EXPERIMENTAL STUDY OF THE VALID RULES ACCORDING TO THE MEASURE M_{GK}

Bruno Bakys RALAHADY and André TOTOHASINA

Department of Mathematics and Computer Science
Ecole Normale Supérieure Pour l'Enseignement Technique (ENSET), University of Antsiranana BP.0 – Madagascar.

ABSTRACT

Many Data Mining researchers are dedicated to researching and optimizing pattern extraction algorithms and / or valid rule generation algorithms. Most of these works condition the candidate patterns according to the minimum support threshold and filter the valid rules according to an arbitrary and subjective threshold of the confidence metric. In this paper, we have introduced another approach, an experimental study of generation of valid rules according to the interest measure M_{GK}, tested with a dataset of didactic research. Data Mining, critical value, interest measures, association rules

KEYWORDS

Data Mining, critical value, interest measures, association rule.

1 MOTIVATIONS AND INTRODUCTION

Extracting association rules is an important task in data mining. Numerous studies have already been carried out on the extraction of frequent patterns, rare patterns, frequent closed patterns, to optimize the treatment time. This work is based on the study of algorithmic complexity in order to make improvements on the processing time and the memory capacity used (Bastide and al. 2000) , (Pasquier and al.1999), (Liu and al.1999) and (Yun and al. 2003).

Concerning the extraction of motifs, the authors finish their writings by evoking tables or curves illustrating the memory capacity absorbed according to the size of the database and the number of patterns or set of frequent patterns generated as a function of minimum support thresholds. During our experiment, we selected the algorithms: APRIORI, Close, A-Close (Agrawal and Srikant.1994).

On the other hand, data mining processing continues on the valid association rules extraction algorithm. On the analysis of the rules of associations, several researchers frame their studies on the mathematical properties of measures (Guillaume 2000 and 2013), categorize and qualify measures in several classes (Feno 2007),.

In this report, we will present the results of the experiments we have carried out with an ASI-MGK tool, which we have developed, and which is able to generate valid rules based on 42 quality

DOI: 10.5121/ijcst.2019.01201

measures, whose valid association rules are filtered according to the critical value of the interest measures M$_{GK}$, obtained from the khi-square independence test at 1 degree of freedom at the risk threshold α.

1.1 Methodology

1.1.1 Binary context and notions of association rules

An association rule extraction context is a $\mathcal{B} = (\mathcal{O}, I, \mathcal{R})$ triplet in which $\mathcal{O}$ is called set of objects (or transactions), I set of attributes (or items), and $\mathcal{R} \subseteq \mathcal{O} \times I$ is a binary relation. An association rule is a pair $(X, Y) \in 2^I \times 2^I$ of patterns, noted: $X \to Y$, where X and Y are of disjoint motives ($X, Y \subseteq I$ and $X \cap Y = \emptyset$), respectively called premise and consequent of the rule. Note for all $X \subseteq I$, $X' = \{e \in \mathcal{O} /, e\mathcal{R}x, \forall x \in X\}$, called the extension of X. In what follows, let $n = |\mathcal{O}|$ and P be the uniform probability over $(\mathcal{O}, \mathcal{P}(\mathcal{O}))$, defined for all $X \subseteq I$, $X' \subseteq \mathcal{O}$ by: $\forall X \in \mathcal{P}(I), P(X) = \frac{cardX'}{n}$

Definition 1. (*Support*) Support for a $X \to Y$ association rule is the proportion of transactions in the database that contain $X \wedge Y$. (Agrawal and Swami. 1993)

$$supp(X \to Y) = P(X' \cap Y'). \tag{1}$$

A rule $r: X \to Y$ is valid according to support if $supp(X \to Y) \geq minsup$

Definition 1. (*Confidence*) The confidence of a $X \to Y$ association rule is the ratio of the number of transactions that contain $X \wedge Y$ to the number of transactions that contain X, (Agrawal and al.1993);

$$conf(X \to Y) = P(Y' | X') = \frac{P(X' \cap Y')}{P(X')}. \tag{2}$$

A rule $r: X \to Y$ is valid according to confidence if $conf(X \to Y) \geq minconf$, or $minconf$ $\in]0,1[$ arbitrarily set by the user.

A $r: X \to Y$ is valid on support-Confidence if $min(supp(X'), supp(Y')) \geq minsupp$ and $conf(X \to Y) \geq minconf$.

Definition 2. *(Lift)* The lift value of an association rule $X \to Y$ is (Brin and al. 1997):

$$lift(X \to Y) = \frac{P(X' \cap Y')}{P(X')P(Y')}. \tag{3}$$

A rule $r: X \to Y$ is valid according to lift if $lift(X \to Y) \geq minlift$, fixed $minlift$.

Lift is connected to confidence and support according to the relation below:

$$Lift(X \rightarrow Y) = \frac{conf(X \rightarrow Y)}{supp(Y)} = \frac{conf(Y \rightarrow X)}{supp(X)} = Lift(Y \rightarrow X). \tag{4}$$

Thus, Lift is a measure of symmetrical quality. This is one of the reasons why lift has been replaced by Conviction defined by (Brin and al. 1997);

$$Conv(X \rightarrow Y) = \frac{P(X' \cap \overline{Y'})}{P(X')P(\overline{Y'})}. \tag{5}$$

Conviction then has the dual advantage of being both asymmetrical and implicative, ie

$$Conv(\overline{Y} \rightarrow \overline{X}) = Conv(X \rightarrow Y), \forall X \rightarrow Y.$$

The validity of a rule according to the measures support, confidence and lift depends on a constant value previously chosen; the threshold is determined a priori by the user, this threshold does not take into account the nature of the data.

Definition 3. (*Valid rule*) A rule between two patterns is valid according to a measure m only depending on their contributions, ie

$r: X \rightarrow Y$ is valid if $m(X \rightarrow Y) > m(X, Y, \alpha)$, where α[1] and $m(X, Y, \alpha)$ is the critical value of m at seiul α, for two patterns X, Y.

Definition 4. (*Correlation coefficient*) Correlation measure of two patterns. (Lavrac end al. 1999).

$$\phi(X \rightarrow Y) = \frac{P(X' \cap Y') - P(X')P(Y')}{\sqrt{P(X')P(X')P(\overline{X'})P(\overline{Y'})}} \tag{6}$$

Its validity depends on the χ^2 and it determines a dependency between two patterns.

Definition 5. (M_{GK})Let X and Y be two reasons for a data mining context. We define the measure M_{GK} by:

$$M_{GK}(X \rightarrow Y) = \begin{cases} \frac{P(Y'|X') - P(Y')}{1 - P(Y')}, & if \;\; P(Y'|X') \geq P(Y') \\ \frac{P(Y'|X') - P(Y')}{P(Y')}, & if \;\; P(Y'|X') < P(Y') \end{cases} \tag{7}$$

Note: $M_{GK}^{f}(X \rightarrow Y) = \frac{P(Y'|X') - P(Y')}{1 - P(Y')}, if \; P(Y'|X') \geq P(Y')$ the favoring component which is responsible for pruning the rules.

[1] A statistical threshold of pruning calculated directly from the data.

Definition 6. (*Critical value M_{GKcr}*) Given a $\mathcal{B}$ database of n transactions, where n_X and n_Y are supports of the patterns X and Y respectively. The critical value $\mathrm{M_{GKcr}}$ for the measure M_{GK}, proposed in (Totohasina and Feno 2008), is obtained in the following way. Consider a contingency table by crossing these two patterns X and Y, using Pearson's chi-square (χ^2) statistic with one degree of freedom, such as:

$$\mathrm{M_{GKcr}}(X \rightarrow Y, \alpha) = \sqrt{\frac{1}{n}\frac{n-n_X}{n_X}\frac{n_Y}{n-n_Y}\chi_{cr}^2}, if \quad X \quad favors \quad Y. \tag{8}$$

By the above definition, a rule $r: X \rightarrow Y$ is valid according to $\mathrm{M_{GK}}$ if $\mathrm{M_{GK}}^f(X \rightarrow Y) \geq \mathrm{M_{GKcr}}(X \rightarrow Y, \alpha)$. .

Definition 1. (*Interesting rules*)

1. A rule $X \rightarrow Y$ is potentially interesting, if the support of its premise is inferior to that of its consequent.
2. If the $X \rightarrow Y$ rule is potentially interesting, then its $\overline{Y} \rightarrow \overline{X}$ will also be;
3. If the $X \rightarrow Y$ rule is potentially interesting, then the $Y \rightarrow X$ rule will no longer be interesting.
4. If the $X \rightarrow \overline{Y}$ rule is potentially interesting, then the $\overline{X} \rightarrow Y$ rule will no longer be interesting.

Proposition 1. Let X and Y be two patterns such that $P(Y|X) \geq P(Y)$ and $P(X) \leq P(Y)$, then $M_{GK}(Y \rightarrow X) \leq M_{GK}(X \rightarrow Y)$.

Proof. Since $P(Y|X) \geq P(Y)$, we have

$$M_{GK}(Y \rightarrow X) = \frac{P(X|Y)-P(X)}{1-P(X)} = \frac{\mathrm{P(X \cap Y)-P(X)P(Y)}}{\mathrm{P(Y)\big(1-P(X)\big)}} = \frac{\mathrm{P(X)}}{\mathrm{P(Y)}}\frac{\mathrm{\big(1-P(Y)\big)}}{\mathrm{\big(1-P(X)\big)}}\frac{\mathrm{P(X \cap Y)-P(X)P(Y)}}{\mathrm{P(X)\big(1-P(Y)\big)}}$$

$$= \frac{P(X)}{P(Y)}\frac{\big(1-P(Y)\big)}{\big(1-P(X)\big)} M_{GK}(X \rightarrow Y), \text{ hence } M_{GK}(Y \rightarrow X) = \frac{P(\overline{Y})}{P(Y)}\frac{P(X)}{P(\overline{X})} M_{GK}(X \rightarrow Y).$$

Now by hypothesis, $P(X) \leq P(Y)$, equals $P(\overline{X}) \geq P(\overline{Y})$, so $P(X)P(\overline{Y}) \leq P(\overline{X})P(Y)$, implies $M_{GK}(Y \rightarrow X) \leq M_{GK}(X \rightarrow Y)$.

1.1.2 Algorithm for extracting valid rules according to M_{GK}

This algorithm is based on Agrawal's $Apriori\ Gen - Rules$ algorithm (Agrawal and Srikant.1994) with some modification of the measure and the threshold used.

Algorithm $Apriori\ Gen - Rules$ using M_{GK}

```
input $\mathcal{F}, \alpha$   $frequents\ item$, risk error
output $\mathcal{R}$   set of association rules
begin
   $\mathcal{R} \leftarrow \emptyset$
   for all $k$-motif $l_k \in \mathcal{F}, k \geq 2$ do
       $H_1 \leftarrow 1 - frequents\ item \in l_k$
       for all $h_1 \in H_1$ do
           $\mathrm{M_{GKcr}} \leftarrow CalculMgkcr(H_1 \leftarrow H_1 - h_1, \alpha)$
           $\mathrm{M_{GK}} \leftarrow CalculMGK(H_1 \leftarrow H_1 - h_1)$
           if $\mathrm{M_{GK}} \geq \mathrm{M_{GKcr}}$ then
               $\mathcal{R} \leftarrow \mathcal{R} \cup \{r: l_k - h_1 \quad h_1\}$
           else
               $H_1 \leftarrow \{H_1 - h_1\}$
           end
       end
       $MGKGen - Rules(l_k, H_1)$
   end
   Return $\mathcal{R}$
end
```

Where $MGKGen - Rules$ is defined below;

Algorithm $MGKGen - Rules$

```
input $l_k, H_m, \alpha$
output $\mathcal{R}$        set of association rules
begin
    $\mathcal{R} \leftarrow \emptyset$
    for all $k$-motif $l_k \in \mathcal{F}, k \geq 2$ do
        $H_1 \leftarrow 1 - frequents\ item \in l_k$
        for all $h_{m+1} \in H_{m+1}$ do
            $\mathrm{M_{GKcr}} \leftarrow$CalculMgkcr($\mathrm{H_1 \leftarrow H_1 - h_{m+1}}, \alpha$)
            $\mathrm{M_{GK}} \leftarrow$CalculMGK($\mathrm{H_1 \leftarrow H_1 - h_{m+1}}$)
            if $\mathrm{M_{GK} \geq M_{GKcr}}$ then
                $\mathcal{R} \leftarrow \mathcal{R} \cup \{\mathrm{r: l_k - h_{m+1} \quad h_{m+1}}\}$
            Else
                $\mathrm{H_1 \leftarrow \{H_{m+1} - h_{m+1}\}}$
            end
        end
        $\mathrm{MGKGen - Rules(l_k, H_{m+1})}$
    end
    Return $\mathcal{R}$
end
```

2 Experimental evaluations

2.1 The numbers of rules extracted

In these two figures (fig 1) we showed the result of comparison of number of valid association rules based on the M_{GK} measure validated according to an error threshold.

- At the left, the number of valid rules according to M_{GK} is compared with the number of valid rules for $minconf = 0.6,\ 0.7$ and 0.8. M_{GK}-valid is comparable with confidance respectively for ($\alpha = 0.05,\ minconf = 0.6$), (0.01, 0.7) and (0.001, 0.8).
- At the right, it is compared with the valid rules according to Lift for $minlift = 1.0,\ 1.05,\ 1.15$ and 1.25

The corresponding alpha values with minlift are ($\alpha = 0.05,\ minlift = 1$), (0.05, 1.05), (0.05, 1) and (0.025, 1.25). $Lift$ is less selective than M_{GK}.

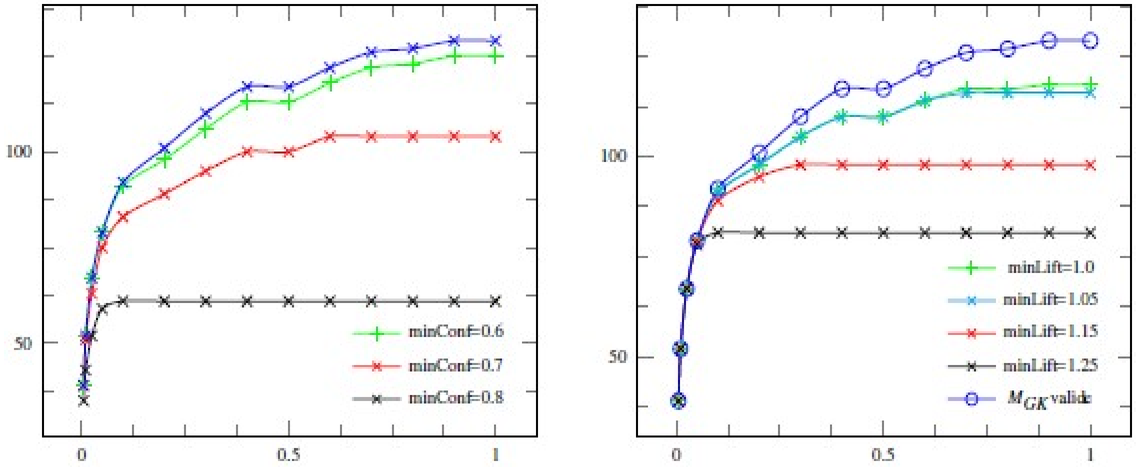

Figure 1: Variation of valid rules according to error thresholds

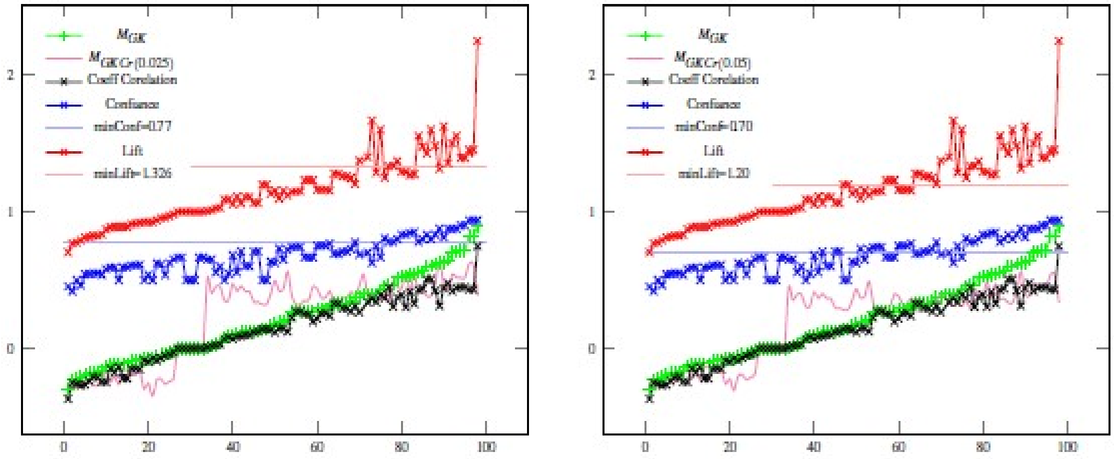

Figure 2: Variation in the number of rules extracted according to the thresholds used

2.2 The thresholds of validity

The figures (fig 2) show the 129 rules extracted from all common patterns greater than $minsup = 0.3$. On the right, the lift and confidence interest measures are compared to the function $\mathrm{M}_{\mathrm{GKcr}}(\alpha)$ for threshold $\alpha = 0.025$. The figure (fig 2) shows us that the rules extracted at this threshold are more pertinent and reliable than the extracted rules using $minlift = 1.32$ and $minconf = 0.77$.

Lift does not allow you to choose between $X \rightarrow Y$ and $Y \rightarrow X$; in the sense of Lift, the rules $X \rightarrow Y$ and $Y \rightarrow X$ are equivalent. The strong variation at the end of the curve indicates its sensitivity to the data size. The representative curve of M_{GK} shows its robustness, its small variation and its normality (its values which are in the interval $[-1,1]$).

Compared with linear correlation coefficient, these two measures obviously have values between $[-1,1]$, they can extract two types of rules; negative rules and positive rules. On the other hand, according to its properties, the linear correlation coefficient does not allow us to distinguish the rules $X \rightarrow Y$ and $Y \rightarrow X$. Which is not the case for the M_{GK} measure. Like lift, it is sensitive to data size. M_{GK} is also more discriminating. It seems that M_{GK} has a lot more interest in these measures.

2.3 Rule status extracted

According to the figure (fig 3), using Confidence with $minconf = 0.77$, Lift $minlif = 1.26$, correlation coefficient and $M_{GK}$when risk of error $\alpha = 0.05\%$, these four measures each generate 19 valid rules. By taking advantage of the dynamic visualization window of the implicit graph of the ASI-MGK tool, after some reorganization, we will analyze the graphs obtained. The graphs produced are oriented graphs of the same order (order 13), of different connected components ie the rules produced are not the same.

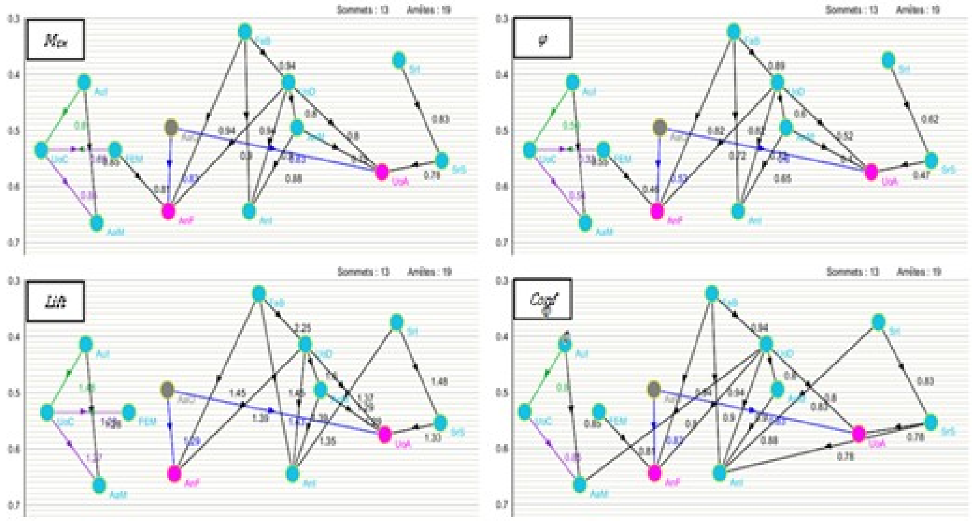

Figure 3: Implicative graph of the 19 extracted rules.

From the point of view, we see that M_{GK} and correlation coefficient produce the same graph.

On the other hand, Lift differs in a few rules; the absence of the rule $\{FEM\} \Rightarrow \{AnF\}$ and the appearance of the rule $\{SrI\} \Rightarrow \{AnI\}$ testify to this.

For that of confidence the difference is very remarkable; the disappearance of the rule $\{AnM\} \Rightarrow \{UoA\}$ and the equivalent rules: $\{UoC\} \Rightarrow \{FEM\}$ and $\{FEM\} \Rightarrow \{UoC\}$ and the rule appearances $\{UoD\} \Rightarrow \{AaM\}$, $\{SrI\} \Rightarrow \{AnI\}$ and $\{SrS\} \Rightarrow \{AnI\}$.

3 CONCLUSIONS

This new valid rule extraction approach introduced in this article is already integrated with ASI-MGK. We showed his ability to eliminate non-informative rules and his interest in graphical analysis of valid association rules using an implicative graph.

This comparative study proves that commonly used measures such as Confidence and Lift are sometimes misleading; hides the valid rule and displays unnecessary rules. We also find that there is a correspondence between these measures at a certain threshold, except that the pruning thresholds remain subjective for the approaches based on the Confidence and Lift measures.

So,why is it not also necessary to look for the critical functions of these measures?

As a futur work, like the critical values of the M_{GK} measure, we will consider developing an abacus of critical values on the other measures of qualities of the rules usually used such as lift and conviction, instead of using subjective thresholds.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami (1993) “Mining association rules between sets of items in large databases”, *In Proc. of the ACM SIGMOD Conference*, volume 22, pages 207-216, Washington,U.S.A.

[2] R. Agrawal and R. Srikant. (1994) “Fast algorithms for mining association rules.”, *In Proc. of the 20th VLDB Conference*, pages 487-499, San Diego, Chile.

[3] S. Brin, R. Motwani, and C. Silverstein (1997) “Beyond market baskets: Generalizing association rules to correlations. *In Proc. of the ACM SIGMOD Conference*, pages 265-276, Tucson, Arizona.

[4] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme and L. Lakhal (2000) “Mining frequent patterns with counting inference”, SIGKDD Explor. Newsl., 2(2):66–75.

[5] B. Liu, W. Hsu, and Y. Ma (1999) “Mining association rules with multiple minimum supports”, *In Proc. of SIGKDD’99 ACM Press*, pages 337–341, New York, NY, USA.

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

[7] J. Han, J. Pei, Y. Yin and R. Mao (2004) “Mining Frequent Patterns without Candidate Generation” *Data Mining and Knowledge Discovery*, vol. 8, pp.53-87.

[8] H. Yun, D. Ha, B. Hwang, and K. Ryu (2003) “Mining association rules on significant rare data using relative support”. *Journal of Systems and Software*, 67(3):181–191.

[9] N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal (1999) Discovering frequent closed itemsets for association rules. LNCS, 1540:398–416.

[10] S. Guillaume, (2000) "Processing of large data. Measure and extraction algorithms and ordinal association rules”, University of Nantes, France, PhD thesis.

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

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

[13] A. Totohasina and D. Feno (2008) "De la qualité des règles d'association: étude comparative des meures MGK et Confiance", *Actes du 9ème colloque Africain sur la recherche en Informatique et Mathématiques Appliquées*, CARI-08, 561-568

[14] A. Totohasina, (2008) "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 thesis.