OPTI2i AND ε-PRECIS METHODS SELECTION ALGORITHM

Nombre Claude Issa[1] , Brou Konan Marcellin[2] , Kimou Kouadio Prosper[3]
[1]Polytechnic Doctoral School
[2]National Polytechnic Institute Houphouet Boigny – Yamoussoukro
[3]Research Laboratory of Computer Science and Technology

ABSTRACT

Since the reference algorithm APRIORI [AGR97], other algorithms for optimizing the extraction of association rules have been developed. But no method is generally better than the others. This article deals with the optimization of closed itemsets in the context of highly correlated data. The work in this article responds to one of the perspectives of our article entitled "A new approach to optimizing the extraction of frequent 2-itemsets". In this previous article, we had obtained interesting optimization results from the 2-itemsets on a context of extraction of scattered data (weakly correlated data). The present article allowed us to obtain interesting results of the 2-itemsets on dense data (strongly correlated).
Our approach was inspired by the research work of {Pas00, CB02, BBR03, CF14]. It has improved the extraction of a concise number of association rules by introducing a margin of error defined by the parameter ε in the formula $\left|\text{Ferm}_{\delta}(S)-\delta\right|<\varepsilon$ (δ an integer, $\delta>0$, $\text{Ferm}_{\delta}(S)$ is the δ-Closure of the item S [CF07]), $\varepsilon = \frac{1}{n_r + 1}$ where n_r is the number of association rules extracted with the δ-strong rule [CF07]). The smaller ε, the more precise and concise the number of association rules extracted. The comparison between the δ-strong rule, based on δ-Closure and δ-free ([Pas00]) and the ε-precise rule that we found showed that our approach is much more precise and eliminates production errors of the really concise number of association rules. From the result obtained on dense data, we wrote an algorithm called SELECTION. The SELECTION algorithm allows you to select the appropriate method for optimizing 2-itemsets, depending on the context of data extraction.

Keywords: Precise and concise number of association rules, dense data, margin of error, itemset

1. INTRODUCTION

We recall that in our article, entitled OPTI2i, we managed to optimize the extraction of frequent 2-itemsets only for contexts of extraction of scattered data. However for dense data, our results have not often been better than those of the APRIORI reference algorithm. We used an already existing method to provide a solution to the optimization of frequent 2-itemsets in the context of dense data. There are several algorithms for extracting knowledge from the data contained in a database. For more than a decade, the method based on closed itemsets has shown its effectiveness over the years, especially on highly correlated data.
We are interested in discovering a concise and even more precise number of rules of association. The discovery of a restricted number of association rules [CF07] helped solve two problems :

- The extraction of redundant association rules generated by the reference algorithm APRIORI [AGR94] and useless for the expert in the field;
- The high time taken to extract association rules and reduce their storage space.

Our approach aims to contribute to the precision of extraction of these association rules by introducing the parameter ε.
The main objective of this research work is for us to manage to select the optimized extraction method of the frequent 2-itemsets from a context of data extraction encountered. To this end, we propose in the first section the study of art on closed itemsets. Section 2 presents the methodological approach

DOI:10.5121/ijcsa.2020.10401

envisaged to facilitate the discovery of knowledge based on association rules. Section 3 discusses the discovery of a precise and more concise number of association rules while optimizing the extraction of frequent closed 2-itemsets. Section 4 proposes the SELECTION algorithm, which allows us to select the appropriate method between OPTI2i and ε-precise according to the context of data extraction. Our article ends with a brief conclusion and the perspectives envisaged.

2. SECTION 1

1.1. STATE OF THE ART

In this part, we will present the algorithms operating only on the extraction of closed itemsets. The best known algorithms in this context are the CLOSE, A-CLOSE and TITANIC algorithms. Here we will not care about the number or the quality of the association rules generated.
Each of these algorithms is characterized by two phases. During the first phase, the infrequent candidates are pruned, that is to say the itemsets having a support strictly inferior to Sup_min (minimal support). Each algorithm introduces new pruning strategies to try to reduce the additional cost of calculating item closings.
The second phase is that of construction relating to an iteration k. It makes it possible to generate the candidates of size k from those retained during the iteration k-1, that is to say of size k-1. The work of CLEMENT FAURE [CF07] based on the δ-strong rule has given effective results in optimizing frequent closed itemsets. His work has also made it possible to extract a concise number of association rules qualitatively meeting the needs of the expert in the field.

1.2. CONTRIBUTION

The δ-strong rule has made it possible, like other previous algorithms based on frequent closed itemsets, to considerably reduce the number of association rules. This rule produced a concise number (i.e. a limited number) of association rules.
While this approach is good, it does not achieve a 100 percent efficiency rate. Thus it can ignore association rules which could be potentially useful to the expert depending on the field studied and the request entered by the user. It is for this reason that we introduce a parameter ε> 0, which will take into account this margin of error (see definition 10) without additional cost and with better extraction time of the frequent closed 2-itemsets.
This contribution is also inspired by the research work of [Pas00, CB02, BBR03]. The former studied the properties of condensed representations using frequent δ-free itemsets and the closure operator (δ-closed). We also saw that it was possible to generate a concise set of association rules from a representation using the free. Another line of research considered the use of so-called δ-strong rules (minimal body, controlled confidence) in the context of classification.
Clement Faure used an algorithm [BBR00] capable of extracting a condensed representation using frequent δ-free itemsets and δ-closure. According to FAURE's work, if the behavior is known when is equal to zero, it is interesting to study the properties of these itemsets and of the rules generated when δ is strictly greater than zero. In our work we have considered a generalization of these different approaches. In fact, we use the results of Clément Fauré's work to minimize the margin of error for generating a concise number of association rules, while retaining the main advantage of algorithms based on so-called closed itemsets. This achievement is based, as we have mentioned in the state of the art, on the pruning of redundant itemsets. Since the OPTI2i method could not give us better results compared to the previous methods only in the contexts of extraction of dense data, we will also treat in this article the selection of the OPTI2i or ε-precise optimization method depending on the context of data extraction used.

3. SECTION 2

2.1. WORKING ENVIRONMENT

In this part, we will present our working framework, that is to say the context in which we will situate our contribution. Let K be the database from section 2 of the state of the art, Ti the transactions of K (Ti⊆ K). At an itemset of cardinal 1, Xi the items (attributes) are different from A. (A c Xi) ⊂ K. The item A if it exists is always located on the left side of the association rule. After setting the scene, we will set out definitions. Definitions 8 to 10 are inspired by the work of Clement Faure et al. [CF07].

2.2. WORKING TOOLS

Our working tools first go through a literature review which allowed us to bring out a state of the art on the discovery of frequent closed itemsets. We also used propositions based on mathematical notions in order to consolidate our work, artificial intelligence and datamining techniques.
The data mining technique used here is the popular method of association rules. The implementation phase of our work is based on the optimization algorithm for extracting the 2-itemsets from the APRIORI reference algorithm and the work of CLEMENT FAURE. As its name suggests, the SELECTION function will help us to select between the optimization method of frequent 2-itemsets OPTI2i and the optimization method of frequent closed 2-itemsets which we will call P-FERM.

2.3. THEORETICAL PHASE

Consider the data from 6 receipts that we obtained with 6 customers from the CDCI YAMOUSSOUKRO supermarket (Ivory Coast): Let BD be the binary database of transactions T and K, all of the items. K = {A, B, C, D, E, F, G, H, I, J}

$t_1 \rightarrow$ {Milk, cookie, soap, rice, sugar }
$t_2 \rightarrow$ { Beer can, oil, blue II razor}
$t_3 \rightarrow$ { Soap, oil, beer can, sugar}
$t_4 \rightarrow$ { Spaghetti, oil, blue II razor}
$t_5 \rightarrow$ { Milk, cookie, sugar, nescafe box}
$t_6 \rightarrow$ { soap, rice, sugar, oil}

A→Milk B→Cookie C→Soap D→Rice E→Sugar F→Bier can
G→ Oil H→Blue II razor
I→Spaghetti J→Nescafe box.

The following table represents the transaction database, where each transaction is a list of items purchased by one of the 6 customers of the supermarket :

Ti	A	B	C	D	E	F	G	H	I	J
t_1	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**
t_2	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**	**0**
t_3	**0**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**0**	**0**
t_4	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**0**
t_5	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**
t_6	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**

Figure 1: list of products

Definition 1: Item

An item corresponds to a product on one of the 6 receipts. In our database BD we have 10 items (A, B, C, D, E, F, G, H, I and J).

Definition 2: Itemset
An itemset is a set of items.
Example: {A} is an itemset of cardinal 1; {A, B} is an itemset of cardinal 2; {A, B, C} is an itemset of cardinal 3;

Definition 3: Support
We call support for an itemset the number of times its items appear together in database transactions. Example: Support ({A}) = 2/6; Support ({A, B}) = 2/6; Support ({A, B, C}) = 1/6

Definition 4: Superset
A superset is a cardinal itemset greater than its sub itemset. Example: {A, B} is a superset of {A}; {A, B} is a superset of {B}.

Definition 5: Frequent itemset
An itemset is said to be frequent if its support is greater than or equal to a minimum support (Sup_min) set by the user. Example: If we consider Sup_min = 2/6, then if we consider only the itemsets of cardinal 1, we can retain the itemsets {A}, {B}, {C}, {D}, {E}, {F}, {G} and {H}. The itemsets {I} and {J} are not frequent.

Definition 6: Closed itemset
A frequent itemset is said to be closed if its support is larger than those of all its supersets.
Example: If we consider Sup_min = 2/6, then the item {A, B} is often closed because none of its supersets has support greater than 2/6. Support ({A, B, C}) = 1/6; Support ({A, B, C, D}) = 2/6; Support ({A, B, C, D, E}) = 1/6; Support ({A, B, C, D, E, F}) = 0; Support ({A, B, C, D, E, F, G}) = 0; Support ({A, B, C, D, E, F, G, H}) = 0; Support ({A, B, C, D, E, F, G, H, I}) = 0; Support ({A, B, C, D, E, F, G, H, I, J}) = 1/6.

Definition 7: Minimum generator
It is the smallest element in an equivalence class.
The association rules extraction approach using closed frequent itemsets was preceded by the association rules extraction approach based on the extraction of frequent itemsets [AGR94]. The algorithms adopting the approach of extracting closed itemsets have made it possible to solve the problems of generation of the many redundant itemsets and often very difficult to interpret by the end user to whom the data belongs. This approach is based on pruning the lattice of closed itemsets, using the closing operators of the Galois connection. A first category of algorithms developed were interested in the discovery of frequent closed itemsets. A second category of algorithms, also based on the extraction of frequent closed itemsetsets, has improved the results obtained by the algorithms of the first category, by considerably and precisely limiting the number of association rules generated.
However if the number of association rules is concise, the fact remains that this number obtained is not very precise. This would result in the loss of certain association rules (knowledge) which could prove potentially useful to the expert
Now let's review the main algorithms allowing the extraction of frequent closed itemsets.

Definition 8 (Rule δ-strong)
Given a database K defined on Items, a minimum frequency threshold γ, and an integer $\delta > 0$, a rule δ-strong on K is an association rule $A \rightarrow X$, and Freq $(A \cup X) \geq \gamma$, Freq (A) - Freq $(A \cup X) \leq \delta$, $A \cup X \subseteq$ Items; with $X \neq \emptyset$.
Note : A δ-strong rule accepts at most, δ items on its right side.
Example of a "1-strong" rule : $AB \rightarrow C$ is a 1-strong, on the other hand $BE \rightarrow A$ is not a 1-strong rule.

Definition 9 (δ-closure)

Let A be an itemset ⊆ Items and δ a positive integer. The fermeture-closure of A, is the largest on set of A defined as follows : $f\text{erm}_{\delta}(A) = \{I \in \text{Items} | \text{Freq}(S) - \text{Freq}(A \cup \{X_i\}) \leq \delta\}$

Note : The BD database in which we have drawn our examples does not allow to obtain enough δ-forte because it does not contain dense data.

Definition 10 (ε-precise)

Let A be an itemset ⊆ Items and δ a positive integer. The δ-closure of A, is the largest on set of A defined as follows : $f\text{erm}_{\delta}(A) = \{I \in \text{Items} | \text{Freq}(S) - \text{Freq}(A \cup \{X_i\}) \leq \delta\}$

And respects the following condition : $|\text{Ferm}_{\delta}(A) - \delta| < \varepsilon$. The smaller ε, the higher and more precise the certainty of extracting interesting itemsets.

Definition 11 (normal law)

In terms of statistical decision and forecasting, the normal law is the most widespread and useful statistical law. It represents many random phenomena. In addition, many other statistical laws can be approached by the normal law, especially in the case of large samples. Its mathematical expression is as follows :

$$n(x) = \frac{n}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Let n_{itfreq} - be the number of itemsets generated with δ-closure. % case = n_{itfreq} / (n_{itfreq} +1). The normal law can be applied to the extraction of δ-closed itemsets. According to the definition, each δ-closed k-itemsets produces a number nk (k≥2). nk are normally distributed around the exact value n_{itfreq}.

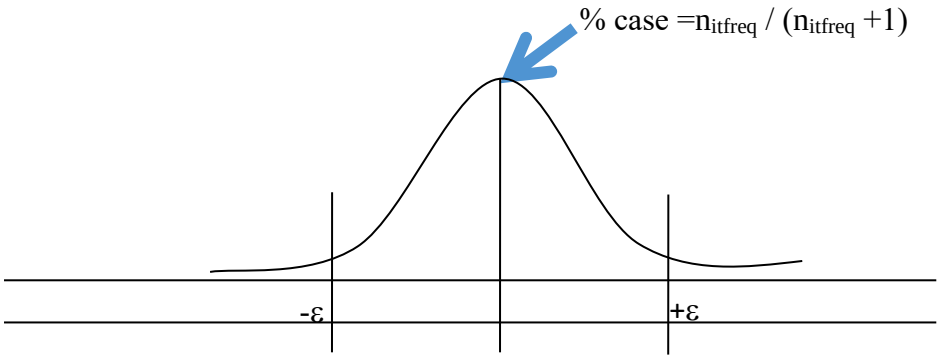

Figure 2 - Normal curve for estimating the margin of error ε

Conditions of use of the normal distribution around the true value

Let nitfreq be the probability sample such as :

- $n_{itfreq} \geq 30$
- $n_{itfreq} * \hat{P}_{\%} \geq 500$
- $n_{itfreq} * (100 - \hat{P}_{\%}) \geq 500$

$$\varepsilon = \left(1 - \frac{n_{itfreq}}{n_{itfreq} + 1}\right) = \frac{1}{(n_{itfreq} + 1)}$$

Replace ε in the margin of error formula. We have :

$$[P_{\varepsilon-}; P_{\varepsilon+}] = \hat{P}_{\%} \pm \frac{1}{(n_{itfreq} + 1)} * \sigma_{\hat{P}_{\%}}$$

Where $\sigma_{\hat{P}_{\%}}$ is called standard error

$$\begin{cases} \sigma_{\hat{P}_{\%}} = \dfrac{\sqrt{P_{\%}(100 - P_{\%})}}{\sqrt{n_{tfreq}}} & \text{So large population of itemsets } (N >= (n_{itfreq}+1)n_{itfreq}) \\ \sigma_{\hat{P}_{\%}} = \dfrac{\sqrt{P_{\%}(100 - P_{\%})}}{\sqrt{n_{tfreq}}} * \sqrt{\dfrac{(N - n_{tfreq})}{(N-1)}} & \text{So small population of itemsets } (N < (n_{itfreq}+1)n_{itfreq}) \end{cases}$$

N - Total number of itemsets; nitfreq - The sample size which represents the number of frequent itemsets obtained with the δ-closure.

The real percentage $P_{\%}$ will take the value of the percentage of the highest confidence ($Conf_{max}$) of the nitfreq.

$(Freq(A) - Freq(A \cup Xi) < \varepsilon + \delta$ (1)

$(Freq(A) - Freq(A \cup Xi) < \dfrac{1}{(n_{itfreq}+1)} + \delta$ (2)

– If n_{itfreq} is high, then $\lim_{n_{itfreq} \to +\infty} \dfrac{1}{(n_{itfreq}+1)} \to 0$,

(2) becomes $(Freq(A) - Freq(A \cup Xi) < \delta \Rightarrow \delta$-closure (3)

– If n_{itfreq} is very small, then $\lim_{n_{itfreq} \to 0} \dfrac{1}{(n_{itfreq}+1)} \to 1$,

(2) becomes $(Freq(A) - Freq(A \cup Xi) < \delta + 1 \Rightarrow \varepsilon$-precise (4)

n_{itfreq}	10	100	1000	10000	100000	1000000
ε	0,09090909	0,00990099	0,000999	9,999E-05	9,9999E-06	1E-06

Table 1 - Estimated value of epsilon as a function of n_{itfreq}

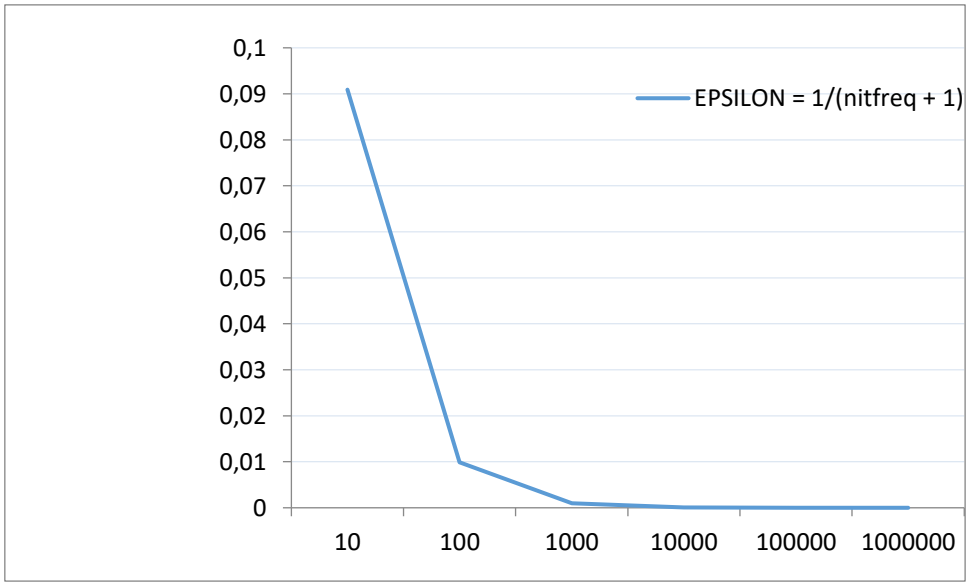

Figure 3 - epsilon trend curve

This curve shows more precisely that ε tends to 0 for all $n_{itfreq} \geq 1000$, $\delta + \varepsilon = \delta$ and ε tends to 1 for all $n_{itfreq} < 1000$, $\delta + \varepsilon = 2\delta$

2.4. PRESENTATION OF THE OPTI2I AND P-FERM ALGORITHMS

ALGORITHM 1: OPTI2I
Input : D - Database
Minsyp, Minconf - Real
K - Integer
Output : Ck - candidate itemsets, Fk - Frequent itemsets
Support, confidence - Real
T - transactions ($t \subset T$)
Start
$K \leftarrow 1$
Ck←candiadats 1-itemsets
Fk←ϕ
For each transaction $t \in D$ Make
If Support.t.items≥Minsup
Then Fk←Ck∪Fk
Else Delete Ck.t.itemsets
EndIf
OPTI2I_Gen (Fk)
EndFor
Sort in descending order frequent t.items $\subseteq$ D
{The frequent 1-itemsets sorted}
{Let's form the frequent 2-itemsets}
$K \leftarrow 2$
If (t.itemsets = 1) and ((t + 1) .itemsets = 1)
Then rule (t.itemsets) $\rightarrow$ ((t + 1) .itemsets) $\leftarrow$ 1
Else If (t.itemsets = 1) and ((t + 1) .itemsets = 0)
or (t.itemsets = 0) and ((t + 1) .itemsets = 1)
Then rule (t.itemsets) $\rightarrow$ ((t + 1) .itemsets) $\leftarrow$ 0
{Transitivity}
Else If (t.itemsets = 1) and ((t + 1) .itemsets = 1)
and (t + 2) .itemsets = 1)
Then rule (t.itemsets) $\rightarrow$ ((t + 2) .itemsets) $\leftarrow$ 1
EndIf
EndIf
EndIf
If Support (t.itemsets) ≥Minsup
Then Fk←Ck∪Fk
Else DeleteCk.t.itemsets
EndIf
OPTI2I_Gen (Fk)
For
Return ∪kFk
End

ALGORITHM 2 : OPTI2I_Gen

Input : t.itemsets
Output : t.candidat, Ck
Begin
For each pair of itemsets
    t.candidat←t.itemsets ∪(t + 1) .itemsets
    Ck←Ck∪t.candidat
EndFor
Return Ck
End

ALGORITHM P-FERM

FFCk - Set of frequent candidate k-itemsets.
FCk - Set of frequent closed delta k-itemsets.
Fk - Set of frequent closed epsilon k-itemsets from FCk
Each element of these sets has three fields :
1) gen: the generator; 2) supp: support
3) deltaferm: δ-closure; 4) epsferm; epsilon closure

Table 2 - Notations and parameters used in P-FERM

Input data
K: Binary database, Sup_min
$Xk \subseteq K$: itemsets located to the right of a ruler
$Ak \subseteq K$: Itemets to the left of a ruler
Delta : Integer
Eps : Real
Output data
$FC = \cup_k FCk$ Set of frequent closed delta itemsets
$F = \cup_k Fk$ Set of frequent precise epsilon itemsets
Begin
{Initialization}
FFC1 = {1-itemsets}
for (k=1 ; FFCk.gen≠ϕ; ;k++) Do
While **(**(Freq(A_k) – Freq(X_{k+1}) ≤ delta) ET (delta > 0) Do
FC_k.deltaferm = ϕ
FFC_k.supp = ϕ
FFC_k = GEN-DELTACLOSURE(FC_k)
EndWhile
Eps = 1 / (FFC_k + 1)
While ABS **(**$Ferm_\delta$(Ai) - δ) <Eps) Do
F_k.epsferm = ϕ
FFC_k = GEN-EPSCLOSURE (F_k)
EndWhile
For all C ∈ FFC_k Do
If C.Supp ≥ Supp_min
Then $FC_k = FC_k \cup C$
EndIf
FFC_k + 1 = GEN-GENERATOR (F_k)
EndFor
EndFor
Return $FC = \cup_k F_k$
End

2.5. THE SELECTION ALGORITHM

```
ALGORITHM SELECT
Var
K1 context of data extraction;
It1: 1-itemsets;
S : Boolean
function SELE (K data extraction context, It: 1-itemsets): Boolean
Var
Sel : Boolean;
Begin
Sel ← False;
If (Support (Iti) ≥ 80% AND (Confidence (Ai → Bi) ≥ 0.75)
Then
P-FERM; // Execution of P-FERM for dense data
Selt ← True;
Else
OPTI2i; // Execution of OPTI2I for scattered data
Sel ← False;
Endif
Return Sel;
End SELE
Begin                    { Main algorithm }
S←SELE (K1, It1);
End
```

Comment

Since the OPTI2i method only optimizes the extraction of 2-itemsets in the context of extracting scattered data, the SELECTION algorithm selects the appropriate algorithm for extracting frequent 2-itemsets. If the extraction context K contains dense data, then the P-FERM algorithm is executed. Otherwise, the OPTI2i scattered data optimization optimization algorithm will run.

4. SECTION 3

RESULTS

Here our experiments are mainly focused on correlated data. The frequent itemsets generated are of sizes k = 1 and k = 2. We will limit ourselves to comparing our ε-precise method of the P-FERM algorithm with δ-free by Clement Faure published in his research work “Discoveries of relevant patterns by the implementation of a Bayesian network : application to industry aeronautics ",

To obtain the results of our work, we program with the PYTHON language, then with the generated data we use Excel software to represent them graphically. The experiments were carried out on the following computer system:

- Core i3 2.4 GHZ processor
- 4 GB RAM
- 500 GB hard drive
- Windows 8.1 operating system
- Office 2013 office software

We used the following four datasets during these experiments:

- C20D10K and C73D10K which are samples from the Public Use Microdata Samples file containing data from the Kansas census carried out in 1990. They consist of 10,000 objects corresponding to the first 10,000 people listed, each object containing 20 attributes (20 items per object and 386 items in total) for C20D10K and 73 attributes (73 items per object and 2,178 items in total) for C73D10K.

Example : Comparison between δ- strong and ε-precise

Let us take the same example as that taken in [Pas00] and [CF07]. It will allow us to compare the δ-strong rule method and our ε-precise approach.

Let be the database K :

Tid	Ti.items
1	{a,c,d}
2	{b,c,e}
3	{a,b,c,e}
4	{b,e}
5	{a,b,c,e}
6	{b,c,e}

Tid	a	b	c	d	e
1	X		X	X	
2		X	X		X
3	X	X	X		X
4		X			X
5	X	X	X		X
6		X	X		X

Figure 4 – Database K

- **The δ-strong rule**
  A rule δ-strong on the database K is an association rule $X \rightarrow Y$, and $Freq(X \cup Y) \geq \gamma$ (γ is a minimum frequency threshold), $Freq(X) - Freq(X \cup Y) \leq \delta$, where $X \cup Y \subseteq$ Items with $Y \neq \emptyset$.
- **Free itemset**
  A free itemset is an itemset that is not included in a closure of its subsets. Example: {a, b} is a free itemset, because Closure (a) = {a, c} and Closure (b) = {b, e}
- **δ-free**
  Let S be a free itemset. S is said to be δ-free if there is no rule δ-strong $X \rightarrow Y$ / $X \cup Y \subseteq S$ with $Y \neq \emptyset$.
- **δ-closure**
  Let S be an itemset ⊆ Items and δ a positive integer, the δ-closure of S is the largest superset of S defined as follows: $Ferm_{\delta}(S) = \{I \subset Items / Freq(S) - Freq(S \cup \{I\}) \leq \delta \}$

- **Let's calculate the frequencies of the k-itemsets**
  Supp_min = 2

k=1		
Freq(a) = 3 Freq(b) = 5	Freq(c) = 5 Freq(d) = 1	Freq(e) = 5
Prune the infrequent 1-itemsets : {d}		
k=2		
Freq(a∪b) = 2 Freq(a∪c) = 3 Freq(a∪e) = 2	Freq(b∪c) = 4 Freq(b∪e) = 4	Freq(c∪e) = 4
k=3		
Freq(a∪b∪c) = 2 Freq(a∪b∪e) = 2	Freq(a∪c∪e) = 2	Freq(b∪c∪e) = 4
k=4		
Freq(a∪b∪c∪e) = 2		

δ=1, 1-exception, i.e. 1-itemset to the right of the rule	
2-item sets	
{ab}	{bc}

$Ferm_1(a) = Freq(a) – Freq(a \cup b) = 1$ {ac} $Ferm_1(a) = Freq(a) – Freq(a \cup c) = 0$ {ae} $Ferm_1(a) = Freq(a) – Freq(a \cup e) = 1$	$Ferm_1(b) = Freq(b) – Freq(b \cup c) = 1$ {be} $Ferm_1(b) = Freq(b) – Freq(b \cup e) = 1$ {ce} $Ferm_1(c) = Freq(c) – Freq(c \cup e) = 1$
3-item sets	
{abc} $Ferm_1(ab) = Freq(ab) – Freq(a \cup b \cup c) = 0$ {abe} $Ferm_1(ab) = Freq(ab) – Freq(a \cup b \cup e) = 0$ {ace} $Ferm_1(ac) = Freq(ac) – Freq(a \cup c \cup e) = 1$ {bce} $Ferm_1(bc) = Freq(bc) – Freq(b \cup c \cup e) = 0$	
4-item sets	
{abce} $Ferm_1(abc) = Freq(abc) – Freq(a \cup b \cup c \cup e) = 0$	
1-fermés fréquents {ab} - {ae} - {bc} - {be} - {ce} - {ace}	fréquentsfermés {ac} - {abc} - {abe} - {bce} - {abce}
<u>δ=2</u>, 2-exceptions, i.e. 2-itemsets to the right of the rule	
3-item sets	
{abc} $Ferm_2(a) = Freq(a) – Freq(a \cup b \cup c) = 1$ {abe} $Ferm_2(a) = Freq(a) – Freq(a \cup b \cup e) = 1$ {ace} $Ferm_2(a) = Freq(a) – Freq(a \cup c \cup e) = 1$ {bce} $Ferm_2(b) = Freq(b) – Freq(b \cup c \cup e) = 1$	
4-item sets	
{abce} $Ferm_1(ab) = Freq(ab) – Freq(a \cup b \cup c \cup e) = 0$	
2-fermés frequents {abc} - {abe} - {ace} - {bce}	fréquentsfermés {abce}
<u>δ=3</u>, 3-exceptions, i.e. 3-itemsets to the right of the rule	
4-item sets	
{abce} $Ferm_3(a) = Freq(a) – Freq(a \cup b \cup c \cup e) = 1$	
3-closed frequent {abce}	Frequent closed { }
δ- closed frequent {ab} - {ae} - {bc} - {be} - {ce} - ~~{ace}~~ {abc} - {abe} - {ace} - {bce} {abce}	Frequent closed {ac} - {abc} - {abe} - {bce} - ~~{abce}~~ {abce}

Figure 5 - Extraction of δ- frequent closed

- **Let's calculate ε**

  Let nitfreq be the sample from database K. (n_{itfreq} corresponds to the number of frequent δ-closed itemsets and nitfreq = 10

  ε = 1 / (nitfreq + 1) = 1 / (10 + 1) = 1/11 = 0.090909….

  If we consider the trend curve of the parameter ε, we see that ε→ 1 (n_{itfreq} = 10 <1000).

  The frequent (δ+ε) -closed are: {abc} - {abe} - {ace} - {bce} and {abce}.

  We can use the combinations of frequent 1-closed in Figure 3.

  So let's group them together: ab + ae = abe (which means that a appears with b and a appears with e. So we have abe together). Likewise ab+bc = abc; bc+be=bce; be+ce=~~bce~~.

Proposition 1:

Let the sets of items $\{a_1, a_2,\ldots, a_k\}$, $\{b_1, b_2,\ldots, b_m\}$ and $\{c_1, c_2,\ldots, a_l\}$ with $k \neq m \neq l$. When two frequent k-itemsets are δ-closed, then for all ε> 0 the sums a_kb_m and b_mc_l produce (k + 1)-itemsets $a_kb_mc_l$.

Let us calculate the Confidences of the association rules of δ-strong and ε-precise

Let's fix Conf_min = 50% :

δ-strong		**ε-precise**
Confiance(a→b) = 66%	Confiance(a→bc)=66%	Confiance(ab→e)=100%
Confiance(a→e) = 66%	Confiance(a→be)=66%	Confiance(a→be)=66%
Confiance(b→c) = 80%	Confiance(a→ce) =66%	Confiance(ab→c) =100%
Confiance(b→e) = 80%	Confiance(b→ce) =80%	Confiance(a→bc) =66%
Confiance(c→e)= 100%	Confiance(a→bce) =66%	Confiance(bc→e) =100%
		Confiance(b→ce) =80%

Figure 6 - Association rules with confidence ≥50% for δ-strong and ε-precise

We note that the precise rule produces fewer association rules than the strong rule. **ε**-precise also shows that the association rules with exactly one exception (1-closure) are more relevant with 100% confidence.

- **Side-by-side storage of items**

After optimizing the extraction of closed frequent itemsets, we will proceed to arrange them side by side. Thus the items from the association rules with the highest confidence will be ranked first and the others will follow. If two association rules have identical trusts, then their premises will be placed one after the other and their consequences will follow immediately.

In the arrangement of items, we will only retain frequent itemsets. Let's use the trusts found in Figure 6 for our rule (i.e. the **ε**-precise rule). The association rules ab→e, ab→c and bc→e have identical Confidences and are the highest. Hence their items are classified as follows: abce or bace. The other association rules b→ce, a→be and a→bc will not participate in the ranking, because the items in premise and in conclusion have already been classified. Thus the articles are classified as follows: a→b→c→e or b→a→c→e.

Proposition 2 :

Let δ-closure of A_i, for all ε→1 the (δ+ε)-closure can be obtained by grouping the (δ-1)-closure of A_i.

Now let's test the response times of the δ-strong and ε-precise rules on the correlated (dense) data sets C20D10K and C73D10K

- **Dataset C20D10K**

Support	δ- closed frequent	δ-strong	ε-precise	Time difference : $D_t = T_\varepsilon - T_\delta$
20	1 255	2,43	2,32	-0,11
15	3 289	3,68	2,51	-1,17
10	8 578	8,15	8,03	-0,12
7,5	15 890	12,72	11,64	-1,08
5	37 123	17,63	16,49	-1,14
2,5	125 257	29,28	28,16	-1,12

Table 3: Response time for C20D10K

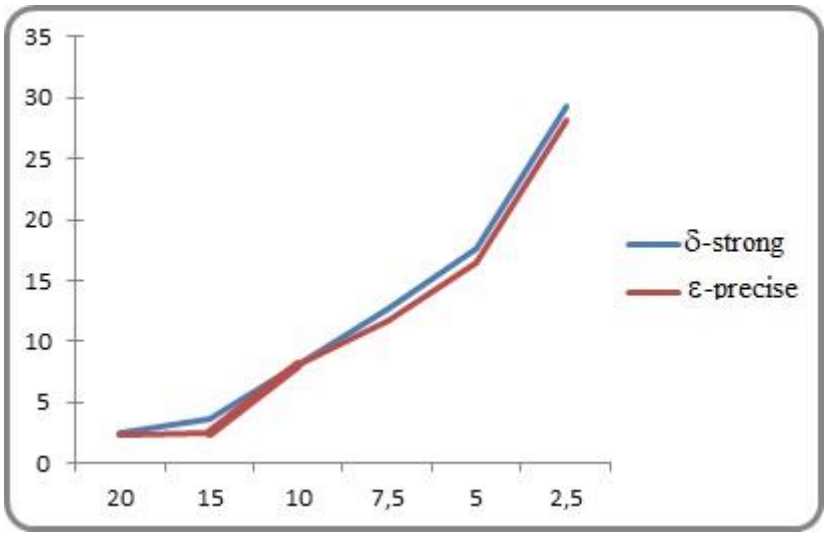

Figure 6: Experimental results for C20D10K

– **Dataset C73D10K**

Support	δ- closed frequent	δ-strong	ε-precise	Time difference : $D_t = T_\varepsilon - T_\delta$
80	10 924	45,00	42,11	-2,89
75	22 346	95,72	93,58	-2,14
70	64 495	185,37	172,13	-13,24
60	427 375	401,68	389,02	-12,66

Table 4: Response time for C73D10K

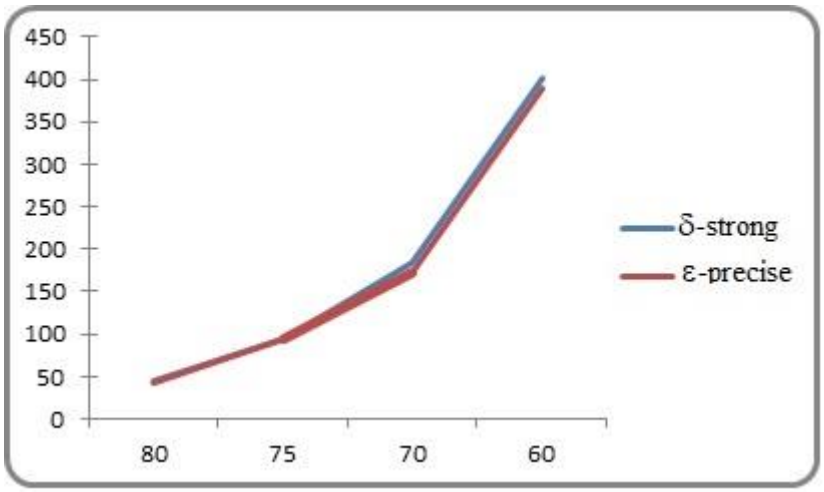

Figure 7: Experimental results for C73D10K

5. DISCUSSION

The color curves in Tables 3 and 4 show response times greater than 0. Figures 6 and 7 show the experimental results obtained by our ε-precise method and the δ-strong method. We observe in these figures that the discovery times of frequent δ-closed itemsets have been significantly improved by the ε-precise rule. The improvement in these times is linked to the accuracy of extracting a more concise number of association rules, thereby considerably eliminating all unnecessary and redundant rules. It should be noted that our rule therefore solves the problem of optimizing the extraction of association rules and also δ-closed itemsets. As we indicated in our state of the art, the main objective of this article is not to compare the ε-precise and δ-strong methods. But to solve the problem of optimizing the frequent closed 2-itemsets that the OPTI2i algorithm did not allow. However, our ε-precise method has improved the δ-strong rule, and therefore provides an overall gain in response time by extracting the association rules for frequent itemsets.

6. CONCLUSION AND PERSPECTIVES

In this paper, the problem of the usefulness and the relevance of the extracted association rules is treated using the δ-closure based on the δ-strong rule. We have proposed a parameter ε which determines a margin of error. The ε parameter made it possible to correct the δ-strong rule, by reducing the concise number of discovered association rules while at the same time removing the redundant rules. The P-FERM algorithm improves the extraction response time of the extraction rules compared to the δ-strong rule. A first perspective of the next works concerns the very significant improvement of the extraction response times of a concise and precise number of association rules. Another perspective could consider improving the δ-strong rule by another method than the one we have proposed in this paper in order to take into account both correlated and uncorrelated datasets. Moreover, the classification of articles side by side is done using the same logic as in our paper entitled "OPTI2i - A new approach to optimize the extraction of frequent 2-itemsets".

REFERENCES

[1]. R. Agrawal, R. Srikant. Fast algorithms for mining association rules in large databases. Proc. VLDB conf., pp 478−499, September 1994.
[2]. N. Pasquier Y. Bastide R. Taouil et L. Lakhal. Pruning closed itemset lattices for association rules. In Proceedings of the 14th Advanced Databases Day (BDA'98), pages 177–196, 1998.
[3]. Zaki M.J., Hsiao C.-J., ChARM: An Efficient Algorithm for Closed Association RuleMining. Technical Report 99-10, Renssaeler Polytecnic Institute, Troy, New York, 1999.
[4]. R. J. Bayardo, R. Agrawal, D. Gunopulos. Constraint-based rule mining in large, dense databases. Proc. ICDE conf., pp 188−197, March 1999.

[5]. N. Pasquier. Datamining : algorithms for extracting and reducing association rules in databases. Doctoral thesis, University of Clermont-Ferrand II, January 2000.
[6]. Bastide Y., Taouil R., Pasquier N., Stumme G., Lakhal L. « PASCAL : a frequent pattern extraction algorithm », Computer Science and Technology, vol. 21, n° 1, 2002, p. 65-95.
[7]. JAOUA A., ELLOUMI S., BENYAHIA S., ALVI F., « Galois connection in fuzzy binary relations: applications for discovering association rules and decision making », Methodos Publisher, 2002.
[8]. KRYSZKIEWICZ M., « Concise Representations of Association Rules », HAND D. J., ADAMS N., BOLTON R., Eds., Proceedings of Pattern Detection and Discovery, ESF Exploratory Workshop, London, UK, vol. 2447 Reading Notes in Computer Science, Springer, September 2002, p. 92 109.
[9]. Boulicaut J.-F., Bykowski A. & Rigotti C. (2003). Free-sets: a condensed representation of boolean data for the approximation of frequency queries. Data Mining and Knowledge Discovery, 7(1), 5–22.
[10]. S. Ben Yahia et E. MephuNguifo. Approaches to extracting association rules based on Galois match. ISI (Information Systems Engineering), Hermes-Lavoisier, 34(9) :23–55, 2004.
[11]. Uno, T., M. Kiyomi, et H. Arimura (2004b). Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In FIMI.
[12]. Wang, J. et G. Karypis (2005). HARMONY: Efficiently mining the best rules for classification. In SIAM'05.
[13]. Calders, T., C. Rigotti, et J.-F. Boulicaut (2006). A survey on condensed representations for frequent sets. In Constraint Based Mining, Springer-Verlag, LNAI, volume 3848, pp. 64–80.
[14]. Fournier-Viger, P., Gomariz, A., Campos, M., Thomas, R.: Fast Vertical Sequential Pattern Mining Using Co-occurrence Information. In: Proc. PAKDD 2014, pp. 40-52. (2014)
[15]. Soulet, A., Rioult, F.: Efficiently Depth-First Minimal Pattern Mining. In: Proc. 18th Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 28–39 (2014)
[16]. M., G, M., Mohamed, M., and Smarandache, F. A new method to solve the problems of totally neutral philosophical programming. Neural calculation and applications, 1-11.
[17]. M., M., G., Fakhry, A.E. et El-Henawy, I. (2018). 2 niveaux de stratégie de regroupement pour détecter et localiser la falsification par transfert de copie dans les images numériques. Outils multimédia et applications, 1-19.
[18]. Abdel M., & Mohamed, M. (2018). Internet of Things (IoT) and its impact on the supply chain: a framework for the creation of intelligent, secure and efficient systems. Next generation IT systems.
[19]. Basset, M., G. Manogaran, M. Mohamed, and Rushdy, E. Internet of things in an intelligent educational environment: support framework in the decision-making process. Competition and calculation: practice and experience, e4515.
[20]. Abdel, M., M., G., Rashad, H., and Zaied, A.N.H. (2018). A comprehensive review of the quadratic assignment problem: variants, hybrids and applications. Ambient Intelligence and Humanized Computing Journal, 1-24.
[21]. Ab, M., G, M., Mohamed, M. and Chilamkurti, N. (2018). Three-way decisions based on neutrosophical sets and the AHP-QFD framework for the problem of supplier selection. Next generation IT systems.