A Novel Algorithm for Mining Negative and Positive Association Rules

Parfait Bemarisika
Laboratory of Mathematics and Computer Science
ENSET-University of Antsiranana, Madagascar.
Email: bemarisikap7 {at} yahoo.fr

André Totohasina
Laboratory of Mathematics and Computer Science
ENSET-University of Antsiranana, Madagascar.
Totohasina

Abstract—**The extraction of association rules is one of the most popular tasks in data mining. According to our point of view, few methods have been dedicated to the extraction of negative association rules, mainly because of the high cost of calculation and the prohibitive number of association rules extracted, several of which are redundant and uninteresting. During the last two years, the research on such extraction takes its size, but it is still a hot topic for researchers in this field. In this work, we shall introduce a new approach to generating negative association rules from positive rules, by using the quality measure called M$_{GK}$. So, we have proposed a new algorithm to improve the computation time and quality of extracted association rules. We conducted experiments using several databases to test the performance of our algorithm.**

Keywords-algorithm; data mining; association rules; positive rules; negative rules; quality measure M$_{GK}$.

I. INTRODUCTION

The research on the extraction of association rules has been renewed in [1] and has progressively developed. According to our point of view, few approaches have been devoted to the extraction of negative rules, even if they have been frequently studied during the last two years. In many situations, it may be interesting to consider the negative rules, that is, taking into account the absence of certain patterns whose fields of application can be described as multiple, including: social sciences, medicine, telecommunication, computer science and didactic of discipline. Interest in negative association rules has been developed in [3], the authors use the well known statistical measure Chi-square, usually described as χ^2, in order to extract rules of correlation between two patterns. We can note some remarkable works that have been elaborated. In [12], the authors combine the support and confidence to generate the frequent patterns with the objective of detecting positive and negative association rules. In [2], the authors propose the discovery of the negative rules of type $X \wedge Y \to \overline{Z}$ or $\overline{X} \wedge Y \to Z$ using an approach based on extracting generalized constraint patterns containing negations of attributes. In [21], the authors present an algorithm generating negative rules of the form: $X \to \overline{Y}$. Based on the literal context, in [20] the authors proposed the computing of positive and negative association rules in order to get rules of the form $\overline{X} \to Y$. In [5], the authors use the couple support-confidence to generate positive and negative rules. Feno and Totohasina [6-15] study the mathematical properties of the measure M_{GK} to allow the extraction of negative rules. In [9], the authors propose generating negative rules from positive rules in a context of logical implications. In [4], the authors use of positive and negative association rules based on support and confidence to improve the accuracy of a classifier. In [18], the authors propose a retrieval algorithm of the negative rules based on couple support-confidence and the measure of linear Correlation, in order to present the importance of decision analysis. Tushar [16] proposes an extraction of the negative rules of algorithm using the following couple support-confidence and measure Conviction, to build up a classification mode. In [10], the authors present an approach to improve the extraction of positive and negative rules based on both a genetic algorithm and the couple support-confidence and the extent of Correlation. In [11], Rakesh and Narayna propose an algorithm for the extraction of positive and negative rules, by a coherent approach, via the couple support-confidence. Ramakrishnudu and Sbramanyam study the negative association rules based on tree function [17]. In [13], Guillaume and Papon proposed an algorithm for extracting negative rules using the couple support-confidence, and the extent of partially M$_G$ (Guillaume Measure) for the rule $\overline{X} \to \overline{Y}$. It is a variant of the Apriori algorithm. As we have noted, the research based on negative rules takes its size. In spite of this approach using only the confidence, various studies [7] have illustrated that we can improve the quality of extracted rules when using other more quality measures. In this work, our main concern shall be dealing with the problem of extracting positive and negative rules of the forms $X \to Y$, $X \to \overline{Y}$, $\overline{X} \to Y$ and $\overline{X} \to \overline{Y}$ by using mainly the quality measure M$_{GK}$. However, this problem is exponential in order to partially resolve in size extracted association rules and of calculations cost, with several redundant or uninteresting rules. It is therefore necessary to define an efficient algorithm to perform this type of knowledge. In this article, we have proposed a new algorithm generating both positive and negative rules according to the measure M$_{GK}$ (Guillaume Khenchaff's Measure). The rest of the paper is organized as following. Some preliminary and motivations are set out in section 2. The section 3 is devoted to the description of the method that we offer. The algorithm proposed is described in section 4. Experimental evaluation will be presented in form of summary in section 5. In conclusion, we shall present our results and research perspectives.

II. Preliminaries and Motivations

In this section, we shall introduce all the basic concepts that will be useful later in this work. Therefore, we deal with a binary context.

Definition 1. *A binary context is a triplet* $K = (T, I, R)$, *where T and I are respectively finite sets of transactions (or objects) and items (or attributes), and* $R \subseteq I \times T$ *is a binary relation between the transaction and the item. A couple* $(i,t) \in R$ *denotes the fact that the transaction* $t \in T$ *contains the item* $i \in I$ *.*

It is assumed that the data *K* to explore are binary, its means we can describe each transaction by means of a finite set of items $I = \{i_1, ..., i_m\}$, also called attributes. Each transaction *T* will be a subset of *I*. Furthermore, it combines each transaction identifier TID (Transaction IDentify): $T = \{t_1, ..., t_n\}$, that is to say $\forall (i,t) \in I \times T$, *t*[*i*]=1 if the item *i* is present in *t* and *t*[*i*]= 0 otherwise. The pattern *X* is a subset of items I $(X \subseteq I)$. Below, the table 1 shows an example of binary context with four items {A, B, C, D} and five transactions {1, 2, 3, 4, 5}. Let $X' = \{t \in T \mid \forall i \in X, iRt\}$ the set of all public entities to all elements *X*, this is the dual of the pattern *X*.

TABLE I. EXAMPLE OF BINARY CONTEXT

TID	A	B	C	D
1	1	1	1	1
2	0	1	0	0
3	1	0	1	0
4	0	1	1	0
5	1	1	0	1

By considering the context of table *I*, for example $X = \{AB\}$, we have: $X' = \{1,5\}$ and its logical negation is presented by $\overline{X'} = \{2,3,4\}$. In the rest of this paper, to simplify the notation, let us denoted $P(Y \mid X) = P(Y' \mid X')$, where X' is the extension (or dual) of the pattern *X*.

Definition 2. *An association rule is a quasi implication of the form* $X \rightarrow Y$ *, where X and Y are disjointed patterns (* $X, Y \subseteq I$ and $X \cap Y = \varnothing$ *) respectively called the premise and the consequent of the rule.*

Example 1. Let us consider a rule *Computer → Printer*. It can be explained as following: basically, if a customer wants to buy a *Computer*, he or she also needs to buy a *Printer*. A typical application of mining association rules is *"analysis of the food basket"*. The objective is to provide receipts to customers in order to help them understanding, for example, their consumption habits, organizing store rays, organizing promotions and manage inventory, in order to improve the profit.

This association may be quantified by a set of quality measures. Both more conventional measures are *support* and *confidence.*

Definition 3. *We define the support of an itemset X, denoted by supp(X), the ratio between the number of t transactions containing X and the total number of transactions in data base K.*

$$\text{supp}(X) = \frac{|\{t \in T : X \subseteq t\}|}{|T|}$$

where, for any finite set A, |A| denotes the cardinality of its elements.

Example 2. According to table I, we have:

$$\text{supp}(AB) = 2/5 = 0.4$$

Definition 4. *The support an association rule X→Y, denoted supp(X→Y), is the proportion transactions in the database containing the itemset X ∪ Y.*

$$\text{supp}(X \rightarrow Y) = \text{supp}(X \cup Y) = p(X \cap Y) = \frac{|t \in T : X \subseteq t, Y \subseteq t|}{|T|}$$

Example 3. According to table I, we have:

$$\text{supp}(C \rightarrow D) = 1/5 = 0.2$$

Definition 5. *The confidence of an association rules X→Y, denoted by conf(X→Y), is the ratio between the number of transactions containing the itemset X∪Y and the number of those containing X. This is a realization of the conditional probability given X' of Y' in the finite uniform of the probability space (T, P(T), P).*

$$conf(X \rightarrow Y) = p(Y \mid X) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{p(X \cap Y)}{p(X)}$$

Where $p(Y \mid X)$ *is the conditional probability of Y given X.*

Example 4. From to table I, we have:

$$conf(A \rightarrow B) = p(B \mid A) = 2/3 = 0.66$$

As we know, trust is the most classical measure in the evaluation process of association rule. However, it cannot itself guarantee the quality of association rule. The example in the following table II illustrates this.

TABLE II. DEFAULT OF CONFIDENCE

	coffee	$\overline{coffee}$	Σ
tea	20	5	25
$\overline{tea}$	70	5	75
Σ	90	10	100

Therefore, let us consider the rule *tea → coffee*. It support is equal to 0.2 and confidence is 0.8. These reasonably high values for measures of support and confidence invite us to think that customers who purchase *tea* also buy *coffee (tea* favors *coffee).* However, the share of customers purchasing *coffee* is independently of the *tea* is *P(coffee|tea)=P(coffee)=0.9, w*hereas the rule tells us that the proportion of customers consuming *tea* and *coffee* is less higher, since it is equal to 0.8. The rule *tea → coffee* that seems interesting is therefore misleading. The confidence measure is insensible to independence. It easily produces the rules of independent association. By regarding this problem of failure, the use of other measures seems therefore necessary.

The following section exposes the description of our method by using another interestingness measure.

III. DESCRIPTION OF THE METHOD

As we have described in the previous section, the confidence measure selects easily uninteresting rules. In addition, the confidence captures only the rules located in the attraction area[1], it is unable to select the rules found in the repulsion[2] area. While several new rules in this second area could be really interesting for the user. Thus, we have introduced another method capable of detecting the rules in both areas. We know that the association rules extraction in such dual area is exponential, mainly due to the number of rules and the cost of calculations. As a result, we shall optimize the following two points:

- *Elimination of uninteresting rules: improving the quality of association rules;*
- *To go through the generation of the whole association rules: improving the cost of calculations.*

We are particularly interested in the rules of generation by the use of the measure M_{GK}[8]. In the first hand, we shall demonstrate how to eliminate the rules having no real interest to the user, in which we added four new properties. In the second hand, we shall focus on how to reduce the different steps of researching set of rules, using the added properties.

As a result, we shall roughly begin with, the measure M_{GK}.

Definition 6. *For all association rules $X \rightarrow Y$, the quality measure M_{GK} is defined as:*

$$M_{GK}(X \rightarrow Y) = \begin{cases} \dfrac{p(Y \mid X) - p(Y)}{1 - p(Y)}, if \quad p(Y \mid X) \geq p(Y) \\ \dfrac{p(Y \mid X) - p(Y)}{p(Y)}, if \quad p(Y \mid X) < p(Y) \end{cases} \qquad (1)$$

Let us notice that WU[19] calls M_{GK} as a conditional probability incriminations ratio (CPIR), and Totohasina[14] calls ION. It is varying to show that $-1 \leq M_{GK} \leq 1$[15].

If $M_{GK}(X \rightarrow Y) = 1$, then the attraction between the premises X and therefore Y is high, that is X and Y are positively dependent. In this case, the rule $X \rightarrow Y$ and $\overline{X} \rightarrow \overline{Y}$ could really be interesting.

If $M_{GK}(X \rightarrow Y) = 0$, that is X and Y are independent, the rule $X \rightarrow Y$ is not interesting.

If $M_{GK}(X \rightarrow Y) = -1$, that is X and Y are stochastically incompatible, there is therefore a strong repulsion between X and Y, i.e. X and Y are negatively dependent. The rules $X \rightarrow \overline{Y}, \overline{X} \rightarrow Y, Y \rightarrow \overline{X}$ and $\overline{Y} \rightarrow X$ that could be interesting. The main mathematical properties characterizing this measure of quality are developed in [6-15]. Therefore, we urge interested reader to check their works.

Example 5 Using the example of table II: *tea → coffee.* Quality as measured M_{GK} is obtained by:

$$M_{GK}(tea \rightarrow coffee) = \frac{p(coffee \mid tea) - p(coffee)}{1 - p(coffee)} = 0$$

This result explains the statistical independence between *tea* and *coffee.* This clearly shows that the rule *tea → coffee* is not interesting. Thus, the measure M_{GK} is indeed sensitive to independence. It censures the independence.

A. Elimination of unintersting association

Definition 7. *A rule $X \rightarrow Y$ is potentially interesting if $p(Y \mid X) \geq p(Y)$, thus $M_{GK}(X \rightarrow Y) \geq 0$ otherwise it is not interesting, we note $\overline{X} \rightarrow Y$.*

The following proposition 1 partition the set of rules divided into two classes.

Proposition 1. *If the rule $X \rightarrow Y$ is potentially interesting, then $Y \rightarrow X, \overline{Y} \rightarrow \overline{X}$ and $\overline{X} \rightarrow \overline{Y}$ will be also, and else, then*

[1] A repulsion area is an area between the incompatibility (a point where $p(Y \mid X) = 0 \Leftrightarrow M_{GK}(X \rightarrow Y) = -1$ and independence ($p(Y \mid X) = p(Y) \Leftrightarrow M_{GK}(X \rightarrow Y) = 0$).

[2] An attraction area is an area between the independence and logical implication ($p(Y \mid X) = 1 \Leftrightarrow M_{GK}(X \rightarrow Y) = 1$).

$X \rightarrow \overline{Y}, \overline{X} \rightarrow Y, \overline{Y} \rightarrow X$ and $Y \rightarrow \overline{X}$ *will be interesting [14].* This proposition indicates that when the rules $X \rightarrow Y$ is potentially interesting health, then the rules $Y \rightarrow X, \overline{Y} \rightarrow \overline{X}$ *and* $\overline{X} \rightarrow \overline{Y}$ are evaluated, when $X \rightarrow Y$ is not interesting, while the rules $X \rightarrow \overline{Y}, \overline{X} \rightarrow Y, \overline{Y} \rightarrow X$ *and* $Y \rightarrow \overline{X}$ are studied.

The following statements describe our method for evaluation of positive and negative rules.

Proposition 2. *If* $X \rightarrow Y$ *with* $p(X) > p(Y) > 1/2$ *is considered as pertinent according to M_{GK}, then* $Y \rightarrow X$ *is also, and if* $X \rightarrow Y$ *with* $p(X) < p(Y)$ *is not interesting, then* $Y \rightarrow X$ *is not interesting also.*

Proof. If $p(Y \mid X) \geq p(Y)$, we have:

$$M_{GK}(Y \rightarrow X) = \frac{p(X)p(\overline{Y})}{p(\overline{X})p(Y)} M_{GK}(X \rightarrow Y) \qquad (2)$$

$$\text{else, } M_{GK}(Y \rightarrow X) = M_{GK}(X \rightarrow Y) \qquad (2')$$

For $p(X) > p(Y) > 1/2 \Leftrightarrow p(\overline{X}) < p(\overline{Y}) < 1/2$, was: $p(X)p(\overline{Y}) > p(\overline{X})p(Y)$. According to the equation (2), we have $M_{GK}(Y \rightarrow X) > M_{GK}(X \rightarrow Y)$. This indicates that if $X \rightarrow Y$ is considered interesting under M_{GK}, then a rule $Y \rightarrow X$ is also interesting. Therefore, according to equation (2'), we have, with $p(Y \mid X) \geq p(Y)$, which shows that if $X \rightarrow Y$ is not interesting, and then a rule $Y \rightarrow X$ is also not interesting.

Proposition 3. *If the rule* $X \rightarrow Y$ *is considered interesting by M_{GK}, then* $\overline{Y} \rightarrow \overline{X}$ *is interesting, and if* $X \rightarrow Y$ *is not interesting, with* $1/2 < p(X) < p(Y)$ *, then* $\overline{Y} \rightarrow \overline{X}$ *will also not interesting.*

Proof. If $X \rightarrow Y$ is M_{GK}-valid, we have:

$$M_{GK}(\overline{Y} \rightarrow \overline{X}) = M_{GK}(X \rightarrow Y) \qquad (3)$$

$$\text{else, } M_{GK}(\overline{Y} \rightarrow \overline{X}) = \frac{p(X)p(Y)}{p(\overline{X})p(\overline{Y})} M_{GK}(X \rightarrow Y) \qquad (3')$$

The equation (3) shows that, in this case X favors Y, if $X \rightarrow Y$ is interesting, then $\overline{Y} \rightarrow \overline{X}$ is also interesting. Second, by (3'), with $1/2 < p(X) < p(Y) \Leftrightarrow p(\overline{Y}) < p(\overline{X}) < 1/2$, was $p(X)p(Y) < p(\overline{X})p(\overline{Y})$, we have: $M_{GK}(\overline{Y} \rightarrow \overline{X}) \leq M_{GK}(X \rightarrow Y)$. This indicates that if $X \rightarrow Y$ is not interesting then $\overline{Y} \rightarrow \overline{X}$ is also not interesting.

Proposition 4. *If* $X \rightarrow Y$ *is, with* $p(X) > p(Y)$, *interesting to M_{GK}, then a rule* $\overline{X} \rightarrow \overline{Y}$ is *also, and if* $X \rightarrow Y$ *, with* $p(X) < p(Y) < 1/2$, is *not interesting, then* $\overline{X} \rightarrow \overline{Y}$ *is also not interesting.*

Proof. If $p(Y \mid X) \geq p(Y)$, we have:

$$M_{GK}(\overline{X} \rightarrow \overline{Y}) = \frac{p(\overline{X})p(Y)}{p(X)p(\overline{Y})} M_{GK}(X \rightarrow Y) \qquad (4)$$

$$\text{Else, } M_{GK}(\overline{X} \rightarrow \overline{Y}) = \frac{p(X)p(Y)}{p(\overline{X})p(\overline{Y})} M_{GK}(X \rightarrow Y) \qquad (4')$$

By taking into account of the following hypothesis: $p(X) > p(Y) \Leftrightarrow p(\overline{X}) < p(\overline{Y})$, we have: $p(X)p(\overline{Y}) > p(\overline{X})p(Y)$. By equation (4), we have $M_{GK}(\overline{X} \rightarrow \overline{Y}) \geq M_{GK}(X \rightarrow Y)$, if $p(X) > p(Y)$. This illustrates that if $\overline{X} \rightarrow \overline{Y}$ is interesting, and then a rule $X \rightarrow Y$ is also interesting. In other words, by hypothesis $p(X) < p(Y) < 1/2$, was: $p(X)p(Y) < p(\overline{X})p(\overline{Y})$. Hence after (4'), we have: $M_{GK}(\overline{X} \rightarrow \overline{Y}) \leq M_{GK}(X \rightarrow Y)$, if $p(X) < p(Y) < 1/2$. This demonstrates that if $X \rightarrow Y$ is not interesting, then the rule $\overline{X} \rightarrow \overline{Y}$ is also not interesting.

Proposition 5. *If* $X \rightarrow \overline{Y}$ *is valid by M_{GK}, then* $\overline{X} \rightarrow Y$, $Y \rightarrow \overline{X}$ *and* $\overline{Y} \rightarrow X$ *are as valid as M*_{GK}.

Proof. Indeed $M_{GK}(X \rightarrow \overline{Y}) = \frac{p(\overline{Y} \mid X) - p(\overline{Y})}{p(Y)} = \frac{p(\overline{X} \mid Y) - p(\overline{X})}{1 - p(\overline{X})}$

we have: $M_{GK}(X \rightarrow \overline{Y}) = M_{GK}(Y \rightarrow \overline{X})$. What shows that if $X \rightarrow \overline{Y}$ is valid in *M*_{GK} (*M*_{GK}-valid), then $Y \rightarrow \overline{X}$ is valid by *M*_{GK}.

Also, $M_{GK}(X \rightarrow \overline{Y}) = \frac{p(\overline{Y} \mid X) - p(\overline{Y})}{p(Y)} = \frac{p(\overline{X} \cap Y) - p(\overline{X})p(Y)}{p(X)p(Y)}$,

$$\text{we have: } M_{GK}(X \rightarrow \overline{Y}) = \frac{p(\overline{X})p(\overline{Y})}{p(X)p(Y)} M_{GK}(\overline{X} \rightarrow Y) \qquad (5)$$

And $M_{GK}(X \rightarrow \overline{Y}) = \frac{p(\overline{Y} \mid X) - p(\overline{Y})}{p(Y)} = \frac{p(X \cap \overline{Y}) - p(X)p(\overline{Y})}{p(X)p(Y)}$, we

$$\text{have } M_{GK}(X \rightarrow \overline{Y}) = \frac{p(\overline{X})p(\overline{Y})}{p(X)p(Y)} M_{GK}(\overline{Y} \rightarrow X) \qquad (5')$$

Identifying (5) and (5'), we have $M_{GK}(\overline{X} \rightarrow Y) = M_{GK}(\overline{Y} \rightarrow X)$. This shows that, if $\overline{X} \rightarrow Y$ is M_{GK}-valid, then $\overline{Y} \rightarrow X$ is valid by *M*_{GK}. Finally, there is in this case of the presence of a rule $X \rightarrow \overline{Y}$ interesting, but a rule $X \rightarrow Y$ not interesting. We are therefore if the number of samples-against is much more than that of examples. So is $1/2 < M_{GK}(X \rightarrow Y)$. It is necessary that $p(X) > 1/2 > p(Y)$ is equivalent to $p(\overline{X}) < 1/2 > p(\overline{Y})$; so $p(\overline{X})p(\overline{Y}) \leq p(X)p(Y)$. So, according to (5), we have:

$M_{GK}(X \to \overline{Y}) \leq M_{GK}(\overline{X} \to Y)$. This completes the proof of the demonstration.

B. The steps of generation of set association rules

We would like to recall that Proposition 1 allows proceeding to the partition of the set of rules in two areas: *attraction area and repulsion area*. In the attraction area (*X* favors *Y*), we study the rules of type $X \to Y, Y \to X, \overline{X} \to \overline{Y}$ and $\overline{Y} \to \overline{X}$. While in the repulsion area (*X* disfavors *Y*), we have analyzed the rules of type $X \to \overline{Y}$, $\overline{X} \to Y$, $Y \to \overline{X}$ and $\overline{Y} \to X$. Indeed, if *X* implies *Y*, we have described according to the propositions 2 and 3 that the rule $X \to Y$, $Y \to X$ and $\overline{Y} \to \overline{X}$ are equivalent. This illustrates that only one rule among the three suggested is efficient for the result of our study. Moreover, we have also been proved that according to the Proposition 4 that if the rule $X \to Y$ is M$_{GK}$-valid, then $\overline{Y} \to \overline{X}$ is also M$_{GK}$-valid. As a result, in the attraction area, we only need to evaluate the rules $X \to Y$ and $\overline{X} \to \overline{Y}$. It is the half of set association rules that are relevant within this area. If *X* disfavors *Y*, we have mentioned in the Proposition 5 that the rules $X \to \overline{Y}$ and $Y \to \overline{X}$ (resp. $\overline{X} \to Y$ and $\overline{Y} \to X$) are equivalent. Finally, in the repulsion area, it is necessary to study the rules *(* $X \to \overline{Y} \wedge Y \to \overline{X}$)or ($\overline{X} \to Y \wedge \overline{Y} \to X$ *)*. It is also half of set association rules in this area. In conclusion, with this new method that we have presented, we can conclude that only half of the set of rules are efficiently analyzed. This situation explains the reason why there is a considerable number of rules to evaluate: hence improving the calculations cost is needful.

We have to expose the description of our method on the problem of extracting negative and positive rules; the resulting algorithms are presented in next section.

IV. About the proposed algorithm

The objective of our algorithm is to generate all relevant rules according to measure M_{GK} from the set of frequent itemsets *F* by Apriori algorithm. Thus, the problem of such generation is restated as following. Given two disjoint patterns *X*, *Y* in *F* and a minimum threshold $\min M_{GK}$ for measure M_{GK}, find the set of valid rules R as:

$$R = \{X \to Y \mid X, Y \subseteq F, X \cap Y = \varnothing : M_{GK}(X \to Y) \geq \min M_{GK}\}$$

The proof of the algorithm is obtained using the proposals we set in the previous section. As its name suggests the extraction algorithm of negative and positive rules (ARNP) generates not only negative but also positive rules. It optimizes the solution of the problem of exact rules $(M_{GK}(X \to Y) = 1)$ and approximate rules $(\min M_{GK} \leq M_{GK}(X \to Y) \leq 1)$. The pseudo-code of this algorithm is presented in Algorithm 1.

Algorithm 1: Algorithm of negative and positive association rules (ARNP)

Input: Set of frequent patterns *F*, $\min M_{GK}$ minimum threshold.

Output: A set of positive and negative rules valid.

```
1: begin
2:  R ← {};
3:   for all $X \in F$ do
4:     for all $Y \in F$ do
5:       $M^f_{GK}(X \to Y) = \frac{p(Y|X) - p(Y)}{1 - p(Y)}, (p(Y|X) \geq p(Y));$
6:       $M^d_{GK}(X \to Y) = \frac{p(Y|X) - p(Y)}{p(Y)}, (p(Y|X) \geq p(Y));$
7:       if ($M^f_{GK}(X \to Y) \geq 0$ $X \to Y$ is interesting) then
8:         if ($M^f_{GK}(X \to Y) \geq \min M_{GK}$) then
9:           $R = R \cup \{(X \setminus Y \to Y) \wedge (\overline{X} \setminus \overline{Y} \to \overline{Y})\};$
10:        end
11:      else ($M^d_{GK}(X \to Y) < 0$, $X \to Y$ is not interesting) then
12:        if ($M^d_{GK}(X \to \overline{Y}) \geq \min M_{GK}$) then
13:          $R = R \cup \{(X \setminus \overline{Y} \to \overline{Y})\};$
14:          if ($M^d_{GK}(\overline{X} \to Y) \geq \min M_{GK}$) then
15:            $R = R \cup \{(\overline{X} \setminus Y \to Y)\};$
16:          end
17:        end
18:      end
19:    end
20:  end
21: return R
22: end
```

Firstly, the ARNP algorithm receives as input a set of frequent patterns *F* from which the valid set of rules R is generated, and a minimum threshold $\min M_{GK}$. Then, the algorithm starts by initializing the set of valid rules (line 2). For each pattern of *X* and *Y* in *F*, then generated all valid rules (lines 3 to 20). For this, the algorithm first defines the measure M_{GK} (lines 5 and 6) to assess the set of rules. Then he begins by identifying the correlation area between *X* and *Y* units. If *X* and *Y* are positive dependency that is to say $M_{GK}(X \to Y) \geq 0$ (line 7), then the algorithm examines the rules type $X \to Y$ and $\overline{X} \to \overline{Y}$ (lines 8 to 10). Otherwise, if *X* and *Y* have a negative dependence *(X* disfavors *Y),* thus $M_{GK}(X \to Y) < 0$ (line 11), then successively algorithm evaluates the rules $X \to \overline{Y}$ and $\overline{X} \to Y$ (lines 12 to 17). The algorithm returns the set R which contains the valid rules (line 21) and stops when all the rules are generated.

V. EXPERIMENTAL EVALUATION

The objective of this section is to assess the feasibility of our approach. We evaluate the effectiveness of our algorithm in interesting us in terms of rules quality and extraction time.

Experimental protocol. We have implemented our algorithm in the R language. Our experiments were performed on a PC with 4 GB of RAM under Windows system. In this context, we conducted a series of experiments on four databases available on the UCI machine learning repository. The characteristics of these databases are shown in the table III.

TABLE III. CARACTERISTICS THE DATABASE EXPERIMENTS

Database	Number of transactions	Number of items
Adult	48842	115
German	1000	71
Income	6876	50
Iris	150	15

We tested our algorithm on four databases as indicated in the table III. For this, we set the threshold $minM_{GK}$ 60% and by varying the *minsup* (minimum threshold for measure "support") of 1% to 5%, the results obtained in below table IV. For each database, we restored the total execution time in seconds (*time column*) and the number of positive rules (*positive column*) and negative rules (*negative column*: rules rarely studied in the literature). The total number of positive and negative rules is summarized in the last column (*sum column*).

TABLE IV. RESULTS OF ALGORITHM

data	minsup.	time(s)	Number of valid rules		
			positive	negative	sum
Adult	1%	124	49925	75	50000
	2%	120	44925	75	45000
	3%	68	24929	71	25000
	4%	40	14941	59	15000
	5%	35	5382	55	5437
German	1%	65	48478	40	48518
	2%	45	44683	40	44723
	3%	22	11835	40	11875
	4%	15	7833	39	7872
	5%	10	4209	37	4246
Income	1%	9	2800	27	2827
	2%	7	2200	27	2227
	3%	6	1325	27	1352
	4%	4	1056	25	1083
	5%	3	553	25	578
Iris	1%	5	2437	59	2496
	2%	4	2000	59	2059
	3%	3	1200	59	1259
	4%	2.2	950	57	1007
	5%	2	500	57	557

Quality of obtained rules. Light of this table IV, on all four bases that we have studied; we find that our algorithm has concluded a number of association rules (positive and negative) is very reasonable. Overall, this situation is more significantly in the class of negative association rules. With a large enough data (*Adult*) and a minimum threshold for the relatively low support *(minsupp* = 1%), we only have 75 rules, a number very homogeneous, so reliable information to the user.

Execution time. The execution time of our algorithm for four games databases is presented numerically in the table IV (positive column and negative column) and graphically in the figure 1. We first note that the execution time changes linearly according to the number of rules whatever the base. We also note that the slope of the curves is determined by the number of attributes datasets. Indeed, the curve of the base *Adult* has the greatest slope, because it contains the largest number of attributes (115 attributes). It is followed by that in *German* with 71 attributes, then by that of *Income* with 50 attributes, and finally by the *Iris* with 15 attributes. Moreover, we see that a fairly minimal support low *(minsupp* = 1%), the highest time does not exceed 140 seconds even with a set whose size is around 75000 rules. This also shows that the effectiveness of our algorithm.

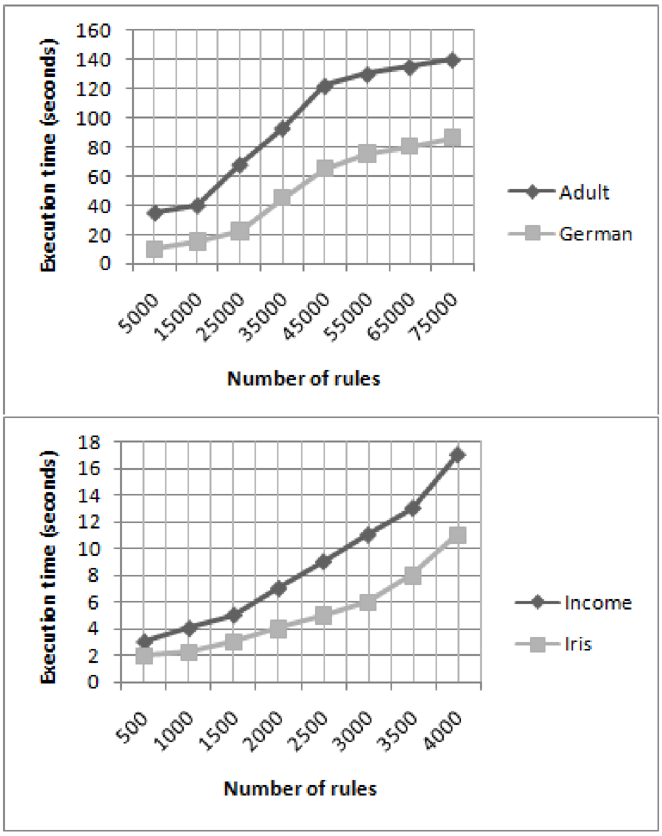

Figure 1. Execution time of the algorithm on four databases

More generally, these experiments demonstrate the feasibility of our approach that provides all valid association rules (positive and negative) within very reasonable size even if the set is large. In conclusion, our algorithm gives satisfactory results both on the quality of generated association rules on the execution time.

VI. Conclusion and future work

We have proposed a new algorithm that could contribute to generating the positive and negative association rules according the measure M_{GK}. Added to the extraction of association rules in attractive area, our approach optimizes the solving of problem of extraction rules in the repulsion area. The results of our experiments have shown the effectiveness of our approach that to reduce the number of rules in the very reasonable computing time. In the near future, we plan addressing two tracks: first, expanding our conclusion and future work of our algorithm to elaborate a tool for generating of an implicative graph under the implicative measure M_{GK}; Second, extending M_{GK} to extract implicative quantitative association rules vs qualitative association rules.

Acknowledgment

This work has been realized in the Laboratory of Mathematics and Computer Science at the « Normal School for Technical Education (ENSET) », of University of Antsiranana, (Madagascar), along with Laboratory of Computer Science and Mathematics (LIM), University of The Reunion (France). We would like to thank the AUF (University Agency of the Francophone), for having financed this Project Horizons Francophone: Fundamental Sciences - Computer Science and Mathematics.

References

[1] Agrawal R, T. Imielinski and A. N. Swami, “Mining association rules between sets of items in large databases”, *Proceedings of the ACM SIGMOD*, pp.207-216, Washington DC, 1993.

[2] Boulicaut J. F, Bykowski A and Jeudy B, “Towards the tractable discovery of association rules with negations”, *Conference on FQAS’00,* pp.425-434, 2000.

[3] Brin S, Motwani R, Silverstein C, “Bayond market baskets: Generalizing association rules to correlation”, *Proceedings of the ACM SIGMOD,* pp. 265-276, 1997.

[4] B. Ramasubbareddy, A. Govardhan, and A. Ramamohanreddy, “Classification based on positive and negative association rules”, *International Journal of Data Engineering (IJDE)*, Issue 2, pp.84–92, 2011.

[5] Cornelis C, P. Yan, X. Zhang and G. Chen, “Mining Positive and Negative Association Rules from Large Databases”, *Proceedings of the IEEE,* pp.613-618, 2006.

[6] Feno D, “Measure of quality the association rules: standardization and characterization of bases”, *University of the Reunion, France,* 2007, PhD thesis.

[7] Guillet F and H. Hamilton, “Quality measures in data mining”, *Springer-Verlag,* 2007.

[8] Guillaume S, “Processing of large data. Measure and extraction algorithms and ordinal association rules”, *University of Nantes, France, 2000, PhD thesis*.

[9] Missaoui R, L. Nourine and Y. Renaud, “Generating positive and negative exact rules using formal concept analysis: problems and solutions”, *ICFCA’08,* 2008.

[10] Nikky Rai, Susheel Jain and Anurag Jain, “Mining interesting positive and negative association rule based on improved genetic algorithm”, *Network and Complex Systems*, vol.3, pp.17–26, 2013.

[11] P. Narayana and D. Rakesh, “Mining positive and negative association rules using coherent approach”, *International Journal of Computer Trends and Technology*, vol.4, 2013.

[12] Savasere A, E. Omiecinki and S. Navathe, “Mining for strong negatives associations in a large database of customer transactions”, *Proceedings of the 14th ICDE’98,* pp.494-502, 1998.

[13] Sylvie Guillaume and P.-A. Papon, “Extraction optimisée de règles d’association positives et négatives (RAPN)”, RNTI, 2013.

[14] Totohasina A, H. Ralambondrainy and J. Diatta, “A unifying vision of quality measures Boolean association rules and an efficient algorithm for extracting association rules implicative”, *University of Antsiranana, Madagascar, 2005.*

[15] Totohasina A, “Contribution to the study of measures of quality of association rules: normalization and constraints in five cases and MGK, properties, composite base and extension rules for applying statistical and physical sciences”, *University of Antsiranana, Madagascar,* HDR, 2008.

[16] Tushar Mani, “Mining negative association rules”, *IOSR Journal of Computer Engineering (IOSRJCE)*, vol.3, Issue 6, pp.43-47, 2012.

[17] T. Ramakrishnudu and R. B. V. Sbramanyam, “Mining positive and negative association rules using fii-tree”, *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 4, pp.147–150, 2013.

[18] Xushan Peng, Ping Cheng and Maoji Wang, “A study of negative association rules mining algorithm based on multi-database”, *Proceedings of the 2012 2nd International Conference on Computer and Information Application (ICCIA)*, 2012.

[19] WU X, C. Zhang and S. Shang, “Mining booth Positive and Negative Association Rules”, *ICML-2002 & ACM TOIS 2004*, University of Vermont

[20] WU X, C. Zhang and S. Shang, “Efficient mining of both positive and negative association rules”, *ACM Transactions on Informations Systems (TOIS)*, vol.22, 2004.

[21] W. G. Teng, M. J. Hisieh, and M. S. Chen, “On the mining of substitution rules for statiscally dependent items”, *Second IEEE International Conference on Data Mining (ICDM’02),* pp.442–449, 2002.