigues basis [9] is not informative. To address this problem, Kryszkiewicz extends this same Duquenne-Guigues basis of exact and approximate association rules. However, this approach presents a significant number of association rules, especially for dense contexts, making it difficult to manipulate the result. Zaki [19] defines a generic basis for non-redundant association rules where it uses the transitivity axiom. However, the proposed definition of redundancy is not very appropriate, which is not conform of informative concept. Recently, [12] offers a new generic basis. However, the definition of a redundant association rules adopted does not guarantee the absolute minimal premise. This approach is then inappropriate on concept of the informative association rules.

The concept of negative basis is introduced in [6]. Despite its notable interest, this approach is not informative, because it selects the premises on pseudo-closed [9] which intuitively returns the maximal elements, so incompatible of minimality. In addition, its formulation of exact negative association rules is not appropriate, which can present a high memory for searching space.

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 notions

In KE concept, a formal context (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 $X \subseteq \mathcal{I}$, $\overline{X} = \neg X = \{t \in \mathcal{T} | \exists i \in X : (i,t) \notin \mathcal{R}\}$ is complementary set of X (i.e. $\mathcal{I} \backslash X$). A subset $X \subseteq \mathcal{I}$ with $k = |X|$ is called k-itemset, where $|\ell|$ denotes the cardinality of ℓ. For $I \subseteq \mathcal{I}$, the set $\phi(I) = I' = \{t \in \mathcal{T} | i\mathcal{R}t, \forall i \in I\}$ is called extension of I, it is the set of transactions that satisfy all items in I. Similarly, for $T \subseteq \mathcal{T}$, the set $\psi(T) = T' = \{i \in \mathcal{I} | i\mathcal{R}t, \forall t \in T\}$ is intension of T. Both functions ϕ and ψ form a Galois connection between $\mathcal{P}(\mathcal{I})$ and $\mathcal{P}(\mathcal{T})$ [7], where $\mathcal{P}(\ell)$ is a powerset lattice of ℓ. The support of $X \subseteq \mathcal{I}$ is $supp(X) = P(X') = \frac{|\phi(X)|}{|\mathcal{T}|}$ where P is the discrete probability on $(\mathcal{T}, \mathcal{P}(\mathcal{T}))$. X is frequent if $supp(X) \geqslant minsup$ where $minsup \in]0,1]$ is a minimum support. Let $\mathcal{F}$ be the set of all frequent on $\mathcal{B}$, i.e. $\mathcal{F} = \{X \subseteq \mathcal{I} | supp(X) \geqslant minsup\}$. $\gamma(X) = \psi o \phi(X)$ is called Galois closure operator. Formally, a rule $X \rightarrow Y$ is said to be exact if $X \subseteq Y$ and $\gamma(X) = \gamma(Y)$. It is approximate if $X \subseteq Y$ and $\gamma(X) \subset \gamma(Y)$. Two itemsets $X, Y \subseteq \mathcal{I}$ are said to be equivalent, denoted by $X \cong Y$, iff $supp(X) = supp(Y)$. The set of itemsets that are equivalent to an itemset X is denoted by $[X] = \{Y \subseteq \mathcal{I} \mid X \cong Y\}$.

Table 1: Context $\mathcal{B}$

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

Property 1 (Antimonotonicity). $\forall X, Y \subseteq \mathcal{I}$, we have $X \subseteq Y \Rightarrow \phi(X) \supseteq \phi(Y)$.

Definition 1 (Closed itemset). *An itemset $X \subseteq \mathcal{I}$ is closed iff $X = \gamma(X)$.*

Property 2 ([2,15,19]). For all $X, Y \subseteq \mathcal{I}$, we have $X \subset Y \Rightarrow \gamma(X) \subset \gamma(Y)$.

Definition 2 (Generators). *An itemset G is said to be generator of a closed itemset $\mathcal{C}$ iff $\gamma(G) = \mathcal{C}$ and $\nexists g \subseteq \mathcal{I}$ with $g \subseteq G$ such that $\gamma(g) = \mathcal{C}$. Consider $0 < minsup \leqslant 1$, we define the set $\mathcal{G}_{\mathcal{C}}$ of all frequent generator itemsets in $\mathcal{B}$ as:*

$$\mathcal{G}_{\mathcal{C}} = \{G \in [\mathcal{C}] | \mathcal{C} \in \mathcal{FC},\ \nexists g \subset G,\ supp(G) \geqslant minsup\}$$

An itemset $G \in [\gamma(G)]$ is called a generator, if G has no proper subset (ordered by $\subseteq$) in $[\gamma(G)]$ given by $[\gamma(G)] = \{Y \subseteq \mathcal{I} | \gamma(Y) = \gamma(G)\}$. In other words, it has no proper subset with the same support (i.e. the same closure).

Property 3 ([2,15,19]). For all $X \subseteq \mathcal{I}$ from the context $\mathcal{B}$, $supp(X) = supp(\gamma(X))$.

Definition 3 (Frequent closed itemsets). *The closed itemset $\mathcal{C}$ is said to be frequent if $supp(\mathcal{C}) \geqslant minsup$. We define the set $\mathcal{FC}$ of all frequent closed as:*

$$\mathcal{FC} = \{\mathcal{C} \in \mathcal{I} | \mathcal{C} = \gamma(\mathcal{C}),\ supp(\mathcal{C}) \geqslant minsup\}$$

Definition 4 (Maximal frequent closed itemsets). *Let $\mathcal{FC}$ be the set of all frequent closed. We define $\mathcal{MC}$ the set of all maximal frequent closed in $\mathcal{B}$ as:*

$$\mathcal{MC} = \{\mathcal{C} \in \mathcal{FC} \mid \nexists \widetilde{\mathcal{C}} \supset \mathcal{C},\ \widetilde{\mathcal{C}} \in \mathcal{FC}\}$$

For all $X, Y \subseteq \mathcal{I}$, the support and confidence of $X \to Y$ are respectively defined by $supp(X \cup Y) = \frac{|\phi(X \cup Y)|}{|\mathcal{T}|}$ and $P(Y'|X') = \frac{P(X' \cap Y')}{P(X')}$. Despite its notable contribution, the pair support-confidence generates a large number of rules many of which are uninteresting and redundant (see [3,4]). So, we use the selective pair support-M_{GK}. We define M_{GK} [16] of a rule $X \to Y$ as:

$$M_{GK}(X \to Y) = \begin{cases} \frac{P(Y'|X') - P(Y')}{1 - P(Y')}, \text{ if } P(Y'|X') > P(Y'),\ P(Y') \neq 1 \\ \frac{P(Y'|X') - P(Y')}{P(Y')}, \text{ if } P(Y'|X') \leqslant P(Y'),\ P(Y') \neq 0. \end{cases} \quad (1)$$

M_{GK} varies in $[-1, 1]$. When $-1 \leqslant M_{GK}(X \to Y) < 0$, then $P(Y'|X') \leqslant P(Y')$ (i.e. X disfavors Y) means that Y is negatively dependent on X. In this case, $X \to Y$ is not interesting rule. When $M_{GK}(X \to Y) = 0$, then $P(Y'|X') = P(Y')$ means that X and Y are independent. In this case, $X \to Y$ is also not interesting rule. When $M_{GK}(X \to Y) > 0$, then $P(Y'|X') > P(Y')$ (i.e. X favors Y) means that Y is positively dependent on X, so $X \to Y$ is interesting.

Formally, a rule $X_2 \to Y_2$ is redundant with respect to $X_1 \to Y_1$ iff (i) ($X_1 \subseteq X_2$ and $Y_1 \supset Y_2$) and (ii) ($supp(X_1 \cup Y_1) = supp(X_2 \cup Y_2)$, $M_{GK}(X_1 \to Y_1) = M_{GK}(X_2 \to Y_2)$). A rule $X \to Y$ is informative if $X \in \mathcal{G}_{\gamma(X)}$ and $Y \in \mathcal{FC}$.

Property 4. Given X and Y two itemsets of $\mathcal{I}$, if (X, Y) is a stochastically independent pair, so $(\overline{X}, Y)$, $(X, \overline{Y})$, ($\overline{X}$ and $\overline{Y}$) are also independent.

Proof. Let $X, Y \subseteq \mathcal{I}$. We have $P(\overline{X'})P(Y') - P(\overline{X'} \cap Y') = (1 - P(X'))P(Y') - (P(Y') - P(X' \cap Y')) = P(X' \cap Y') - P(X')P(Y')$. So, if $P(X' \cap Y') = P(X')P(Y')$, then $P(\overline{X'})P(Y') = P(\overline{X'} \cap Y')$. We have the same result for $(X, \overline{Y})$ and $(\overline{X}, \overline{Y})$, then replacing Y with $\overline{Y}$. □

Then, no rule can be interesting if (X, Y) is stochastically independent.

Property 5. For all $X, Y \subseteq \mathcal{I}$, we have $M_{GK}(X \to Y) = -M_{GK}(X \to \overline{Y})$.

Proof. (1) If $P(Y'|X') > P(Y')$, we have $P(Y'|X') + P(\overline{Y}|X) = 1 \Leftrightarrow P(\overline{Y}|X) = 1 - P(Y'|X') \leqslant 1 - P(Y') = P(\overline{Y})$. So, $M_{GK}(X \to Y) + M_{GK}(X \to \overline{Y}) = \frac{P(Y'|X') - P(Y')}{1 - P(Y')} + \frac{P(\overline{Y}|X) - P(\overline{Y})}{P(\overline{Y})} = \frac{P(Y'|X') - P(Y')}{1 - P(Y')} + \frac{1 - P(Y'|X') - 1 + P(Y')}{1 - P(Y')} = 0$. (2)
If $P(Y'|X') \leqslant P(Y')$, we have $P(Y'|X') + P(\overline{Y'}|X') = 1 \Leftrightarrow P(\overline{Y}|X) = 1 - P(Y'|X') \geq 1 - P(Y') = P(\overline{Y'})$. Thus, $M_{GK}(X \to Y) + M_{GK}(X \to \overline{Y}) = \frac{P(Y'|X') - P(Y')}{P(Y')} + \frac{P(\overline{Y'}|X') - P(\overline{Y'})}{1 - P(\overline{Y'})} = \frac{P(Y'|X') - P(Y')}{1 - P(Y')} + \frac{1 - P(Y'|X') - 1 + P(Y')}{P(Y')} = 0$. □

It follows that if the value of M_{GK} for the rule $X \to Y$ is strictly negative, we conclude that it is the rule $X \to \overline{Y}$ which will be pertinent and we have the value of this rule without having to recalculate it thanks to this notable property.

Property 6. For all $X, Y \subseteq \mathcal{I}$, we have: (1) $P(Y'|X') > P(Y') \Rightarrow 0 < M_{GK}(X \rightarrow Y) \leqslant 1$, and (2) $P(Y'|X') \leqslant P(Y') \Rightarrow -1 \leqslant M_{GK}(X \rightarrow Y) \leqslant 0$.

Proof. (1) Since $P(Y'|X') > P(Y')$ we have $P(Y'|X') - P(Y') > 0 \overset{P(Y')\neq 1}{\Leftrightarrow} \frac{P(Y'|X')-P(Y')}{1-P(Y')} > 0$. For each $X, Y \subseteq \mathcal{I}$: $P(Y'|X') \leqslant 1 \Leftrightarrow P(Y'|X') - P(Y') \leqslant 1 - P(Y') \overset{P(Y')\neq 1}{\Leftrightarrow} \frac{P(Y'|X')-P(Y')}{1-P(Y')} \leqslant 1$. So, $0 < M_{GK}(X \rightarrow Y) \leqslant 1$. (2) Since $P(Y'|X') \leqslant P(Y')$ we have $P(Y'|X') - P(Y') \leqslant 0 \Rightarrow -P(Y') \leqslant P(Y'|X') - P(Y') \leqslant 0 \overset{P(Y')\neq 0}{\Leftrightarrow} -1 \leqslant \frac{P(Y'|X')-P(Y')}{P(Y')} \leqslant 0 \Leftrightarrow -1 \leqslant M_{GK}(X \rightarrow Y) \leqslant 0$. □

From this property 6, we conclude that M_{GK} offers a robust and efficient tool to prune systematically the unintersting association rules (i.e. $-1 \leqslant M_{GK} \leqslant 0$). Note that the first component of M_{GK} (cf. equation (1)) is implicative but the second not, only the first will be active. This procedure is thus more consistent with the classic causal interpretation of an association rule. Let $n = |\mathcal{T}|$, $n_X = |\phi(X)|$, $n_Y = |\phi(Y)|$, $n_{X\wedge Y} = |\phi(X \cup Y)|$ and $n_{X\wedge\overline{Y}} = |\phi(X \cup \overline{Y}|$. Also, $N_{X\wedge\overline{Y}}$ indicates the random variable which generates $n_{X\wedge\overline{Y}}$, and $N_{X\wedge Y}$ that which generates $n_{X\wedge Y}$. For any association rule $X \rightarrow Y$, we define M_{GK} as:

$$M_{GK}(X \rightarrow Y) \overset{P(\overline{Y'})\neq 0}{=} 1 - \frac{P(\overline{Y'}|X')}{P(\overline{Y'})} = 1 - \frac{nn_{X\wedge\overline{Y}}}{n_X n_{\overline{Y}}} \tag{2}$$

In principle, such an association rule $X \rightarrow Y$ will be all the better if it contains large examples $N_{X\wedge Y}$ and relatively few counter-examples $N_{X\wedge\overline{Y}}$.

4 Mining Informative Association Rules

Our approach is based on the same theoretical basis as the approach proposed by [2]. However, our extraction strategy is different in terms of the measure used and the rules extracted. Indeed, our approach uses the selective couple support-M_{GK}, and simultaneously extracts the positive and negative rules. The current version [17] of the support-M_{GK} approach uses the critical value $\sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n_Y}{n-n_Y}\chi^2(\alpha)}$, i.e. a rule $X \rightarrow Y$ will be valid if $M_{GK}(X \rightarrow Y) \geqslant \sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n_Y}{n-n_Y}\chi^2(\alpha)}$ where $\chi^2(\alpha)$, at the risk of error $\alpha \in]0, 1]$, is the statistic of Chi-square to a degree of freedom. Despite its indisputable interest, this critical value can nevertheless present some faults. A low value (natural choice) of α leads to a high critical value which exponentially exceeds the real value of M_{GK}. This rejects the robust association rules. Conversely, a relatively large value of α leads to a very low critical value. This accepts the very bad rules.

In order to overcome these limits, we introduce a new technique which consists in modeling the number $N_{X\wedge\overline{Y}}$ of counter-examples of this rule $X \rightarrow Y$. In principle, such a rule is all the better when the number $N_{X\wedge\overline{Y}}$ of counter-examples to its formal validity is zero or sufficiently small. To do this, it should be noted that $\frac{\partial M_{GK}}{\partial n_{X\wedge\overline{Y}}} = -\frac{1}{\frac{n_X n_{\overline{Y}}}{n}}$ (cf. equation 2), which shows that M_{GK}

decreases if the counter-examples $n_{X\wedge\overline{Y}}$ increase and all the more quickly as the $\frac{n_X n_{\overline{Y}}}{n}$ is relatively small. It would therefore be reasonable to compare the counter-examples $n_{X\wedge\overline{Y}}$ to the $\frac{n_X n_{\overline{Y}}}{n}$. For lack of space, we were unable to detail our technique and only present here a very abstract way. Therefore, consider M_{GK} (cf. equation (2) as the realization of a random variable $\mathcal{K}$ defined by $\mathcal{K} = 1 - \frac{nN_{X\wedge\overline{Y}}}{N_X N_{\overline{Y}}}$. The latter takes its values in $[0,1]$. Its values are therefore fractions of the integer values whose numerators are Poissonian, and we obtain:

$$P\left(\mathcal{K} \geqslant \alpha\right) = P\left(1 - \frac{nN_{X\wedge\overline{Y}}}{N_X N_{\overline{Y}}} \geqslant \alpha\right) = P\left(N_{X\wedge\overline{Y}} \leqslant \frac{N_X N_{\overline{Y}}}{n}(1-\alpha)\right) \tag{3}$$

However in the contingency, the random variable N_X (resp. $N_{\overline{Y}}$) takes the fixed value and equal to n_X (resp. $n_{\overline{Y}}$), which brings the (3) to the equation (4):

$$P\left(\mathcal{K} \geqslant \alpha\right) = P\left(N_{X\wedge\overline{Y}} \leqslant \frac{n_X n_{\overline{Y}}}{n}(1-\alpha)\right) \tag{4}$$

We then establish the formula of $N_{X\wedge\overline{Y}}$ under the assumption H_0 of independence. Let $U, Z \subseteq \mathcal{I}$ be chosen randomly and independently of the same cardinals of X and Y respectively. The modeling of $N_{X\wedge\overline{Y}}$ depends on the distribution of $|\phi(Z \cup \overline{U})|$ which follows a Poisson distribution of parameter $\frac{n_X n_{\overline{Y}}}{n}$ [13]. We center and reduce the $N_{X\wedge\overline{Y}}$ into the variable $Q(X,\overline{Y})$ called contingent observation, we get $Q(X,\overline{Y}) = \dfrac{N_{X\wedge\overline{Y}} - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{\frac{n_X n_{\overline{Y}}}{n}}} \sim \mathcal{N}(0,1)$. From where we get

$$P\left(N_{X\wedge\overline{Y}} \leqslant \frac{n_X n_{\overline{Y}}}{n}(1-\alpha)\right) = P\left(Q(X,\overline{Y})\sqrt{\frac{n_X n_{\overline{Y}}}{n}} + \frac{n_X n_{\overline{Y}}}{n} \leqslant \frac{n_X n_{\overline{Y}}}{n}(1-\alpha)\right).$$

Let $q(X,\overline{Y}) = \dfrac{\frac{n_X n_{\overline{Y}}}{n}(1-\alpha) - \frac{n_X n_{\overline{Y}}}{n}}{\sqrt{\frac{n_X n_{\overline{Y}}}{n}}}$, contingent realization of $Q(X,\overline{Y})$. So,

$$\begin{aligned} P\left(\mathcal{K} \geqslant \alpha\right) = P\left(Q(X,\overline{Y}) \leqslant q(X,\overline{Y})\right) &= \Phi(q(X,\overline{Y})) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{q(X,\overline{Y})} e^{-\frac{t^2}{2}}\, \mathrm{d}t, \end{aligned} \tag{5}$$

where Φ is the normal distribution function $\mathcal{N}(0,1)$. This result (equation (5)) is called M_{GK} `Gaussianized`. It will be used especially in the evaluation of approximate association rules but also in that of the exact ones. If we want $X \to Y$ to be likely to be significant, it is necessary that the number $N_{X\wedge\overline{Y}}$ of counterexamples is not greater than the corrective factor $\frac{n_X n_{\overline{Y}}}{n}(1-\alpha)$ and will be all the better as the probability $P\left(N_{X\wedge\overline{Y}} \leqslant \frac{n_X n_{\overline{Y}}}{n}(1-\alpha)\right)$ will be close to 1. More formally, once we have specified the degree of dependence $(P(Y'|X') - P(Y'))$ of the candidate association rule $X \to Y$ and the different thresholds

($0 < minsup \leqslant 1$ and $0 < \alpha \leqslant 1$) the proposed approach must generate the set:

$$\left\{ X \rightarrow Y \text{ informative sur } \mathcal{B} \mid \left\{ \begin{array}{l} (i)\ supp(X \cup Y) \geqslant minsup \\ (ii)\ N_{X \wedge \overline{Y}} \leqslant \frac{n_X n_{\overline{Y}}}{n}(1-\alpha) \end{array} \right. \right\}$$

This task will be carried out in two stages: (i) Extraction of all frequent itemsets on a context $\mathcal{B}$, (ii) Generation of all informative association rules of the type $X \rightarrow Y$ from these frequent itemsets whose number $N_{X \wedge \overline{Y}}$ of counter-examples is not greater than the corrective factor $\frac{n_X n_{\overline{Y}}}{n}(1-\alpha)$ for a α chosen in $]0, 1]$.

4.1 Mining all frequent itemsets

From a context $\mathcal{B}$, this first phase consists in determining $\mathcal{GC}$, $\mathcal{FC}$ and $\mathcal{MC}$. Our main motivation lies in the absence of an autonomous algorithm to determine these three subsets. So, we propose GCM (GeneratorClosedMaximal), an autonomous algorithm allowing to collect these three subsets simultaneously. Close algorithm [15] partly solves this problem: it returns the closed ones while here we want to list all the frequent ones, whether they are generators, closeds or maximal closeds. We therefore modified Close by adding the concepts of maximal frequent and their generators. These two concepts can be derived from a frequent closed using the definitions 2 and 4. In addition, we introduce an efficient technique to calculate the supports that the following theorem 1 exposes.

Theorem 1. *The support of a k-nongenerator ($k \geqslant 3$) is a minimum support of its $(k-1)$-itemsets: if X is a k-nongenerator, then $supp(X) = min\{supp(\widetilde{X}) | \widetilde{X} \subset X\}$.*

Proof. Let X and Z two itemset of $\mathcal{I}$ such that $Z \subseteq X$. Since $Z \subseteq X$, we have $\phi(Z) \supseteq \phi(X) \Rightarrow supp(Z) \geqslant supp(X)$. If X is not generator, it exists $\widetilde{X} \subseteq X$ such that $supp(\widetilde{X}) = supp(X)$. However, $supp(Z)$ is minimal in $\mathcal{I}$, so $supp(Z) \leqslant supp(\widetilde{X})$. Finally, $supp(X) = supp(Z) = min\{supp(\widetilde{X}) | \widetilde{X} \subset X\}$. □

Thus, the support of a k-nongenerator ($k \geqslant 3$) can be derived from $(k-1)$ subsets (proofs are omitted due to lack of space, we can consult [3,4]). From table 1, we have $supp(ABC) = min(\{sup(AB), sup(AC), sup(BC)\}) = min\{2/6, 3/6, 4/6\} = 2/6$. This Theorem 1 is therefore very central: it avoids systematic access in a context made by existing approaches. GCM browses by level the search space where the enumeration follows the theorems 2 and 3 below.

Corollary 1. *Let X a frequent itemset on formal context $\mathcal{B}$. The itemset X is a generator iff $supp(X) \neq min\{supp(\widetilde{X}) | \widetilde{X} \subset X\}$.*

Proof. Let X be a generator. Let $\widetilde{X}$ be a frequent itemset of length $k-1$ with minimum support and a subset of X. Then, $\widetilde{X} \subset X \Rightarrow \phi(\widetilde{X}) \supseteq \phi(X)$. If $\phi(\widetilde{X}) = \phi(X)$, then $supp(\widetilde{X}) = supp(X)$ and X is not a generator. Moreover, it is not the element with the smallest support, whose closure is $\gamma(X)$. This concludes that $\phi(\widetilde{X}) \supseteq \phi(X)$ and hence, $supp(X) \neq min\{supp(\widetilde{X}) | \widetilde{X} \subset X\}$. On the other hand, if $supp(X) \neq min\{supp(\widetilde{X}) | \widetilde{X} \subset X\}$, the itemset X is then the smallest element of the closure $\gamma(X)$. Hence, X is a generator. □

Theorem 2 (Intuitive in [1]). *(1) All subsets of a frequent itemset are frequent. (2) All supersets of an infrequent itemset are infrequent.*

Proof. (1) Let $X, Y \subseteq \mathcal{I}$ such that $X \in \mathcal{F}$ and $Y \subseteq X$. Since $Y \subseteq X$, we have (by antimonotonicity of support) $\phi(Y) \supseteq \phi(X) \Rightarrow supp(Y) \geqslant supp(X) \geqslant minsup \Rightarrow Y \in \mathcal{F}$. (2) Let $X, Y \subseteq \mathcal{I}$ such that $X \notin \mathcal{F}$ and $Y \supseteq X$. Since $Y \supseteq X$, we have $\phi(Y) \subseteq \phi(X) \Rightarrow supp(Y) \leqslant supp(X) \leqslant minsup \Rightarrow Y \notin \mathcal{F}$. □

Theorem 3. *Given an itemset $X \subseteq \mathcal{I}$. If X is generator, then $\forall Y \subseteq X$ is also generator. If X is not generator, then Z is not generator, $\forall Z \supseteq X$.*

Proof. Let $X, Z \subseteq \mathcal{I}$ such that $X \subseteq Z$. It exists $Y \subseteq \mathcal{I}$ such that $X \cap Y = \emptyset$ and $Z = X \cup Y$. Consider X is not generator, it admits then a proper subset T which is equivalent to $T \subseteq X$ and $T \approx X$ implie that $T \cup Y \approx X \cup Y$. By hypothesis, $X \cap Y = \emptyset$, so $T \cup Y \subseteq X \cup Y$, The itemset Z is equivalent to a proper subset $T \cup Y$, so it is not generator. The contrapose gives the result. □

These results will be synthesized in the algorithm 1 below. The GCM algo-

Algorithm 1 GCM

Require: A formal context $\mathcal{B}$, A minimum support threshold $minsup \in]0, 1]$.
Ensure: $\mathcal{G}_\mathcal{C}$ all frequent generators, $\mathcal{FC}$ all frequent closed, $\mathcal{MC}$ all maximal closed.
1: $\mathcal{FCC}_1$.GENERATORS ← {1-itemsts}
2: **for all** $(k \leftarrow 1; \mathcal{FCC}_k$.GENERATORS $\neq \emptyset; k++)$ **do**
3: $\quad \mathcal{FCC}_k.closure \leftarrow \emptyset$; $\mathcal{FCC}_k.support \leftarrow 0$;
4: $\quad \mathcal{FCC}_k \leftarrow$ GENERATECLOSURES$(\mathcal{FCC}_k)$
5: $\quad$ **for all** (candidate itemsets $c \in \mathcal{FCC}_k$) **do**
6: $\quad\quad$ **if** $(supp(c) \geqslant minsup)$ **then**
7: $\quad\quad\quad \mathcal{FC}_k \leftarrow \mathcal{FC}_k \cup \{c\}$
8: $\quad\quad$ **end if**
9: $\quad$ **end for**
10: $\quad \mathcal{FCC}_{k+1} \leftarrow$ GENERATEGENERATORS$(\mathcal{FC}_k)$
11: $\quad \mathcal{FCC}_{k+1} \leftarrow$ GENERATEMAXIMAL$(\mathcal{FC}_k)$
12: **end for**
13: $\mathcal{FC} \leftarrow \bigcup_{j=1}^{k-1} \{\mathcal{FC}_j$.CLOSURE, $\mathcal{FC}_j$.support$\}$
14: **return** $\mathcal{FC}$

rithm takes as input a context $\mathcal{B}$ and a minimum support *minsup*. It simultaneously returns three subsets $\mathcal{G}_\mathcal{C}$, $\mathcal{FC}$ and $\mathcal{MC}$ in three recursive procedures. This choice of decomposition of the algorithms is motivated by the parallelization of these three procedures during the implementation to get these three types of frequent itemsets. Due to lack of space, we could not detail this algorithm 1 and only present a very global way. Thus, the procedure GENERATECLOSURES (line 4) takes as input $\mathcal{FCC}_k$ the k-frequent closed candidates, and generates the frequent closed ones. The procedure GENERATEGENERATORS (line 10) takes as input $\mathcal{FC}_k$ the k-frequent closed itemsets, and generates all frequent generators. The GENERATEMAXIMAL procedure (line 11) also takes $\mathcal{FC}_k$ as input, and generates all maximal frequent itemsets. Figure 1 shows its example of execution with a small context from table 1 and $minsup = 2/6$. The set $\mathcal{FCC}_1$ initialized with the list of all 1-itemsets. The 1-itemset D is pruned from $\mathcal{FCC}_1$ to $\mathcal{FC}$

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

$\mathcal{FCC}_1$		$supp$
A	AC	3/6
B	BE	5/6
C	C	5/6
D	ACD	min(3/6,1/6,1/6)=1/6
E	BE	5/6

Pruning the infrequent $\longrightarrow$

$\mathcal{G}_C$	$\mathcal{FC}$	$\mathcal{MC}$	$supp$
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	min(2/6,2/6,2/6,4/6)=2/6
AE	ABCE	min(2/6,2/6,2/6,4/6)=2/6
BC	BCE	min(4/6,5/6,4/6)=4/6
CE	BCE	min(4/6,5/6,4/6)=4/6

Pruning infr. $\longrightarrow$

$\mathcal{G}_C$	$\mathcal{FC}$	$\mathcal{MC}$	$supp$
AB	ABCE	ABCE	2/6
AE	ABCE	ABCE	2/6
BC	BCE	-	4/6
CE	BCE	-	4/6

Fig. 1: Example of execution of GCM on context $\mathcal{B}$ with a $minsup = 2/6$

because it is not frequent. The 1-itemset C is also own closure, it's added in $\mathcal{FC}$. Six candidates AB, AC, AE, BC, BE and CE are created. Thus, AC and BE are closures of A and, B and E, so added to $\mathcal{FC}$. ABE is created but it's pruned because its subset $BE \notin \mathcal{FC}C_2$. BCE is closure of BC and CE, so added to $\mathcal{FC}$. BC and CE are then generators, so added in $\mathcal{G}_C$. Finally, $ABCE$ is both closed and maximal of AB and AE, so it is added both in $\mathcal{FC}$ and $\mathcal{MC}$.

4.2 Mining all informative association rules

From three subsets $\mathcal{G}_C$, $\mathcal{FC}$ and $\mathcal{MC}$, the second phase of our approach allows us to extract the set of all informative association rules of the form $X \rightarrow Y$ where $X \in \mathcal{G}_C$ and $Y \in \mathcal{FC}$ (or $\mathcal{MC}$). During the selection phase of association rules, many uninteresting and redundant rules can be generated. Fortunately, using Property 6, the uninteresting association rules are systematically pruned. It remains to study in the following the redundant association rules. For this, we rely on a new concept called *bases of rules*. We propose a new concept in which we define the four informative bases IBE^+, IBA^+, IBE^- and IBA^-.

From each base, we first identify the degree of dependence of the itemsets, and then measure its intensity using the measure M_{GK}. For any rule $X \rightarrow Y$, the degree of dependence is obtained by measuring the difference in confidence $P(Y'|X')$ to the probability $P(Y')$, i.e. $P(Y'|X') - P(Y')$. If the difference is positive (i.e. $P(Y'|X') > P(Y')$), then Y is positively dependent on X, this means that $X \rightarrow Y$ is interesting rule. Otherwise, $X \rightarrow Y$ is not interesting because we have more counter-examples than examples. In this case, the negative rule $X \rightarrow \overline{Y}$ which could be interesting. This avoids the independence of 8 rules (total number) $X \rightarrow Y$, $Y \rightarrow X$, $\overline{X} \rightarrow \overline{Y}$, $\overline{Y} \rightarrow \overline{X}$, $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $\overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$, and leads the algorithm into two recursives substeps. Indeed, if X and Y are positively dependent, then these are the four rules $X \rightarrow Y$, $\overline{X} \rightarrow \overline{Y}$, $Y \rightarrow X$ and $\overline{Y} \rightarrow \overline{X}$ which will be evaluated. Otherwise, these are the four opposite rules $X \rightarrow \overline{Y}$, $\overline{X} \rightarrow Y$, $Y \rightarrow \overline{X}$ and $\overline{Y} \rightarrow X$ which will be evaluated.

As we demonstrated in [4], we only retain 4/8 rules $X \rightarrow Y$, $\overline{X} \rightarrow \overline{Y}$, $X \rightarrow \overline{Y}$ and $\overline{X} \rightarrow Y$, the other four rules are semantically redundant. Among these four

retained, we study only two $X \to Y$ and $X \to \overline{Y}$ because $\overline{X} \to \overline{Y}$ (resp. $\overline{X} \to Y$) can be derived from $X \to Y$ (resp. $X \to \overline{Y}$) [4]. This gives a 75% reduction of the search space. Based on two rules, we then formalize our four informative bases. The first concerns a base of positive exact association rules. The similar approaches has been developed in [2,6]. However, these approaches are not informative. Thus, we propose the new base IBE^+ which selects the premise (resp. consequent) in $\mathcal{G}_\mathcal{C}$ set of generators (resp. $\mathcal{FC}$ set of frequent closed).

Definition 5. *Let $\mathcal{FC}$ be the set of all frequent closed and, for each frequent closed itemset $\mathcal{C}$, $\mathcal{G}_\mathcal{C}$ denotes the set of generators of closed $\mathcal{C}$, we have:*

$$IBE^+ = \{G \to \mathcal{C}\backslash G \mid G \in \mathcal{G}_\mathcal{C},\ \mathcal{C} \in \mathcal{FC},\ G \neq \mathcal{C}\} \tag{6}$$

From Table 1, finded that $A \subset AC$ and $\gamma(A) = \gamma(AC)$ implies that $A \to C$ is candidate. We recall, if $P(Y'|X') > P(Y')$ then, because $P(Y'|X') = 1 \Leftrightarrow P(Y'|X') - P(Y') = 1 - P(Y') \Leftrightarrow \frac{1-P(Y')}{1-P(Y')} = 1 \Leftrightarrow M_{GK}(X \to Y) = 1$. Therefore, $P(C'|A') = \frac{|\phi(A \cup C)|}{|\phi(A)|} = \frac{|\phi(A)|}{|\phi(A)|} = 1 \Leftrightarrow M_{GK}(A \to C) = 1$, hence $A \to C \in IBE^+$.

Like any (informative) basis, the IBE^+ basis is a minimal. In order to derive the other informative association rules, we propose the theorem 4 below.

Theorem 4. *(i) All valid positive exact rules and their supports can be derived from the IBE^+ basis. (ii) All rules in the IBE^+ are non-redudant exact rules.*

Proof. (i) Let $r_1 : X_1 \to Y_1\backslash X_1$ be a positive exact rule between two frequent itemsts with $X_1 \subset Y_1$. Since $M_{GK}(r_1) = 1$, we have $supp(X_1) = supp(Y_1)$. By Property 5, we derived that $supp(\gamma(X_1)) = supp(\gamma(Y_1)) \Rightarrow \gamma(X_1) = \gamma(Y_1) = \mathcal{C}$. The itemset $\mathcal{C}$ is a frequent closed itemset (i.e. $\mathcal{C} \in \mathcal{FC}$) and, Obviously, there exists a rule $r_2 : G \to \mathcal{C}\backslash G \in IBE^+$ 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 $\gamma(X_1) = \gamma(Y_1) = \mathcal{C}$, we deduce that $supp(r_1) = supp(\gamma(X_1)) = supp(\gamma(Y_1)) = supp(\mathcal{C}) = supp(r_2)$. Since $G \subseteq X_1 \subset Y_1 \subseteq \mathcal{C}$, then the rule r_1 can be derived from the rule r_2.

(ii) Let $r_2 : G \to \mathcal{C}\backslash G \in IBE^+$. According to definition 5, we have $G \in \mathcal{G}_\mathcal{C}$ and $\mathcal{C} \in \mathcal{FC}$. We demonstrate that there is no other rule $r_3 : X_3 \to Y_3\backslash X_3 \in IBE^+$ such as $supp(r_3) = supp(r_2)$, $M_{GK}(r_3) = M_{GK}(r_2)$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$ then, according to definition 2, we have $\gamma(X_3) \subseteq \gamma(G) = \mathcal{C} \Rightarrow X_3 \notin \mathcal{G}_\mathcal{C}$ and then $r_3 \notin IBE^+$. If $\mathcal{C} \subseteq Y_3$, we have $\mathcal{C} = \gamma(\mathcal{C}) = \gamma(G) \subset Y_3 = \gamma(Y_3)$. From definition 2, we deduce that $G \notin \mathcal{G}_{Y_3}$ and conclude that $r_3 \notin IBE^+$. □

The second base is the base of positive approximate rules (i.e. $0 < M_{GK}(X \to Y) < 1$). The existing bases [2,6] using the pseudo-closed [2,9,15] are not informative. Thus, we propose the new informative base IBA^+, which selects the premise in $\mathcal{G}_\mathcal{C}$ and the conclusions in other $\mathcal{FC}$ containing this current closed F.

Definition 6. *Let $\mathcal{FC}$ the set of frequent closed and, for each frequent closed $\mathcal{C}$, $\mathcal{G}_\mathcal{C}$ indicates the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$IBA^+ = \{G \to \mathcal{C}\backslash G | (G, \mathcal{C}) \in \mathcal{G}_{\gamma(G)} \times \mathcal{FC}, \gamma(G) \subset \mathcal{C}, \frac{n_G n_{\overline{\mathcal{C}}}}{n}(1-\alpha) \geqslant N_{G\overline{\mathcal{C}}}\} \tag{7}$$

Since $BE \subset ABCE$, then $B \rightarrow ACE$ and $E \rightarrow ABC$ are candidates. Here, $n = 6$, $n_B = n_E = 5$, $n_{\overline{ACE}} = n_{\overline{ABE}} = 4$, $n_{B\overline{ACE}} = n_{E\overline{ACE}} = 5 - 2 = 3$, and $\sqrt{\frac{n_E n_{\overline{ABC}}}{n}} = 1.82$. Consider $\alpha = 1\%$, we have $\frac{n_B n_{\overline{ACE}}}{n}(1-\alpha) = \frac{n_E n_{\overline{ABC}}}{n}(1-\alpha) = 3.3 > 3 = n_{B\overline{ACE}}$ and $P\left(N_{B\wedge\overline{ACE}} \leqslant \frac{n_B n_{ACE}}{n}(1-\alpha)\right) = 0.49 \Rightarrow \{B \rightarrow ACE, E \rightarrow ABC\} \in IBA^+$, i.e. the association rules $B \rightarrow ACE$ and $E \rightarrow ABC$ are valid (99% likely) in IBA^+ basis with probability 0.49.

Lemma 1. *For all $X, Y, T, Z \subseteq \mathcal{I}$, such that X favors Y and Z favors T, and $X \cap Y = Z \cap T = \emptyset$, and $X \subset Z \subseteq \gamma(X)$, and $Y \subset T \subseteq \gamma(Y)$. Then, $supp(X \cup Y) = supp(Z \cup T)$ and $M_{GK}(X \rightarrow Y) = M_{GK}(Z \rightarrow T)$.*

Proof. $\forall X, Y, T, Z \subseteq \mathcal{I}$, $supp(X \cup Y) = \frac{|\phi(X \cup Y)|}{|\mathcal{T}|} = \frac{|\phi(X) \cap \phi(Y)|}{|\mathcal{T}|}$ and $supp(Z \cup T) = \frac{|\phi(X \cup T)|}{|\mathcal{T}|} = \frac{|\phi(Z) \cap \phi(T)|}{|\mathcal{T}|}$. Since $X \subset Z \subseteq \gamma(X)$ and $Y \subset T \subseteq \gamma(Y)$, we have $supp(X) = supp(Z)$ and $supp(Y) = supp(T)$ implies $supp(X \cup Y) = supp(Z \cup T)$. From $supp(X) = supp(Z)$ and $supp(Y) = supp(T)$, we have $P(Y'|X') = P(T'|Z') \Leftrightarrow P(Y'|X') - P(Y') = P(T'|Z') - P(Y')$ equivalent to $\frac{P(Y'|X') - P(Y')}{1 - P(Y')} = \frac{P(T'|Z') - P(T')}{1 - P(T')} \Leftrightarrow M_{GK}(X \rightarrow Y) = M_{GK}(Z \rightarrow T)$. □

Theorem 5 below is proposed in order to derive the other valid rules.

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

Proof. (*i*) Let $r_1 : X_1 \rightarrow Y_1 \backslash X_1$ be a valid positive approximate rule between two frequent itemsts with $X_1 \subset Y_1$. Since $M_{GK}(r_1) < 1$, we also have $\gamma(X_1) \subset \gamma(Y_1)$. For any frequent itemsets 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 $r_2 : G_1 \rightarrow (\gamma(G_2)) \backslash G_1 \in IBA^+$. We show that the rule r_1 and its support can be derived from the rule r_2, its support and its M_{GK}. 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 (cf. Lemma 1), $supp(G_1) = supp(X_1)$ and $supp(G_2) = supp(Y_1) = supp(\gamma(G_2))$. According to Lemma 1, we have $supp(X_1 \cup Y_1) = supp(G_1 \cup \gamma(G_2))$ (i.e. $supp(r_1) = supp(r_2)$) and we thus deduce that $M_{GK}(X_1 \rightarrow Y_1) = M_{GK}(G_1 \rightarrow \gamma(G_2))$ (i.e. $M_{GK}(r_1) = M_{GK}(r_2)$).

(*ii*) Let $r_2 : G \rightarrow \mathcal{C} \backslash G \in IBA^+$. According to definition 6, we have $\mathcal{C} \in \mathcal{FC}$ and $G \in \mathcal{G}_C$ such as $C \subset \mathcal{C}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow Y_3 \backslash X_3 \in IBA^+$ such as $supp(r_3) = supp(r_2)$, $M_{GK}(r_3) = M_{GK}(r_2)$, $X_3 \subseteq G$ and $\mathcal{C} \subseteq Y_3$. If $X_3 \subseteq G$ then, by definition 2, we have $\gamma(X_3) \subset \gamma(G) = C$ and then $X_3 \notin \mathcal{G}_C$. We deduce that $supp(X_3) > supp(G)$ and then $M_{GK}(r_3) < M_{GK}(r_2)$. If $\mathcal{C} \subseteq Y_3$, we have (by definition 1) $\mathcal{C} = \gamma(\mathcal{C}) \subset Y_3 = \gamma(Y_3)$. We deduce that $supp(\mathcal{C}) > supp(Y_3)$ and conclude that $M_{GK}(r_2) > M_{GK}(r_3)$. □

The third model concerns a base of negative exact association rules (i.e. $M_{GK}(X \rightarrow \overline{Y}) = 1$). A reference approach [6] for this is not appropriate: it selects a premise from to maximal elements which is not adapted of minimality

concept, so it's not informative. Thus, we propose an informative base IBE^- which selects a premise in generator of a maximal frequent itemset $\mathcal{M} \in \mathcal{MC}$ and a consequent from to dualization in $\mathcal{I}$ of this maximal $\mathcal{M}$ (i.e. $\mathcal{I}\backslash\mathcal{M}$).

Definition 7. *Given $\mathcal{MC}$ the set of maximal frequent itemset and, for each $\mathcal{M} \in \mathcal{MC}$, $\mathcal{G}_{\mathcal{M}}$ denotes the set of generators of $\mathcal{M}$, we have:*

$$IBE^- = \{G \rightarrow \{\overline{y}\} \mid G \in \mathcal{G}_{\mathcal{M}},\ \mathcal{M} \in \mathcal{MC},\ y \in \mathcal{I}\backslash\mathcal{M}\} \tag{8}$$

For example from Table 1 if $minsup = 0.2$, we finded $\mathcal{MC} = \{ABCE\}$ and $\mathcal{I}\backslash\mathcal{M} = \{\overline{ABCE}\} = \{D\}$. We see that AB and AE are generators of $ABCE$ implies that $AB \rightarrow \overline{D}$ and $AE \rightarrow \overline{D}$ are candidates. Indeed, $supp(AB\overline{D}) = supp(AB) - supp(ABD) = 2/6 - 0 \Rightarrow P(\overline{D'}|\{AB\}') = \frac{2/6}{2/6} = 1$ equivalent to $M_{GK}(AB \rightarrow \overline{D}) = 1 \Rightarrow AB \rightarrow \overline{D} \in IBE^-$. Likewise for rule $AE \rightarrow \overline{D}$.

Corollary 2. *For all $X, Y \subseteq \mathcal{I}$: $supp(X \cup Y) = 0 \Leftrightarrow M_{GK}(X \rightarrow \overline{Y}) = 1$.*

Proof. Since $supp(X \cup Y) = 0$, we have $|\phi(X \cup Y)| = 0$ (because $|\phi(X)| \neq 0$) equivalent to $P(Y'|X') = 0 \Leftrightarrow P(\overline{Y}|X) = 1$ equivalent to $P(\overline{Y}|X) - P(\overline{Y}) = 1 - P(\overline{Y})$ equivalent to $\frac{P(\overline{Y}|X) - P(\overline{Y})}{1 - P(\overline{Y})} = 1$ equivalent to $M_{GK}(X \rightarrow \overline{Y}) = 1$ □

Theorem 6. *(i) All valid negative exact association rules, their supports and M_{GK}, can be derived from the rules of the IBE^- basis. (ii) All association rules in the IBE^- basis are non-redudant negative exact association rules.*

Proof. (i) Let $r_1 : X_1 \rightarrow \overline{Y}_1\backslash X_1$ be a valid negative exact rule between two frequent itemsts with $X_1 \subset \overline{Y_1} \subseteq \mathcal{M}$ (i.e. $Y_1 = \mathcal{I}\backslash\mathcal{M}$), where $\mathcal{M}$ is a frequent maximal itemset. Since $M_{GK}(r_1) = 1$, we have, according to Corollary 2, $supp(X_1 \cup Y_1) = 0 \Rightarrow supp(X_1 \cup \overline{Y}_1) = supp(X_1) = supp(\overline{Y_1})$. By Property 5, we derived that $supp(\gamma(X_1 \cup \overline{Y_1})) = supp(\gamma(X_1)) = supp(\gamma(\overline{Y_1})) \Rightarrow \gamma(X_1 \cup \overline{Y}_1) = \gamma(X_1) = \gamma(\overline{Y_1}) = \mathcal{M}$ (a). Obviously, there exists a rule $r_2 : G \rightarrow \overline{y}\backslash G \in IBE^+$ such that G is a generator of $\mathcal{M}$ for which $G \subseteq X_1$ and $G \subseteq \overline{Y_1}$, and thus, according to definition 7, $G \subseteq \overline{y}$. We show that the rule r_1 and its supports can be derived from the rule r_2 and its supports. Since $G \rightarrow \overline{y}\backslash G \in IBE^+$, we have, according to Corollary 2, $supp(G \cup \overline{y}) = supp(G)$. By Property 5, we deduced that $supp(\gamma(G \cup \overline{y})) = supp(\gamma(G)) = supp(\gamma(\overline{y})) \Rightarrow \gamma(G \cup \overline{y}) = \gamma(G) = \gamma(\overline{y}) = \mathcal{M}$ (a'). From (a) and (a'), we have $\gamma(G \cup \overline{y}) = \gamma(X_1 \cup \overline{Y_1}) \Rightarrow supp(G \cup \overline{y}) = supp(X_1 \cup \overline{Y_1})$. Since $G \subseteq X_1 \subset \overline{Y_1} \subset \overline{y} \subseteq \gamma(G) = \mathcal{M}$, we have $supp(G) = supp(X_1) = supp(\overline{Y_1}) = supp(\overline{y}) = supp(\mathcal{M})$ and thus $M_{GK}(X_1 \rightarrow \overline{Y_1}) = M_{GK}(G \rightarrow \overline{y})$.

(ii) Let $r_2 : G \rightarrow \overline{y}\backslash G \in IBE^-$ be a valid negative exact rule. According to definition 7, we have $G \in \mathcal{G}_{\mathcal{M}}$ and $y \in \mathcal{I}\backslash\mathcal{M}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y}_3\backslash X_3 \in IBE^-$ such as $supp(r_3) = supp(r_2)$, $M_{GK}(r_3) = M_{GK}(r_2)$, $X_3 \subseteq G$ and $\overline{y} \subseteq \overline{Y_3}$. If $X_3 \subseteq G$ then, according to definition 2, we have $\gamma(X_3) \subseteq \gamma(G) \subset \gamma(\overline{y}) = \mathcal{M} \Rightarrow X_3 \notin \mathcal{G}_{\mathcal{M}}$ and then $r_3 \notin IBE^-$. If $\overline{y} \subseteq \overline{Y_3}$ then, according to definition 1, we have $\gamma(G) \subset \gamma(\overline{y}) \subseteq \gamma(\overline{Y_3}) = \mathcal{M}$. From definition 2, we deduce that $G \notin \mathcal{G}_{\overline{Y_3}}$ and conclude that $r_3 \notin IBE^-$. □

The fourth model address on negative approximate association rules (i.e. $M_{GK}(X \to \overline{Y}) < 1$). A similar approach is the one defined in [6]. However, this approach uses the pseudo-closed [2,15] which is not informative. To tackle this notable limitation, we propose the new informative base IBA^- which selects both premise and conclusion in generator of incomparable closed itemsets.

Definition 8. *Let $\mathcal{FC}$ be the set of frequent closed itemsets and, for each frequent closed $\mathcal{C}$, $\mathcal{G}_\mathcal{C}$ is the set of generators of $\mathcal{C}$. Consider $0 < \alpha \leqslant 1$, we have:*

$$IBA^- = \{G \to \overline{g} | (G, g) \in \mathcal{G}_{\gamma(G)} \times \mathcal{G}_{\gamma(g)}, \gamma(G) \subsetneq \gamma(g), \frac{n_G n_{\overline{g}}}{n}(1-\alpha) \geqslant N_{G\overline{g}}\} \quad (9)$$

For example from Table 1 if $minsup = 0.2$, we have $[AC] = \{A, AC\}$ and $[BE] = \{B, E, BE\}$. We see that $AC \subsetneq BE$ and mutually disfavors (i.e. AC favors $\overline{BE}$), so $A \to \overline{B}$ and $A \to \overline{E}$ are candidates. Here, $n = 6$, $n_A = 3$, $n_B = n_E = 5$, $n_{AB} = n_{AE} = 2$ and $\sqrt{\frac{n_A n_B}{n}} = 1.58$. Consider $\alpha = 1\%$, we have $\frac{n_A n_B}{n}(1-\alpha) = \frac{n_A n_E}{n}(1-\alpha) = 2.5 > 2 = n_{AB}$ and $P\left(N_{A \wedge B} \leqslant \frac{n_A n_B}{n}(1-\alpha)\right) = 0.49 \Rightarrow \{A \to \overline{B}, A \to \overline{E}\} \in IBA^-$, i.e. the association rules $A \to \overline{B}$ and $A \to \overline{E}$ are valid (99% likely) in IBA^- basis with probability 0.49.

Lemma 2. *$\forall X, Y, T, Z \subseteq \mathcal{I}$ such that X disfavors Y, Z disfavors T, $Z \subseteq \gamma(X)$, $T \subseteq \gamma(Y)$: (i) $supp(X \cup Y) = supp(Z \cup T)$; (ii) $M_{GK}(X \to \overline{Y}) = M_{GK}(Z \to \overline{T})$.*

Proof. (i) Because $Z \subseteq \gamma(X)$ and $T \subseteq \gamma(Y)$, we have $supp(X) = supp(Z)$ and $supp(Y) = supp(T)$. Thus, $supp(X \cup Y) = \frac{|\phi(X \cup Y)|}{|\mathcal{T}|} = \frac{|\phi(X) \cap \phi(Y)|}{|\mathcal{T}|} = \frac{|\phi(Z) \cap \phi(T)|}{|\mathcal{T}|} = \frac{|\phi(Z \cup T)|}{|\mathcal{T}|} = supp(Z \cup T)$. (ii) Since $supp(X) = supp(Z)$, $supp(Y) = supp(T)$ and $supp(X \cup Y) = supp(Z \cup T)$, we have $P(Y'|X') = P(T'|Z') \Leftrightarrow P(\overline{Y}|X) = P(\overline{T}|Z)$ equivalent to $P(\overline{Y}|X) - P(\overline{Y}) = P(\overline{T}|Z) - P(\overline{Y})$ equivalent to $\frac{P(\overline{Y}|X) - P(\overline{Y})}{1 - P(\overline{Y})} = \frac{P(\overline{T}|Z) - P(\overline{T})}{1 - P(\overline{T})}$, hence $M_{GK}(X \to \overline{Y}) = M_{GK}(Z \to \overline{T})$. □

Theorem 7. *(i) All valid negative approximate association rules, their supports and M_{GK}, can be derived from the rules of IBA^-. (ii) All association rules in the IBA^- basis are non-redudant negative approximate association rules.*

Proof. (i) Let $r_1 : X_1 \to \overline{Y_1} \backslash X_1$ be a valid negative approximate rule between two frequent itemsts with $X_1 \subset \overline{Y_1}$. Since $0 < M_{GK}(r_1) < 1$, we also have $\gamma(X_1) \subset \gamma(\overline{Y_1})$. For any frequent itemsets 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 $X_1 \subseteq \gamma(G_1) \subset \overline{Y_1} \subset \overline{G_2} \subseteq \gamma(\overline{Y_1}) = \gamma(\overline{G_2})$ and the rule $r_2 : G_1 \to \overline{G_2} \backslash G_1 \in IBA^-$. We show that the rule r_1 and its support can be derived from the rule r_2, its support and its M_{GK}. Since $G_1 \subset X_1 \subseteq \gamma(G_1)$ and $G_2 \subset Y_1 \subseteq \gamma(G_2)$, we have, according to Lemma 2, $supp(X_1 \cup Y_1) = supp(G_1 \cup G_2)$ implies $|\phi(X_1) \cap \phi(Y_1)| = |\phi(G_1) \cap \phi(G_2)|$ (a). Since $G_1 \subset X_1 \subseteq \gamma(G_1) \subset \overline{Y_1} \subset \overline{G_2} \subseteq \gamma(\overline{Y_1})$, we have $supp(G_1) = supp(X_1) = supp(\gamma(G_1)) = supp(\overline{Y_1}) = supp(\overline{G_2})$. Therefore, we have from (a), $|\phi(X_1) \cap \phi(\overline{Y_1})| = |\phi(G_1) \cap \phi(\overline{G_2})| \Leftrightarrow |\phi(X_1 \cup \overline{Y_1})| = |\phi(G_1 \cup \overline{G_2})|$ implies $supp(X_1 \cup \overline{Y_1}) = supp(G_1 \cup \overline{G_2})$, and thus $M_{GK}(X_1 \to \overline{Y_1}) = M_{GK}(G_1 \to \overline{G_2})$.

(*ii*) Let $r_2 : G \rightarrow \overline{g} \backslash G \in IBA^-$. According to definition 8, we have $G \in \mathcal{G}_C$ and $g \in \mathcal{G}_\mathcal{C}$ and $C \subsetneq \mathcal{C}$. We demonstrate that there is no other rule $r_3 : X_3 \rightarrow \overline{Y_3} \backslash X_3 \in IBA^-$ such that $supp(r_3) = supp(r_2)$, $M_{GK}(r_3) = M_{GK}(r_2)$, $X_3 \subset G$ and $Y_3 \subset g$ (i.e. $\overline{Y_3} \supset \overline{g}$). If $X_3 \subset G$ then, according to definition 2, we have $\gamma(X_3) \subset \gamma(G) = C$ and then $X_3 \notin \mathcal{G}_\mathcal{C}$. We deduce that $supp(X_3) > supp(G)$ and then $M_{GK}(r_3) < M_{GK}(r_2)$. If $\overline{g} \subset \overline{Y_3}$, we have $\gamma(\overline{g}) \subset \gamma(\overline{Y_3})$. We deduce that $supp(\overline{g}) > supp(\overline{Y_3})$ and conclude that $M_{GK}(r_2) > M_{GK}(r_3)$. □

4.3 NONRED algorithm

NonRed algorithm is composed of four algorithms (Algorithm 1, Algorithm 2, Algorithm 3 and Algorithm 4). The principal procedure (Algorithm. 2) takes as input the minimal generators $\mathcal{G}_\mathcal{C}$, the frequent closed $\mathcal{FC}$, the maximal frequent itemsets $\mathcal{MC}$ and, the minimum thresholds *minsup* and α. It returns all informative positive and negative association rules containing exact and approximate rules, denoted $\mathcal{IB}$. Its efficiency mainly stems from the economic calculation of the M_{GK} of all candidate rules as explained by the following theorem 8.

Theorem 8 (Pruning space). *Let $X_1 \rightarrow Y \backslash X_1$ and $X_2 \rightarrow Y \backslash X_2$ be two rules, we have: $X_2 \subseteq X_1 \subseteq Y \Rightarrow M_{GK}(X_2 \rightarrow Y \backslash X_2) \leqslant M_{GK}(X_1 \rightarrow Y \backslash X_1)$.*

Proof. By antimonotonicity of support, we have $X_2 \subseteq X_1 \Rightarrow supp(X_1) \leqslant supp(X_2)$ (i.e. $|\phi(X_1)| \leqslant |\phi(X_2)|$). Since $X_2 \subseteq X_1 \subseteq Y$, we have $P(Y'|X'_1) = \frac{|\phi(X_1 \cup Y)|}{|\phi(X_1)|} = \frac{|\phi(X_1) \cap \phi(Y)|}{|\phi(X_1)|} = \frac{|\phi(Y)|}{|\phi(X_1)|} \geqslant \frac{|\phi(Y)|}{|\phi(X_2)|} = P(Y'|X_2)$. From $P(Y'|X'_2) \leqslant P(Y'|X'_1)$, we have $P(Y'|X'_2) - P(Y') \leqslant P(Y'|X'_1) - P(Y') \Leftrightarrow \frac{P(Y'|X'_2) - P(Y')}{1 - P(Y')} \leqslant \frac{P(Y'|X'_1) - P(Y')}{1 - P(Y')}$ equivalent to $M_{GK}(X_2 \rightarrow Y \backslash X_2) \leqslant M_{GK}(X_1 \rightarrow Y \backslash X_1)$. □

Theorem 8 infers that, for all $\widetilde{X} \subseteq X$, if $X \rightarrow Y \backslash X$ does not valid, neitheir does $\widetilde{X} \rightarrow Y \backslash \widetilde{X}$. And, if $X \rightarrow \overline{Y} \backslash X$ is not valid, then $\widetilde{X} \rightarrow \overline{Y} \backslash \widetilde{X}$ will not be. For example, if $A \rightarrow BCD$ is valid, then $AB \rightarrow CD$ and $ABC \rightarrow D$ are valid. If $A \rightarrow \overline{BCD}$ is not valid, then $AB \rightarrow \overline{CD}$ and $ABC \rightarrow \overline{D}$ will not be valid.

These optimizations are summarized in algorithm 2. The latter consists of two main parts. The first part (lines 2-16) generates the IBE^+ basis the IBA^+ basis. The second part (lines 17-32) corresponds to the generation of IBE^- basis and IBA^- basis. For all closed C (line 2) and generator G of this closed C (line 3) which respectively represent the premise and the consequence of the rule, we check if these two itemsets are positively dependent (line 4). We also check if they belong to the same equivalence class (line 5). From line 6, we check that the generator G is not a single element in its equivalence class (i.e. $\gamma(G) \neq G$), and also check if an itemset $G \cup C$ of a candidate rule $G \rightarrow C$ is frequent. All elements of Base (line 7) must satisfy these conditions, so this candidate is eligible for giving the exact positive rule. If these two itemsets are not in the same equivalence class (line 9), we generate the approximate positive rules (lines 10-14). We collect another closed $\mathcal{C}$ containing the closure of G (i.e. $\gamma(G)$) (line 10), and then check if the candidate satisfies the constraints *minsup*

Algorithm 2 NonRedBase

Require: $\mathcal{G}_C$, $\mathcal{FC}$, $\mathcal{MC}$, minimum threshold $minsup$ and α.
Ensure: Base, a set of Informative basis of association rules.

1: Base = $\emptyset$;
2: **for all** ($C \in \mathcal{FC}$) **do**
3: **for all** ($G \in \mathcal{G}_C$) **do**
4: **if** ($P(C'|G') > P(C')$) **then**
5: **if** ($\gamma(G) = C$) **then**
6: **if** ($G \neq \gamma(G)$ && $supp(G \cup C) \geq minsup$) **then**
7: Base $\leftarrow$ Base $\cup \{G \rightarrow C \backslash G\}$; /* Positive Exact Rules-IBE^+ */
8: **end if**
9: **else**
10: **for all** ($\mathcal{C} \in \mathcal{FC} \mid \mathcal{C} \supset \gamma(G)$) **do**
11: **if** ($supp(G \cup \mathcal{C}) \geq minsup$ && $\frac{n_G n_{\overline{\mathcal{C}}}}{n}(1-\alpha) \geq N_{G\overline{\mathcal{C}}}$) **then**
12: Base $\leftarrow$ Base $\cup \{G \rightarrow \mathcal{C} \backslash G\}$; /* Positive Approximate Rules-IBA^+ */
13: **end if**
14: **end for**
15: **end if**
16: **else**
17: **for all** ($\mathcal{M} \in \mathcal{MC}$) **do**
18: $\mathcal{G}_{\mathcal{M}} = generator(\mathcal{M})$;
19: **for all** ($G \in \mathcal{G}_{\mathcal{M}}$) **do**
20: **for all** ($y \in \mathcal{I} \backslash \mathcal{M}$) **do**
21: **if** ($supp(G \cup \{\overline{y}\}) \geq minsup$) **then**
22: Base $\leftarrow$ Base $\cup \{G \rightarrow \{\overline{y}\} \backslash G\}$; /* Negative Exact Rules-IBE^- */
23: **end if**
24: **end for**
25: **end for**
26: **end for**
27: **for all** ($g \in \mathcal{G}_{\gamma(g)} | \gamma(G) \subsetneq \gamma(g)$) **do**
28: **if** ($supp(G \cup \overline{g}) \geq minsup$ && $\frac{n_G n_{\overline{g}}}{n}(1-\alpha) \geq N_{G\overline{g}}$) **then**
29: Base $\leftarrow$ Base $\cup \{G \rightarrow \overline{g} \backslash G\}$; /* Negative Approximate Rules-IBA^- */
30: **end if**
31: **end for**
32: **end if**
33: **end for**
34: **end for**
35: PruneRedRules(Base)
36: **return** Base

and α (line 11). If so, the candidate is eligible and added in Base (line 12). This completes the first part. The second part also consists of two sub-parts. The first sub-part (lines 17-26) corresponds to the procedure for mining the exact negative association rules. The second sub-part corresponds to the procedure for mining approximate negative rules (lines 27-31). So, for a maximal $\mathcal{M}$ (line 17), we generate its generator $\mathcal{G}_{\mathcal{M}}$ (line 18). For a dual y (dual of $\mathcal{M}$) (line 20), we check if the support of the rule $G \rightarrow \overline{y}$ is frequent (line 21). If this is the case, the Base will be updated (line 22). Finally, for a generator g which does not belong to the same equivalence class of G (line 27), we check if the support and M_{GK} of a candidate rule $G \rightarrow \overline{g}$ are frequent (line 28). If so, this rough rule is valid and added in Base (line 29). The procedure PruneRedRules (cf. Algorithm 3) removes the redundancies. So, after initialization (line 1), consider two association rules $X \rightarrow Y \backslash X$ et $\widetilde{X} \rightarrow Y \backslash \widetilde{X}$ (lines 2-3). We check if $\widetilde{X} \subseteq X$, then $X \rightarrow Y \backslash X$ is redundant with respect to $\widetilde{X} \rightarrow Y \backslash \widetilde{X}$, so pruned (line 5). Next, consider two negative rules $X \rightarrow \overline{Y} \backslash X$ et $\widetilde{X} \rightarrow \overline{Y} \backslash \widetilde{X}$ (lines 9-10). We then

Algorithm 3 PruneRedRules

Require: Base Base of informative rules
1: Base=Base;
2: **for all** $(X \rightarrow Y \backslash X \in$ Base$)$ **do**
3: **for all** $(\widetilde{X} \rightarrow Y \backslash \widetilde{X} \in$ Base$)$ **do**
4: **if** $(\widetilde{X} \subset X)$ **then**
5: Base $\leftarrow$ Base$\backslash \{X \rightarrow Y \backslash X\}$
6: **end if**
7: **end for**
8: **end for**
9: **for all** $(X \rightarrow \overline{Y} \backslash X \in$ Base$)$ **do**
10: **for all** $(\widetilde{X} \rightarrow \overline{Y} \backslash \widetilde{X} \in$ Base$)$ **do**
11: **if** $(\widetilde{X} \subset X)$ **then**
12: Base $\leftarrow$ Base$\backslash \{X \rightarrow \overline{Y} \backslash X\}$
13: **end if**
14: **end for**
15: **end for**
16: **return** Base

check if $\widetilde{X} \subseteq X$, then $X \rightarrow Y \backslash X$ is redundant, therefore pruned in Base (line 12). AllValidRules algorithm (Algorithm 4) derives *all valid informative rules*. It first derives the rules of type $X \rightarrow Y \backslash X$. Indeed, if $X \rightarrow Y \backslash X$ is

Algorithm 4 AllValidRules

Require: Base Base of informative rules, and $minsup$.
Ensure: All_{RUL} (*All Valid Association Rules*).
1: All_{RUL} = Base;
2: **for all** $(X \rightarrow Y \backslash X \in$ Base$)$ **do**
3: **if** $(supp(\overline{X} \cup \overline{Y}) \geqslant minsup)$ **then**
4: $All_{RUL} \leftarrow All_{RUL} \cup \{\overline{X} \rightarrow \overline{Y} \backslash \overline{X}\}$
5: **end if**
6: **end for**
7: **for all** $(X \rightarrow \overline{Y} \backslash Y \in$ Base$)$ **do**
8: **if** $(supp(\overline{X} \cup Y) \geqslant minsup)$ **then**
9: $All_{RUL} \leftarrow All_{RUL} \cup \{\overline{X} \rightarrow Y \backslash \overline{X}\}$
10: **end if**
11: **end for**
12: **return** All_{RUL}

valid, then $\overline{X} \rightarrow \overline{Y} \backslash \overline{X}$ is also valid (concept of derivability [4]) and added in All_{RUL} (Algorithm 4 line 4). The last step address for deriving the rules of type $X \rightarrow \overline{Y} \backslash X$ (Algo. 4 lines 7-11). For this, if $X \rightarrow \overline{Y} \backslash X$ is valid, then $\overline{X} \rightarrow Y \backslash \overline{X}$ is also valid (derivability [4]), and added in All_{RUL} (Algorithm 4 line 9).

We present the complexity of the principal algorithm NonRedBase. Its complexity remains linear in $|\mathcal{FC}| \times |\mathcal{G}_{\mathcal{C}}|$, and is in $\mathcal{O}(|\mathcal{FC}||\mathcal{G}_{\mathcal{C}}|(5^m - 2(3^m)))$. Indeed, the line 2 (resp. line 3) is in $\mathcal{O}(|\mathcal{FC}|)$ (resp. $\mathcal{O}(|\mathcal{G}_{\mathcal{C}}|)$). It has two recursives tests. The first test (lines 4-15) is in $C_1 = \mathcal{O}(|\mathcal{FC}||\mathcal{G}_{\mathcal{C}}|)$, and the second test (lines 17-31) in $C_2 = \mathcal{O}(|\mathcal{MC}|+|\mathcal{G}_{\mathcal{C}}|)$. C_1 is more complex than C_2 (i.e. $C_2 \leqslant C_1$). For a m-itemset, the cardinality of association rules is equal to $2^{2m} - 2^{m+1}$. Which gives $C_m^{m-1}(2^{2(m-1)} - 2^m)$ for a $(m-1)$-itemset, $C_m^{m-2}(2^{2(m-2)} - 2^{(m-1)})$ for a $(m-2)$-itemset, and so an. In sum, $\sum_{k=2}^{m} C_m^k(2^{2k} - 2^{k+1}) = \sum_{k=2}^{m} C_m^k 4^k -$

$2\sum_{k=2}^{m} C_m^k 2^k = \left[\sum_{k=0}^{m} C_m^k 4^k - (1+4m)\right] - 2\left[\sum_{k=0}^{m} C_m^k 2^k - (1+2m)\right]$. For all $x \in \mathbb{R}$, $\sum_{k=0}^{m} C_m^k x^k = (1+x)^m$, then $\sum_{k=2}^{m} C_m^k (2^{2k} - 2^{k+1}) = \mathcal{O}(5^m - 2(3^m))$. Finally, the time complexity of NonRedBase is in $\mathcal{O}(|\mathcal{FC}||\mathcal{G}_{\mathcal{C}}|(5^m - 2(3^m)))$.

5 Experimental evaluation

In this section, we present the performance of NonRed algorithm with respect to Bastide [2] and Feno [6] approaches. Our approach is implemented in a PC with an Intel Core i3 processor running at 4GB, under Windows (64bit), conducted on four reference databases (or datasets) (cf. Table 2): T10I4D100K[2] and T20I6D100K (cf. footnote 4), C20D10K[3] and Mushrooms (cf. footnote 5). The

Table 2: Database characteristics

Database	Number of transaction	Number of items	Average size of transaction
T10I4D100K	100 000	1 000	10
T20I6D100K	100 000	1 000	20
C20D10K	10 000	386	20
Mushrooms	8 416	128	23

table 2 represents the variation of the number of association rules obtained by our algorithm with respect to those of Bastide and Feno approaches on different datasets for various *minsup* at fixed $\alpha = 0.05$ (for NonRed and Feno approach) and $minconf = 0.8$ (Bastide approach). We also denote by "-" a subset which could not generated. Recall that Bastide contains no negative rules. So, the

Table 3: Number of all valide informative association rules

Dataset	*minsup*	Bastide et al.				Feno et al.				NonRed			
		$\|E^+\|$	$\|A^+\|$	$\|E^-\|$	$\|A^-\|$	$\|E^+\|$	$\|E^-\|$	$\|A^+\|$	$\|A^-\|$	$\|E^+\|$	$\|E^-\|$	$\|A^+\|$	$\|A^-\|$
T10I4D100K	10%	0	11625	-	-	0	0	20555	1256	0	0	725	52
	20%	0	8545	-	-	0	0	15656	1058	0	0	545	34
	30%	0	3555	-	-	0	0	12785	954	0	0	355	25
T20I6D100K	10%	115	71324	-	-	95	98	71899	3897	115	103	1804	56
	20%	76	57336	-	-	66	91	45560	2705	76	95	1403	38
	30%	58	45684	-	-	43	63	41784	1887	58	63	1175	27
C20D10K	10%	1125	33950	-	-	975	255	34588	11705	1125	285	1856	182
	20%	997	23821	-	-	657	135	25582	8789	997	185	1453	123
	30%	967	18899	-	-	567	98	19581	4800	967	101	1221	97
Mushrooms	10%	958	4465	-	-	758	289	4150	3887	958	304	1540	89
	20%	663	3354	-	-	554	178	2944	2845	663	198	1100	78
	30%	543	2961	-	-	444	109	2140	1987	543	115	998	39

exact negative rules E^- (resp. approximate rules A^-) are absent, which gives a notable loss of information for Bastide. For each algorithm, no exact positive E^+ (resp. approximate A^-) association rule is generated from sparse database

[2] http://www.almaden.ibm.com/cs/quest/syndata.html
[3] http://kdd.ics.uci.edu/

T10I4D100K since, for $minsup \leq 30\%$. The reason is that all frequent itemsets are frequent closed itemsets. On the other databases, we can observe that the total number of exact positive and exact negative rules is very reasonable, for all *minsup*. For this, Feno represents number smaller than Bastide and NonRed. The explanation is that Feno uses the pseudo-closed which returns a reduced number of itemsets and thus, it is the same for the association rules, but it's not informative. Whereas Bastide and NonRed algorithms generate the informative association rules. In this case, the premise is selected from to the set of generator itemset which is more dense with respect to the set of pseudo-closed.

On the dense and strongly correlated datasets (C20D10K and Mushrooms), it is very visible that our algorithm, NonRed, is much more selective than Bastide and Feno, for all *minsup*. For example, with the dense database C20D10K at good *minsup* (1%), NonRed only extracts $|A^+| = 1856$ against $|A^+| = 33950$ and $|A^+| = 34588$ for Bastide and Feno respectively (cf. Table 3), let a difference of more than 32000 approximate rules. The main reason is associated with a pruning strategy. Using the pruning strategy (cf. Property 6), NonRed can prune the uninteresting rules, this is not the case for Bastide. In addition, NonRed includes an efficient technique for pruning weakly correlated rules (i.e. close to independence), based on the constraint of M_{GK} minimal, this is not the case for Feno approach. The latter uses a strategy based on critical value, while this critical value is not very selective, it accepts often weakly correlated rules.

We present in the following the execution times of our algorithm, compared to those of literature (Bastide and Feno approaches, in particular). However, this comparison is still very difficult, for several reasons. First, Feno approach, only approach (to our knowledge) for informative basis of positive and negative association rules, does not take into account the frequent itemsets mining, while this step considerably affects the execution times. For this, the frequent itemsets are extracted in other algorithms, which are not taken into account in its cursus. Thus, Feno approah is not comparable of our approach. On the other hand, Bastide approach could not integrate the negative exact and approximate association rules (E^- and A^-), while these subsets could to give very high complexity. For this situation, we partially compare NonRed and Bastide. More precisely, this comparative study only concerns execution times for positive exact E^+ and approximate A^+ association rules. The results of this will be represented in Fig. 2 by varying the *minsup* at fixed $\alpha = 0.05$ and $minconf = 0.6$. On sparse datasets (T10I4D100K and T20I6D100K), the execution times of NonRed and Bastide algorithms are almost identical for positive exact association rules E^+, at the lowest *minsup* (cf. Fig. 2a and 2b). On approximate association rules A^+, it is very obvious that our algorithm, NonRed, is better than Bastide approach (cf. Fig. 2a, 2b). The explanation is that all frequent itemsets are frequent closed itemsets, which complicates the task of Bastide who performs more operations than NonRed algorithm to determine the closures and approximate rules. Even if NonRed algorithm is linear on the number of frequent closed, it benefits a significant reduction for number of accesses to the datasets to determine the supports and frequent closed itemsets, its execution time is then low.

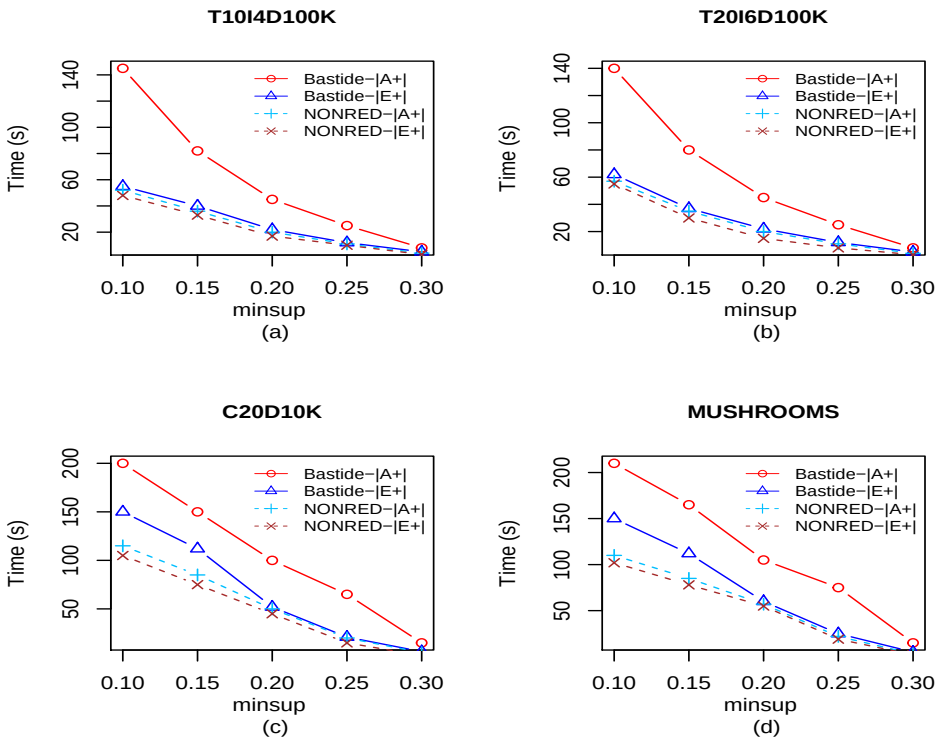

Fig. 2: Response times by varying *minsup* at fixed $\alpha = 0.05$ and $minconf = 0.6$

On dense and strongly correlated datasets (C20D10K and MUSHROOMS), NONRED algorithm is once again better than Bastide, for approximate association rules (cf. Fig. 2c and 2d). The main reason is also associated with the pruning strategy. Using the pruning strategies (cf. Property 6, Theorems 4, 5, 6 and 7), NONRED can reduce considerably the search space, this is not the case for Bastide, and NONRED ends quickly (cf. Fig. 2c and 2d). Bastide obtains the less performance: there is also no pruning strategy for uninteresting association rules, while variations of these rules affect execution times, considerably. As a result, the search space can be browsed in its entirety, which considerably penalizes the execution times for Bastide. This is no longer the case for exact association rules E^+ where Bastide joins NONRED, for *minsup* of 20% to 30%.

6 Conclusion

We presented and evaluated our approach for mining informative association rules. This approach simultaneously generates both the positive and negative

association rules using the frequent closed itemsets and maximal itemsets, and their generators. Experiments have shown that this approach is more efficient than the existing approaches of the literature. The prospective for future work relate to the interactive visualization of the association rules extracted. Another perspective would be to extend this work in paradigm of multidimensional rules.

References

1. Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In Proceedings of 20th VLDB Conference, Santiago Chile, 487–499 (1994).
2. Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., Lakhal, L.: Mining Minimal Non-Redundant Association Rules using Frequent Closed Itemsets. In CL'2000 international conference Computational Logic, pp. 972–986 (2000).
3. Bemarisika, P.: Extraction de règles d'association selon le couple support-M_{GK}: Graphes implicatifs et Applications en didactique des mathématiques. PhD thesis, Université d'Antananarivo (Madagasar), (2016).
4. Bemarisika, P., Ramanantsoa, H., Totohasina, A.: An Efficient Approach for Extraction Positive and Negative Association Rules from Big Data. Proceedings of International Cross Domain Conference for Machine Learning & Knowledge Extraction (CD-MAKE 2018), pp. 79–97 (2018).
5. Durand, N., Quafafou, M.: Approximation of Frequent Itemset Border by Computing Approximate Minimal Hypergraph Transversals. DaWaK, 357–368 (2014).
6. Feno, D.R., Diatta, J., Totohasina, A.: Galois Lattices and Based for M_{GK}-valid Association Rules. Proceedings of CLA 2006, 127–138 (2006).
7. Ganter, B., Wille, R.: FCA: Mathematical foundations. Springer-Verlag (1999).
8. Giacomo, K., Alexandre B.: Average Size of Implicational Bases. Dmitry I. Ignatov, Lhouari Nourine (Eds.), CLA 2018, pp. 37–45 (2018).
9. Guigues, J.L., Duquenne, V.: Familles minimales d'implications informatives résultant d'un tableau de donnés binaires. Maths et Sci. Humaines, 5–18 (1986).
10. Guillaume, S.: Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales. PhD thesis, Universté de Nantes (France), 2000.
11. Kryszkiewicz, M.: Concise representations of association rules. In Hand, D.J., Adams, N.M., Boltonet, R.J. (Eds), 92–103 (2002).
12. Latiri, C., Haddad, H., Hamrouni, T.: Towards an effective automatic query expansion process using an association rule mining approach. IIS, 209–247 (2012).
13. Lerman, I. C.: Classification et analyse ordinale des données. Dunod, (1981).
14. Mannila, H., Toivonen, H.: Levelwise Search and Borders of Theories in Knowledge Discovery. In Proc. on Data Mining Knowledge Discovery, 241–258 (1997).
15. Pasquier, N.: Extraction de Bases pour les Règles d'Association à partir des Itemsets Fermés Fréquents. CNRS (2000).
16. Totohasina, A., Ralambondrainy H.: ION, A pertinent new measure for mining information from many types of data. In IEEE, SITIS, 202–207 (2005).
17. Totohasina, A., Feno, D. R.: De la qualité de règles d'association: Etude comparative des mesures M_{GK} et Confiance. In CARI (2008).
18. Wu, X., Zhang, C., S. Zhang, S.: Efficient mining of both positive and negative association rules. In ACM Transactions on information Systems 3, 381–405 (2004).
19. Zaki, M.J.: Mining Non-Redundant Association Rules. Proc. of KDDM (2004).