



Un nouveau logiciel de traitement de données basé sur les graphes implicatifs

Josoa Michel Tovohery, Ralahady Bruno Bakys, Totohasina André, Daniel Rajaonasy

► To cite this version:

Josoa Michel Tovohery, Ralahady Bruno Bakys, Totohasina André, Daniel Rajaonasy. Un nouveau logiciel de traitement de données basé sur les graphes implicatifs. REVUT Scientific Journal, 2021. hal-03522299

HAL Id: hal-03522299

<https://hal.science/hal-03522299>

Submitted on 12 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Un nouveau logiciel de traitement de données basé sur les graphes implicatifs

TOVOHERY Josoa Michel¹, RALAHADY Bruno Bakys²,
TOTOHASINA André³ & FENO Daniel Rajaonasy⁴

¹Ecole Doctorale Thématique « Science, Culture, Société et Développement » de l'Université de TOAMASINA, josoamicheltovohery@gmail.com

²ENSET- Université d'Antsiranana, ralahadybru@yahoo.fr

³ENSET- Université d'Antsiranana, andre.totohasina@gmail.com

⁴Faculté de Droit, d'Économie, de Gestion, et de Mathématiques, Informatique et Applications
Université de TOAMASINA, fenodaniel2@yahoo.fr

Résumé :

Bien que la mesure M_{GK} soit potentiellement découverte depuis l'année 1975 sous l'appellation de « Certitude Factor », le premier logiciel de graphe implicatif basé sur la mesure M_{GK} n'était apparu qu'en 2017 sous l'appellation SDD-GI-MGK-TCP. C'est un logiciel de Science De Données et de Graphe Implicatif utilisant à la fois la mesure M_{GK} et la mesure TCP, et permettant de découvrir des connaissances souvent cachées, mais pertinentes et utiles de type « Si X, alors Y » ou « Si prémissé, alors conséquence » ou « Si cause, alors effet ». La mesure TCP a été inventée afin de trouver une interprétation mathématique de M_{GK} en termes de Taux de Croissance de Probabilité. La motivation d'usage de la mesure M_{GK} est le fait qu'elle est normalisée et implicative au sens de la logique formelle, ce qui s'avère plus cohérent à la fréquente interprétation causale des résultats obtenus. De plus, parmi les 47 mesures non normalisées existantes dans la littérature, les 30 mesures sont M_{GK} -normalisables, y compris « l'indice d'implication » du logiciel CHIC de Gras. C'est-à-dire, on obtient exactement l'expression de la mesure M_{GK} , après avoir normalisé ces 30 mesures. Ainsi, M_{GK} apparaît comme le noyau de la plupart des mesures des règles d'association existantes dans la littérature. Ensuite, nous montrons aussi que la mesure Confiance et Lift du logiciel TANAGRA de Ricco Rakotomalala ne sont pas meilleures que la mesure M_{GK} . Ainsi, ce papier présentera les principes de fonctionnement et le manuel d'utilisation de ce nouveau logiciel de traitement de données. Une comparaison avec le logiciel CHIC et Tanagra y sera également faite. À titre indicatif, des chercheurs et entrepreneurs ont déjà trouvé des succès et intérêts sur l'amélioration de leurs techniques de marketing en utilisant un outil informatique comme tel.

Mots clés : MGK, Graphe implicatif, CHIC, Tanagra, règle d'association

Abstract :

Although the M_{GK} measurement has been discovered since 1975, the first implicative graph software based on the M_{GK} measurement appeared only in 2017 under the name SDD-GI-MGK-TCP. It is a Data Science and an Implicative Graph Software using both MGK and TCP measurement to discover knowledge that is often hidden but relevant and useful, which can be interpreted as "If X, then Y" or "If premise, then consequence" or "If cause, then

conclusion". The TCP measurement was invented in order to find a mathematical interpretation of MGK in terms of Probability Growth Rate. The reason for using the MGK measure is the fact that it is standardized and implicative in the sense of formal logic, which justifies the causal interpretation of the results obtained. In addition, among the 47 non-standardized measures existing in the literature, the 30 measures are MGK- standardizable, including among the «implication index» measure of the Gras's CHIC software. That is to say, we obtain the exact expression of the MGK measurement, after having normalized these 30 measurements. Thus, MGK appears as the kernel of most measures of association rules existing in the literature. Next, we also point out that the Confidence and Lift measurement of Ricco Rakotomalala's TANAGRA software is not better than the MGK measurement. Thus, this paper will present the operating principles and the user manual of this new data processing software. A comparison with the CHIC and Tanagra software will be made. Researchers and entrepreneurs have already found success and interest in improving their marketing techniques by using a computer tool as such.

Key words : MGK, Implicative graphe, CHIC, Tanagra, Association rules

1 Introduction

La mesure M_{GK} a été découverte indépendamment par quatre groupes de chercheurs dans trois continents, et a eu subi quatre nominations différentes. Elle a été potentiellement découverte, pour la première fois, aux Etats-Unis, par Shortliffe et Buchman en 1975 sous l'appellation « Measure of Increased Belief (or Disbelief) » ou « Certitude Factor » en médecine (Shortliffe & Buchanan, 1975). En Europe, la dénomination « M_{GK} » a été proposé par Guillaume en l'an 2000 (Guillaume, 2000). À Madagascar, donc en Afrique, la même mesure fut découverte indépendamment par Totohasina en 2003 sous la dénomination « ION » (Indice d'implication Orienté et Normalisé), enfin par Wu et ses collaborateurs en 2004 sous l'appellation « CPIR » (Conditional Probability Increment Ratio) aux Etats – Unis (Wu et al. 2004).

Des algorithmes d'extraction des règles d'association valides et des bases des règles d'association valides selon la mesure M_{GK} ont été élaborés dans les travaux de thèses suivants : (Feno, 2007), (Bemarisika, 2016), (Ramanantsoa, 2016) et (Rakotomalala, 2019).

La plus grande motivation de l'usage de la mesure MGK est, d'abord, le fait qu'elle est implicative et normalisée (propriété utile pour qu'un graphe soit implicatif), après le fait que la plupart des mesures non normalisées sont M_{GK} -normalisables, c'est-à-dire on obtient l'expression de la mesure MGK après avoir normalisé ces mesures. Nous signalons qu'une liste de 30 mesures M_{GK} -normalisables parmi les 61 mesures recensées dans (Grissa, 2004) est offerte dans (Armand, 2019).

Bien que la découverte potentielle de la mesure M_{GK} date de l'année 1975, le premier logiciel de graphe implicatif basé sur M_{GK} était apparu en juin 2017. C'était le logiciel SDD-GI-MCK-TCP, élaboré dans le laboratoire des mathématiques de l'ENSET de l'université d'Antsiranana par Tovohery et Totohasina (Tovohery, 2017). Juste après, Ralahady Bruno

Bakys a développé le logiciel ASI-MGK (Ralahady & Totohasina, 2019). Ainsi, ce papier a pour but de partager les principes et le manuel d'utilisation du logiciel SDD-GI-MGK-TCP.

La suite de ce papier est organisée comme suit : la section 2 présentera la mesure M_{GK} , son interprétation et le manuel d'utilisation de ce nouveau logiciel ; la section 3 exposera un exemple de résultat de traitement effectué un jeu de données réelles ; la section 4 discutera des intérêts et limites de notre nouveau logiciel face aux logiciels CHIC et Tanagra ; enfin, la section 5 terminera ce papier par une conclusion et nos perspectives de recherches.

1. Matériels et méthodes

Tout d'abord, le terme SDD – GI – MGK – TCP signifie « Science De Donné – Graphe Implicatif – MGK – TCP ». Donc, notre nouveau logiciel est conçu pour extraire des règles d'association qualitatives selon la mesure MGK et TCP. Ainsi, il semble très intéressant de parler, d'abord, des règles d'association, de ces deux mesures de qualité des règles d'association et du graphe implicatif avant de passer au manuel d'utilisation de notre nouveau logiciel en question.

1.1 Règles d'association

Le concept des règles d'association a trouvé sa genèse dans le domaine de marketing (Agrawal et al., 1993). Le principe se résume comme suit : soit D l'ensemble de toutes les transactions effectuées (ou tous les paniers des clients). Une transaction $t \in D$ est un ensemble d'articles (en anglais « itemset ») achetés par un client. Soient I l'ensemble de tous les articles vendus dans le supermarché et X une partie de I . X est appelé un itemset, ou un ensemble d'articles, ou encore un motif. Une transaction $t \in D$ contient X si t contient tous les articles (items) dans X . Selon la définition donnée par R. Agrawal et al., on appelle règle d'association une implication partielle de type $X \Rightarrow Y$, où X et Y sont deux itemsets, tels que $X \cap Y = \emptyset$. Le *Support* d'un motif X est le nombre de transactions qui contiennent tous les articles dans X divisé par le nombre total de transactions dans D . Le *support* « s » de la règle $X \Rightarrow Y$ est le nombre des transactions dans D qui contiennent à la fois X et Y divisé par le nombre total de transactions dans D . La *Confiance* « c » de la règle $X \Rightarrow Y$ est le support de la règle $X \Rightarrow Y$ divisé par le support de X . On dit : « si X , alors Y », avec une confiance de $c \times 100\%$. Le *Support* et la *Confiance* sont souvent exprimés en pourcentage.

1.2 Mesure M_{GK}

Définition 1. On définit la mesure de qualité de règle d'association M_{GK} par :

$$M_{GK}(X \Rightarrow Y) = \begin{cases} M_{GK}^f(X \Rightarrow Y) = \frac{P_{X'}(Y') - P(Y')}{1 - P(Y')}, & \text{si } P_{X'}(Y') > P(Y'); \\ M_{GK}^d(X \Rightarrow Y) = \frac{P_{X'}(Y') - P(Y')}{P(Y')}, & \text{si } P_{X'}(Y') \leq P(Y'). \end{cases} \quad (1)$$

où X' et Y' sont respectivement les intensions des motifs X et Y . On a : $P(X') = \frac{n_X}{n}$, $P(Y') = \frac{n_Y}{n}$ et $P_{X'}(Y') = \frac{n_{XY}}{n}$, avec n_X , n_Y et n_{XY} sont respectivement le nombre de

transactions contenant X , le nombre de transactions contenant Y et le nombre de transaction contenant à la fois X et Y , tandis que n est le nombre total de toutes les transactions.

Définition 2. On appelle intension du motif X , l'ensemble $T \subset D$ de toutes les transactions t contenant le motif X .

1.2.1 Propriétés des composantes de la mesure de qualité M_{GK}

- 1) M_{GK}^f est implicative : pour tous motifs X et Y , $M_{GK}^f(X \rightarrow Y) = M_{GK}^f(\bar{Y} \rightarrow \bar{X})$;
- 2) M_{GK}^f est non symétrique : pour tous motifs X et Y , $M_{GK}^f(X \rightarrow Y) \neq M_{GK}^f(Y \rightarrow X)$;
- 3) M_{GK}^f est normalisé : pour tous motifs X et Y , $0 \leq M_{GK}(X \rightarrow Y) \leq 1$.
- 4) M_{GK}^d est symétrique, non implicatif et non normalisé.

Ainsi, nous décidons d'utiliser seulement la composante M_{GK}^f quel que soit le cas. Dans le cas où $P_{X'}(Y') \leq P(Y')$, on travaille avec la règle négative $\bar{X} \rightarrow Y$.

1.2.2 Interprétations

- 1) Si $P_{X'}(Y') > P(Y')$ et $M_{GK}^f(Y \rightarrow X) < M_{GK}^f(X \rightarrow Y)$, alors on dit que X est le plus à favoriser Y que Y favorise X . Dans ce cas, si on veut Y , alors il faut favoriser X .
- 2) Si $P_{X'}(Y') > P(Y')$ et $M_{GK}^f(Y \rightarrow X) = M_{GK}^f(X \rightarrow Y)$, alors on dit que X et Y se favorisent mutuellement.
- 3) Si $P_{X'}(Y') \leq P(Y')$ et $M_{GK}^f(Y \rightarrow \bar{X}) < M_{GK}^f(\bar{X} \rightarrow Y)$, alors on dit que X est le plus à défavoriser Y , que Y défavorise X . Dans ce cas, si on veut éliminer Y , alors il faut favoriser X .
- 4) Si $P_{X'}(Y') \leq P(Y')$ et $M_{GK}^f(Y \rightarrow \bar{X}) = M_{GK}^f(\bar{X} \rightarrow Y)$, alors on dit que X et Y se défavorisent mutuellement.

1.2.3 Validation des Règles selon la mesure de qualité M_{GK}

La relation entre la mesure M_{GK} et la valeur critique de la loi de Khi – deux à 1 degré de liberté χ_1^2 au seuil critique α est définie par :

$$M_{GK}^f\text{-Critique}(X \Rightarrow Y, \alpha) = \sqrt{\frac{1}{n} \times \frac{n - n_X}{n_X} \times \frac{n_Y}{n - n_Y} \times \chi_1^2(\alpha)} \quad (2)$$

Comme les valeurs critiques de Khi – deux sont tabulées pour tout seuil critique α , alors il est ainsi facile de calculer la valeur critique de la mesure M_{GK} .

Prise de décision : La règle $X \Rightarrow Y$ est validée selon la mesure M_{GK} au niveau de confiance $(1 - \alpha)100\%$, si et seulement si $M_{GK}^f\text{-Critique}(X \Rightarrow Y, \alpha) < M_{GK}^f(X \Rightarrow Y)$.

Remarques : on a une quasi – implication selon la mesure M_{GK} si les règles $X \Rightarrow Y$ et $Y \Rightarrow X$ sont toutes validées selon la mesure M_{GK} . De plus, si on a une quasi – implication et $M_{GK}^f(Y \Rightarrow X) < M_{GK}^f(X \Rightarrow Y)$, alors on retient seulement la règle $X \Rightarrow Y$ et on rejette la règle $Y \Rightarrow X$, pour éviter la redondance et pour simplifier la prise de décision.

1.3 Mesure TCP

La mesure de qualité Taux de Croissance de Probabilité (TCP) a été adoptée afin de trouver une interprétation mathématique de la mesure M_{GK} en termes de taux de croissance.

1.3.1 Taux de croissance

Définition 3. D'après (Mazerolle, 2005), le taux de croissance est défini par :

$$\text{Taux de croissance} = \frac{\text{Valeur d'arrivée}}{\text{Valeur de départ}} - 1 \quad (3)$$

Ainsi, on a les trois interprétations possibles suivantes :

- a) Si Taux de croissance > 0 , alors il y a une croissance;
- b) Si Taux de croissance < 0 , alors il y a une décroissance;
- c) Si Taux de croissance $= 0$, alors il y a une constance (c'est le cas statuquo).

1.3.2 Taux de croissance de probabilité

On considère les Figures 1 et 2.

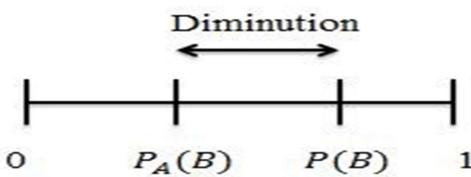


Figure 1 : Cas défavorable

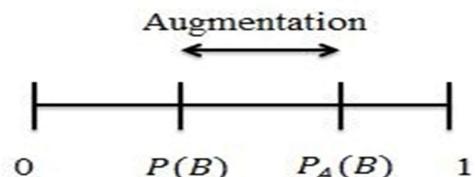


Figure 2 : Cas favorable

On considère une transaction t . Soient A l'événement d'avoir le motif X dans t et B l'événement d'avoir le motif Y dans t .

La Figure 2 montre bien que la probabilité pour que B soit réalisé augmente lors que A soit réalisé. Donc, on a $P_A(B) > P(B)$, ce qui donne l'interprétation : les deux événements A et B se favorisent l'un et l'autre. Alors, si l'on pose, à l'instar de la définition du taux de croissance, « Valeur de départ » = $P(A)$ et « valeur d'arrivée » = $P_B(A)$, alors on pose la définition suivante.

Définition 4. On définit le « taux de croissance de la probabilité » de l'événement B lors que l'événement A soit réalisé par :

$$\text{TCP}(A \Rightarrow B) = \frac{P_A(B)}{P(A)} - 1 \quad (4)$$

Interprétation :

- a) Si $\text{TCP}(A \Rightarrow B) > 0$, alors A favorise B.
- b) Si $\text{TCP}(A \Rightarrow B) < 0$, alors A défavorise B.
- c) Si $\text{TCP}(A \Rightarrow B) = 0$, alors A et B sont indépendants.

Remarques :

- a) Comme TCP est la différence entre le rapport $\frac{P_A(B)}{P(A)}$ et 1, donc il nous donne le comportement de $\frac{P_A(B)}{P(A)}$ vis – à – vis de 1. Ainsi, la mesure TCP mesure la dépendance entre deux événements aléatoires.
- b) On note que $\text{TCP}(A \Rightarrow B) = \text{TCP}(X \Rightarrow Y)$.

1.3.3 Propriétés de la mesure de qualité TCP

- a) TCP est symétrique : $\text{TCP}(X \Rightarrow Y) = \text{TCP}(Y \Rightarrow X)$;
- b) TCP n'est pas implicatif : $\text{TCP}(X \Rightarrow Y) = \text{TCP}(\bar{Y} \Rightarrow \bar{X})$;
- c) TCP n'est pas normalisé : $\text{TCP}(X \Rightarrow Y) \in [-1, +\infty[$.

1.4 Relation entre les mesures de qualité M_{GK} et TCP

Proposition 1. Pour deux motifs X et Y , on a :

- a) $M_{GK}^d(X \Rightarrow Y) = \text{TCP}(X \Rightarrow Y)$ si $P_{X'}(Y') \leq P(Y')$;
- b) $M_{GK}^f(X \Rightarrow Y) = \text{TCP}(X \Rightarrow Y) \times \frac{P(Y')}{1-P(Y')}$, si $P_{X'}(Y') > P(Y')$.

Remarque : Comme $M_{GK}^f(X \Rightarrow Y) \in [0 ; 1]$ et $\text{TCP}(X \Rightarrow Y) \in [-1 ; +\infty[$, alors, d'après l'expression (b), le rapport $\frac{P(Y')}{1-P(Y')}$ est le facteur de normalisation du taux de croissance TCP .

Proposition 2. Si l'on pose $k(P(Y')) = \frac{P(Y')}{1-P(Y')}$, alors on a :

$$M_{GK}(X \Rightarrow Y) = \begin{cases} \text{TCP}(X \Rightarrow Y) \times k(P(Y')), \text{ si } P_{X'}(Y') > P(Y') ; \\ \text{TCP}(X \Rightarrow Y) \text{ si } P_{X'}(Y') \leq P(Y'). \end{cases} \quad (5)$$

Ainsi, la quantité $M_{GK}(X \Rightarrow Y)$ apparaît comme un taux de croissance normalisé et orienté de la probabilité de l'événement B (*i.e.* $Y \in t$) sous la réalisation de A (*i.e.* $X \in t$). D'où la proposition 3 suivante.

Proposition 3. Pour deux motifs X et Y , on a :

- a) $P_{X'}(Y') = P(Y') - P(Y') \times \text{TCP}(X \Rightarrow Y) = P(Y') - P(Y') \times M_{GK}(X \Rightarrow Y)$
si $P_{X'}(Y') \leq P(Y')$; c'est – à – dire, si X est contenu dans une transaction t , alors la probabilité pour que Y soit dans t diminue de $P(Y') \times M_{GK}(X \Rightarrow Y)$.
- b) $P_{X'}(Y') = P(Y') + P(Y') \times \text{TCP}(X \Rightarrow Y) = P(Y') + P(Y') \frac{M_{GK}(X \Rightarrow Y)}{k(P(Y'))}$
 $= P(Y') + (1 - P(Y')) \times M_{GK}(X \Rightarrow Y)$ si $P_{X'}(Y') > P(Y')$; c'est – à – dire, si X est contenu dans une transaction t , alors la probabilité pour que Y soit dans t augmente de $(1 - P(Y')) \times M_{GK}(X \Rightarrow Y)$.

Proposition 4. On définit la valeur critique de la mesure TCP au risque α fixé pour deux motifs X et Y par :

$$\text{TCP_Critique}(X \Rightarrow Y, \alpha) = \sqrt{\frac{1}{n} \times \frac{n - n_X}{n_X} \times \frac{n - n_Y}{n_Y} \times \chi_1^2(\alpha)} \quad (6)$$

Proposition 5. La mesure TCP mesure la dépendance entre deux motifs X et Y . Ainsi, on adopte la règle de décision suivante :

- 1) Pour le cas favorable ($P_{X'}(Y') > P(Y')$), on a :
 - a) si $\text{TCP}(X \Rightarrow Y) > \text{TCP_Critique}(X \Rightarrow Y, \alpha)$, alors la dépendance positive est significative selon la mesure TCP au niveau de confiance $(1 - \alpha)100\%$;
 - b) sinon, la dépendance positive n'est pas significative selon TCP au niveau de confiance $(1 - \alpha)100\%$.
- 2) Pour le cas défavorable ($P_{X'}(Y') \leq P(Y')$), on a :
 - a) Si $\text{TCP}(X \Rightarrow \bar{Y}) > \text{TCP_Critique}(X \Rightarrow \bar{Y}, \alpha)$, alors la dépendance négative est significative selon la mesure TCP au niveau de confiance $(1 - \alpha)100\%$;
 - b) sinon, la dépendance négative n'est pas significative selon TCP au niveau de confiance $(1 - \alpha)100\%$.

Proposition 6. Si la dépendance entre deux motifs X et Y est significative selon la mesure TCP, alors on a une quasi – implication selon la mesure M_{GK}^f , c'est – à – dire les deux règles $X \Rightarrow Y$ et $Y \Rightarrow X$ sont toutes validées selon la mesure M_{GK}^f . La réciproque est aussi vraie.

Preuve. (1) En effet, la dépendance entre deux motifs X et Y est significative au niveau de confiance $(1 - \alpha)100\%$ si $\text{TCP}(X \Rightarrow Y) > \text{TCP_Critique}(X \Rightarrow Y, \alpha)$. Comme $0 < n_Y < n$, donc $\frac{n_Y}{n - n_Y} > 0$. Donc, si la dépendance est significative au niveau de confiance $(1 - \alpha)100\%$, alors on a : $\text{TCP}(X \Rightarrow Y) \times \frac{n_Y}{n - n_Y} > \text{TCP_Critique}(X \Rightarrow Y, \alpha) \times \frac{n_Y}{n - n_Y}$.

Cependant, on a :

- a) $M_{GK}^f(X \Rightarrow Y) = \text{TCP}(X \Rightarrow Y) \times \frac{n_Y}{n - n_Y}$ et
- b) $M_{GK\text{-Critique}}^f(X \Rightarrow Y, \alpha) = \text{TCP_Critique}(X \Rightarrow Y, \alpha) \times \frac{n_Y}{n - n_Y}$.

Ainsi, on obtient $M_{GK}^f(X \Rightarrow Y) > M_{GK\text{-Critique}}^f(X \Rightarrow Y, \alpha)$. En exploitant la symétrie de la mesure TCP, on a aussi $M_{GK}^f(Y \Rightarrow X) > M_{GK\text{-Critique}}^f(Y \Rightarrow X, \alpha)$. D'où la quasi – implication selon M_{GK}^f .

(2) Supposons qu'on a une quasi – implication selon la mesure M_{GK}^f :

- a) $M_{GK}^f(X \Rightarrow Y) > M_{GK\text{-Critique}}^f(X \Rightarrow Y, \alpha)$ et
- b) $M_{GK}^f(Y \Rightarrow X) > M_{GK\text{-Critique}}^f(Y \Rightarrow X, \alpha)$.

En divisant (a) par la quantité positive $\frac{n_Y}{n - n_Y}$ et (b) par $\frac{n_X}{n - n_X}$, alors on a :

$\text{TCP}(X \Rightarrow Y) > \text{TCP_Critique}(X \Rightarrow Y, \alpha)$ et $\text{TCP}(Y \Rightarrow X) > \text{TCP_Critique}(Y \Rightarrow X, \alpha)$. Or,

on a : $\text{TCP}(X \Rightarrow Y) = \text{TCP}(Y \Rightarrow X)$. Ainsi, on a une dépendance significative selon la mesure TCP. D'où la proposition 6.

Ainsi, la mesure TCP est supposée comme le noyau principal de la mesure de qualité M_{GK} . Donc, il est confirmé que M_{GK} incarne, entre autre, la sémantique du taux de croissance orienté et normalisé.

1.5 Mesure Intensité d'Implication

La mesure intensité d'implication est la mesure de qualité des règles d'association utilisée dans le logiciel CHIC de R. Gras et al. Ainsi, il s'avère bien d'en parler.

1.5.1 Indice d'implication

Définition 5. On définit l'indice d'implication entre deux motifs X et Y par :

$$q(X, \bar{Y}) = \frac{n_{X\bar{Y}} - \frac{n_X \cdot n_{\bar{Y}}}{n}}{\sqrt{\frac{n_X \cdot n_{\bar{Y}}}{n}}} \quad (7)$$

où $n_{X\bar{Y}}$ désigne le nombre des transactions contenant X et ne contenant pas Y , et $n_{\bar{Y}} = n - n_Y$ est le nombre des transactions ne contenant pas de Y .

D'après (Gras, 2004), la quantité $q(X, \bar{Y})$ est retenu comme indicateur de la non – implication de X sur Y . Plus $q(X, \bar{Y})$ est grand, plus l'implication $X \Rightarrow Y$ est douteuse. Ensuite, R. Gras a noté que la variable aléatoire $q(X, \bar{Y})$ suit approximativement la loi normale pour $\frac{n_X \cdot n_{\bar{Y}}}{n} \geq 3$. De plus, nous signalons que la mesure $q(X, \bar{Y})$ est M_{GK} -normalisable (Armand, 2019, p. 134).

1.5.2 Intensité d'implication

Définition 6. On définit la mesure Intensité d'Implication (I.I) du motif X sur le motif Y par :

$$I.I(X \Rightarrow Y) = \frac{1}{\sqrt{2\pi}} \int_{q(X, \bar{Y})}^{+\infty} e^{-t^2/2} dt \quad (8)$$

Prise de décision : l'implication $X \Rightarrow Y$ est validée au niveau de confiance $(1 - \alpha)100\%$ si et seulement si $I.I(X \Rightarrow Y) \geq 1 - \alpha$

Interprétation : $I.I(X \Rightarrow Y)$ mesure « l'étonnement de constater la petitesse des contre-exemples en regard du nombre surprenant des instances de 'l'implication » (Gras, 2004).

Remarque : si $n_Y \rightarrow n$, alors $I.I(X \Rightarrow Y) \rightarrow 0$ (Gras, 2004). C'est-à-dire, si l'implication $X \Rightarrow Y$ tend à être trivial, alors $I.I(X \Rightarrow Y)$ devient petit.

1.6 Mesure Lift et Confiance

La mesure Lift et la mesure confiance sont les mesures de qualité des règles d'association utilisées dans le logiciel Tanagra de R. Rakotomalala.

Définition 7. On définit la Confiance d'une règle d'association $X \Rightarrow Y$ par :

$$\text{Confiance}(X \Rightarrow Y) = \frac{P(X' \cap Y')}{P(X')} = P_{X'}(Y') \quad (9)$$

Interprétation : Confiance($X \Rightarrow Y$) = la probabilité conditionnelle sachant X de Y .

Définition 8. On définit l'intérêt (Lift) d'une règle d'association $X \Rightarrow Y$ par :

$$\text{Lift}(X \Rightarrow Y) = \frac{P_{X'}(Y')}{P(Y')} = \frac{P(X' \cap Y')}{P(X')P(Y')} \quad (10)$$

1.7 Graphe implicatif

Selon (Couturier & Gras, 2005, p. 681), un graphe implicatif est « un graphe sur lequel les variables (règles d'association) qui possèdent une intensité d'implication supérieure à un certain seuil sont reliées par une flèche représentant l'implication ». Ensuite, « un graphe implicatif traduit graphiquement l'ensemble du réseau des relations quasi-implicatives entre les variables » (Couturier et al, 2003). La Figure 3 suivante montre un exemple d'un graphe implicatif issu d'un traitement avec le logiciel CHIC.

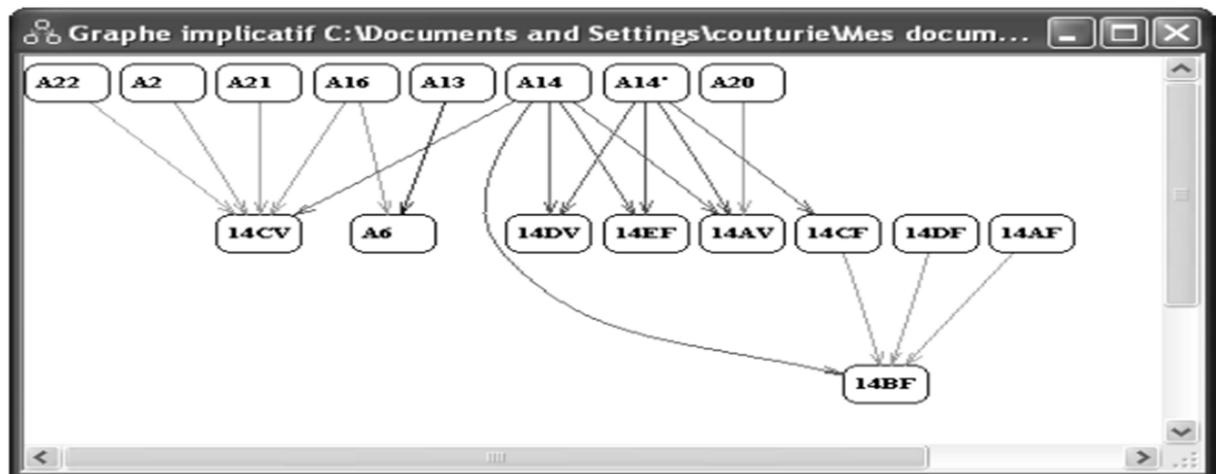


Figure 3 : Un exemple d'un graphe implicatif issu du logiciel CHI.
Source : (Couturier & Gras, 2005, p. 682)

Afin de mathématiser cette définition, nous proposons la définition suivante.

1.7.1 Graphe implicatif

Définition 9. Un graphe implicatif G est la donnée d'un triplet $G = (M ; I ; v)$ tels que :

- a) M est un ensemble fini des motifs,
- b) I est un ensemble de couples ordonnés des motifs deux à deux distincts $(m_i ; m_j) \in M^2$,
- c) v est une mesure de qualité des règles d'association non symétrique (valuation).

Donc, un graphe implicatif G est un graphe orienté élémentaire et valué, dont la valuation est une mesure de qualité des règles d'association non symétrique. Le couple $(m_i ; m_j)$ désigne la règle d'association $m_i \Rightarrow m_j$.

1.7.2 Comment dessiner un graphe implicatif ?

Considérons un espace probabilisé (Ω, τ, P) . La propriété clé de notre approche est la suivante : pour tous événements A et B de τ tels que $P_A(B) > P(B)$:

- a) Si $P(A) < P(B)$, alors on retient la règle $A \Rightarrow B$ (car $M_{GK}^f(A \Rightarrow B) > M_{GK}^f(B \Rightarrow A)$);
- b) Si $P(A) = P(B)$, alors on retient la règle $A \Leftrightarrow B$ (car $M_{GK}^f(A \Rightarrow B) > M_{GK}^f(B \Rightarrow A)$).

Remarque : Ces deux propriétés sont aussi admises pour la mesure Confiance.

Ainsi, notre approche consiste à dessiner les motifs en cascade : celui qui a le support le plus faible est en haut et l'autre qui a le support plus grand en bas. Après, on les lie par une flèche selon leurs liaisons (voir Figure 8).

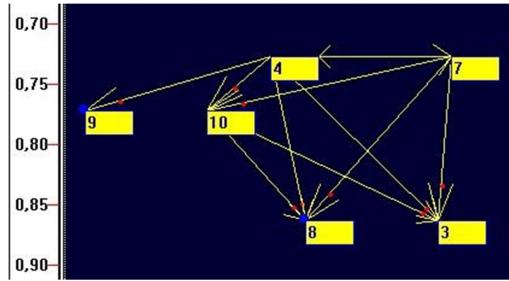


Figure 8 : Exemple de résultat de notre approche

1.8 Présentation du logiciel SDD-GI-MGK-TCP

Le logiciel SDD-GI-MGK-TCP est un logiciel permettant de faire une extraction de connaissances à partir d'une base de données binaires. Ce logiciel utilise l'algorithme Apriori (Agrawal et al., 1993) pour sélectionner les motifs fréquents de la base de données. Après, pour extraire les règles d'association valides, le logiciel laisse deux choix de mesure à l'utilisateur, telles que la mesure Confiance, qui n'est autre que probabilité conditionnelle et la mesure M_{GK}^f . Enfin, il retourne un résultat sous forme d'un graphe implicatif avec un tableau récapitulatif et les étapes d'exécution effectuées pour aider l'utilisateur. La mesure TCP est utilisée pour signaler (ou juger) la pertinence des règles d'association valides par ces deux mesures.

1.8.1 Interface utilisateurs

L'interface utilisateur de ce logiciel est composé de onze objets et d'un menu, tels que quatre champs de saisie, un combo, cinq boutons et un tableau (voir Figure 9).

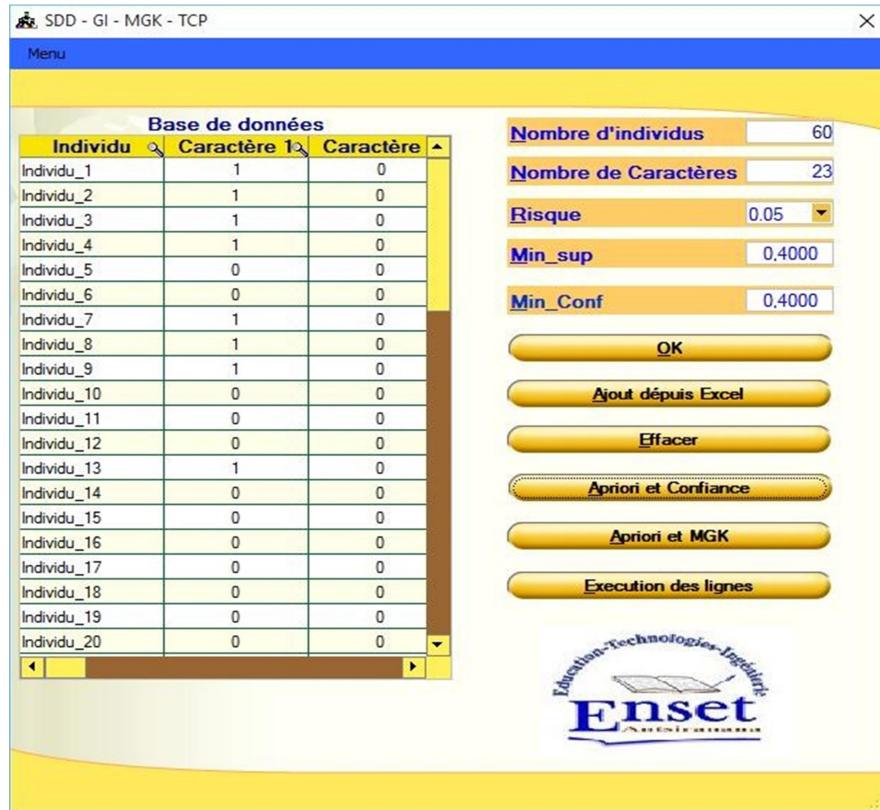


Figure 9 : Interface utilisateurs du logiciel SDD-GI-MGK-TCP

Tout d'abord, nous signalons que l'entrée des données à traiter avec le logiciel SDD-GI-MGK-TCP peut se faire en deux choix (options) :

- 1) Saisie directe des données : on saisit directement les données à traiter sur le tableau «Base de donnée» ;
- 2) Ajout depuis Excel : on charge un fichier Excel format .xls (Classeur Excel 97- 2003) contenant le jeu de données à traiter.

Maintenant, nous allons décrire les rôles des objets de notre interface utilisateur :

Nombre d'individus : Lors de l'application de la première option, le champ de saisie « Nombre d'individus » permet à l'utilisateur de saisir le nombre de lignes de la table des données à traiter, c'est-à-dire le nombre de toutes les transactions ou le nombre de tous les individus enquêtés.

Nombre de caractères : Lors de l'application de la première option, le champ de saisie « Nombre de caractères » permet à l'utilisateur d'introduire dans le logiciel le nombre de colonnes des données, c'est-à-dire le nombre des motifs ou items à étudier sur chaque transactions effectuées.

Le combo Risque : Le combo (ou liste déroulante) « Risque » permet à l'utilisateur de fixer le seuil d'erreur α effectué par le logiciel lors de la filtration des règles pertinentes à l'aide des mesures M_{GK}^f et TCP. Ainsi, toutes les règles affichées sont valide au niveau de confiance $(1 - \alpha)100\%$. La valeur par défaut du combo Risque est 0.05.

Min_supp : Le champ de saisie « Min_supp » permet à l'utilisateur de fixer le support minimum des motifs étudiés. Cela est nécessaire lors de la sélection des motifs fréquents par l'algorithme Apriori. Sa valeur par défaut est 0.4.

Le bouton OK : Après avoir fixé le nombre d'individus n et le nombre m des items à étudier, un clic à ce bouton permet à l'utilisateur d'avoir un tableau vide de n lignes et de m colonnes. De plus, c'est le seul bouton permettant à l'utilisateur d'écrire sur le tableau « Base de données ».

Le bouton Ajout depuis Excel : Ce bouton permet à l'utilisateur d'importer dans SDD-GI-MGK-TCP un fichier de format .xls (Classeur Excel 97- 2003) contenant les données à traiter.

Le bouton Effacer : Ce bouton permet à l'utilisateur d'effacer le tableau « Base de données » et de réinitialiser le logiciel.

Le bouton Apriori et Confiance : Un clic à ce bouton permet à l'utilisateur d'exécuter l'algorithme Apriori et l'algorithme « Association Rule Generation » selon (Wu & Kumar, 2009, p. 65) mais modifié. La modification apportée est sur le sens d'implication, tel que $\forall A, B \in (\Omega, \tau, P)$, si les deux règles $A \Rightarrow B$ et $B \Rightarrow A$ sont valides, alors si $P(A) < P(B)$, alors on retient seulement la règle $A \Rightarrow B$. Toutefois, si $P(A) = P(B)$, alors on retient la règle $A \Leftrightarrow B$. Cette modification a été faite afin de diminuer la redondance des règles d'association valides (filtration). Dans ce cas, le logiciel donne à la fois les étapes d'exécution et le graphe implicatif adéquat selon le *minsupp* et le *minconf* fixés.

Le bouton Apriori et MGK : Un clic à ce bouton permet à l'utilisateur d'exécuter l'algorithme Apriori et l'algorithme « Association Rule Generation » deux fois modifié. Autre que le premier changement précédent, on a changé la filtration la mesure *Confiance* et *minconf* par la mesure M_{GK}^f et $M_{GK-Critique}^f$.

Le tableau Base de données : Ce tableau permet à l'utilisateur de saisir les données à traiter, après avoir donné le nombre d'individus et le nombre de caractères à étudier lors de la première option. De plus, il permet de visualiser les données importées à partir d'un fichier format .xls chargé lors de la deuxième option.

Menu : Tout d'abord, le menu est composé de quatre sous-menus tels que « Importer une base de données », « Visualiser le dernier traitement effectué », « A propos » et « Quitter ».

Le sous-menu « Importer une base de données » est l'équivalent du bouton « Ajout depuis Excel ». Ensuite, le sous-menu « Visualiser le dernier traitement effectué » permet à l'utilisateur de revoir le dernier traitement qu'il a fait avec le logiciel, après avoir quitté le logiciel. Ainsi, le logiciel mémorise les dernières actions de l'utilisateur. Après, le sous-menu « A propos » permet à l'utilisateur de connaître davantage le logiciel, tels que sa version courante, sa description, la capacité en mémoire occupée par le logiciel et son concepteur. Enfin, le sous-menu « Quitter » permet à l'utilisateur de quitter le logiciel après avoir fini ses travaux.

1.9 Résultats

Dans ce papier, nous allons traiter les notes des élèves de la série C, candidat à l'examen du baccalauréat de l'année 2017 des régions DIANA et SAVA – Madagascar (Centres :

Antsiranana I et II, Ambanja, Ambilobe, Andapa, Antalaha, Nosy Be, Sambava et Vohemar). Notre jeu de données contient 90 notes d'élèves série C et de 09 rubriques tels que la moyenne générale (MG) et les notes des 8 matières obligatoires pour la série C :

- a) 04 Matières littéraires : Malagasy (MAL), Français (FRS), Philosophie (PHI), Histoire – géographie (HG) ;
- b) 03 matières scientifiques : Mathématiques générales (MatG), Physique Chimie (PC), Science de la vie et de la terre (SN) ;
- c) et Education Physique et Sportive (EPS).

La transformation de notre jeu de données en binaire consiste à mettre 0 si la note de l'élève sur la matière considérée est strictement inférieure à la moyenne, sinon 1.

Nous avons traité notre jeu de données avec trois logiciels de traitement de volume de données différentes : SDD-GI-MGK-TCP version 1.0, Tanagra version 1.4.50 (libre) et CHIC version 6.0 (payant).

Les Figures 10, 11 et 12 présentent les résultats des traitements effectués.

Interprétation de la Figure 10 : Il est confirmé au niveau de confiance 99.99% que tous les étudiants ayant eu la moyenne supérieure ou égale à 10/20 (MG) et ayant eu la moyenne en Education Physique et Sportive (EPS) ont tous des moyennes en mathématiques générales (MatG).

Cette interprétation nous montre que la matière « mathématiques générales » est la matière clé des élèves de la série C des régions DIANA et SAVA, car ceux qui ont eu réussi son baccalauréat (MG), ont presque la moyenne en mathématiques.

Interprétation de la Figure 11 : Ceux qui ont eu la moyenne en Français (FRS ou 4) et en Histoire Géographie (HG ou 5) ont la moyenne en Malagasy (MAL ou 3), puis en EPS (8). Plus un motif se place en haut, plus le nombre des élèves ayant eu réussi cette matière est faible. La matière « mathématiques générales » est absent, car on a *support* (*MatG*) = 0.378 < *MinSup* = 0.4.

Interprétation de la Figure 12 : Ceux qui sont antécédent impliquent les conséquents. Par exemple, la règle (1) peut s'interpréter comme : ceux qui ont réussi les matières FRS et SN ont réussi la matière HG.

Interprétation de la Figure 13 : Le logiciel SDD-GI-MGK-TCP confirme avec la mesure M_{GK} , au niveau de confiance 90%, que si l'élève est fort en HG et SN, alors il l'est en FRS.

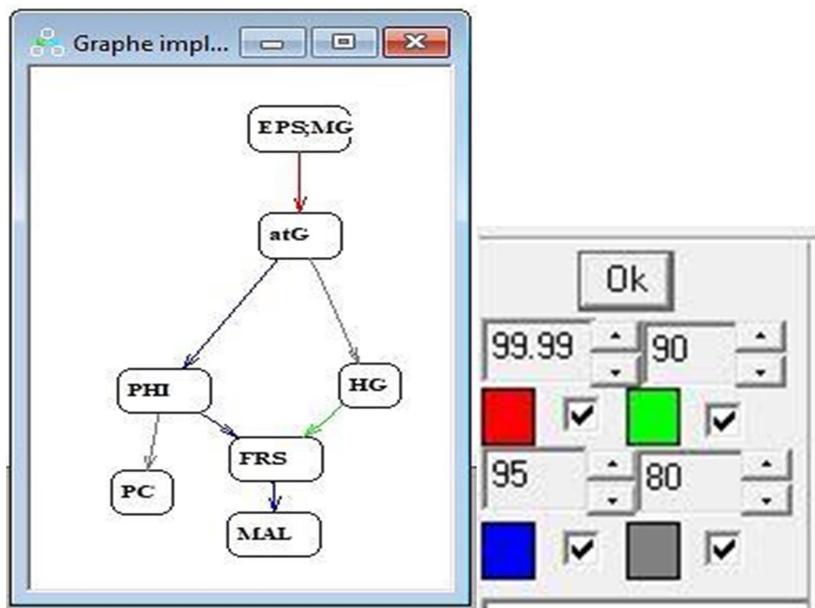


Figure 10 : Résultat du traitement avec CHIC (Graphe implicatif)

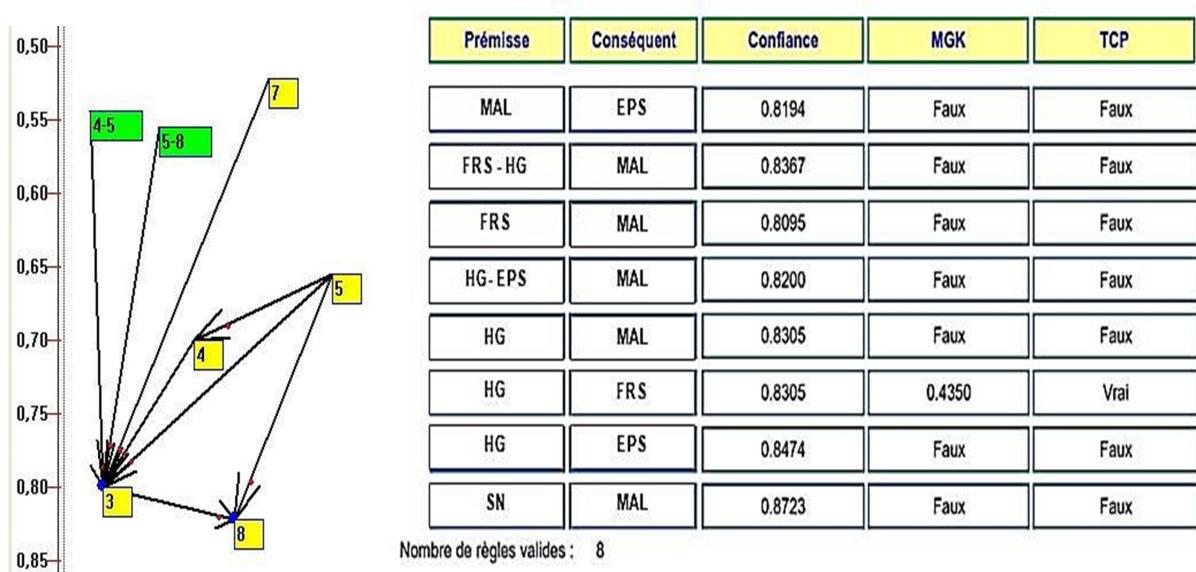


Figure 11 : Résultat du traitement avec SDD-GI-MGK-TCP avec la mesure Confiance sous les paramètres suivants : MinSup = 0.4, MinConf = 0.2, Risque pour MGK et TCP = 0.05
(Code du graphe implicatif : MAL = 3, FRS = 4, HG = 5, SN = 7 et EPS = 8)

RULES

Number of rules : 23					
N°	Antecedent	Consequent	Lift	Support (%)	Confidence (%)
1	"FRS=true" - "SN=true"	"HG=true"	1,37288	40,000	90,000
2	"HG=true" - "SN=true"	"FRS=true"	1,31868	40,000	92,308
3	"SN=true"	"HG=true"	1,26578	43,333	82,979
4	"MAL=true" - "FRS=true"	"HG=true"	1,22632	45,556	80,392
5	"EPS=true" - "FRS=true"	"HG=true"	1,22034	44,444	80,000
6	"SN=true"	"FRS=true"	1,21581	44,444	85,106
7	"PHI=true"	"FRS=true"	1,21581	44,444	85,106
8	"MAL=true" - "HG=true"	"FRS=true"	1,19534	45,556	83,673
9	"HG=true"	"FRS=true"	1,18644	54,444	83,051
10	"EPS=true" - "HG=true"	"FRS=true"	1,14286	44,444	80,000
11	"SN=true"	"MAL=true"	1,09043	45,556	87,234
12	"FRS=true" - "HG=true"	"MAL=true"	1,04592	45,556	83,673
13	"HG=true"	"MAL=true"	1,03814	54,444	83,051
14	"HG=true"	"EPS=true"	1,03069	55,556	84,746
15	"EPS=true" - "HG=true"	"MAL=true"	1,02500	45,556	82,000
16	"MAL=true" - "HG=true"	"EPS=true"	1,01765	45,556	83,673
17	"FRS=true"	"MAL=true"	1,01190	56,667	80,952
18	"PHI=true"	"MAL=true"	1,01064	42,222	80,851
19	"SN=true"	"EPS=true"	1,00920	43,333	82,979
20	"PHI=true"	"EPS=true"	1,00920	43,333	82,979
21	"EPS=true" - "FRS=true"	"MAL=true"	1,00000	44,444	80,000
22	"MAL=true"	"EPS=true"	0,99662	65,556	81,944
23	"FRS=true" - "HG=true"	"EPS=true"	0,99283	44,444	81,633

Figure 12 : Résultat du traitement avec Tanagra sous les paramètres suivants :
 Support = 0.4, Confiance = 0.8, Lift = 0.5 (pour favoriser Confiance)

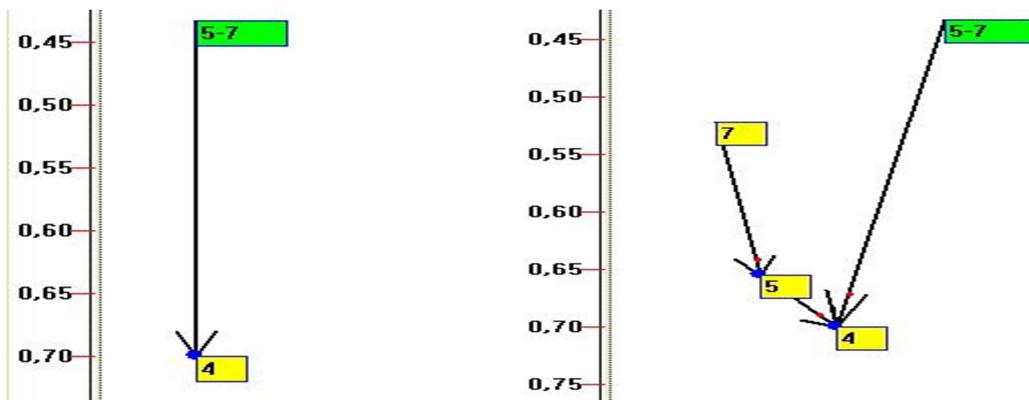


Figure 13 : Résultat de SDD-GI-MGK-TCP avec la mesure MGK au risque $\alpha = 0.1$ (à gauche) et $\alpha = 0.2$ (à droite), Code du graphe implicatif : FRS = 4, HG = 5, SN = 7.

2 Discussion

Dans cette section, nous allons lancer notre discussion sur la qualité des mesures utilisés par ces logiciels, puis sur la qualité des résultats obtenus par ces trois logiciels.

2.1 Comparaison des mesures de qualité utilisées par CHIC, Tanagra et SDD-GI-MGK-TCP

En 2013, Grissa a étudié les comportements des 61 mesures de qualité existantes dans la littérature. Il nous a offert un tableau permettant de quantifier les qualités de ces mesures considéré selon 19 propriétés non subjectives (Propriété 3- Propriété 21). Selon (Grissa 2013, p. 70-71), on a le tableau récapitulatif suivant :

Mesure	P.3	P.4	P.5	P.6	P.7	P.8	P.9	P.10	P.11	P.12	P.13	P.14	P.15	P.16	P.17	P.18	P.19	P.20	P.21	TOTAL
I.Implication	1	1	1	1	1	1	1	0	0	1	1	2	1	0	0	0	1	1	0	14
MGK	1	1	1	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	1	12
Lift	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	9
Confiance	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	8

Tableau 1 : Tableau récapitulatif des comportements des mesures de qualité des règles d'association considérées selon (Grissa 2013).

Nous déclarons que les évaluations de la mesure M_{GK} sur les propriétés P.8, P.19 et P.20 effectuées par Grissa sont erronées.

Preuves :

P.8 : M_{GK}^f est une mesure décroissante en fonction de la taille du conséquent (n_Y). En effet, on a : $M_{GK}^f(X \Rightarrow Y) = \frac{P_X(Y) - P(Y)}{1 - P(Y)} = \frac{n \cdot n_{XY} - n_X \cdot n_Y}{n_X(n - n_Y)}$. On a : $\frac{\partial M_{GK}^f(X \Rightarrow Y)}{\partial n_Y} = \frac{n(n_{XY} - n_X)}{n_X(n - n_Y)^2}$. Comme $n_{XY} \leq n_X$, alors $\frac{\partial M_{GK}^f(X \Rightarrow Y)}{\partial n_Y} \leq 0$. Ainsi, $P.8(M_{GK}^f) = 1$.

P.19 : La mesure M_{GK} est une mesure fondée sur un modèle probabiliste. En effet, on considère l'espace probabilisé (Ω, τ, P) , tel que $\forall X \in \tau, P(X) = \frac{\text{card}(X)}{\text{card}(\Omega)}$ (modèle uniforme). Ainsi, on a les deux propriétés suivantes :

- a) si $P_X(Y) > P(Y)$, alors on a : $0 \leq P_X(Y) - P(Y) \leq 1 - P(Y)$. Ce qui donne : $0 \leq \frac{P_X(Y) - P(Y)}{1 - P(Y)} \leq 1$;
- b) si $P_X(Y) \leq P(Y)$, alors on a : $-P(Y) \leq P_X(Y) - P(Y) \leq 0$. Donc, on a : $-1 \leq \frac{P_X(Y) - P(Y)}{P(Y)} \leq 0$.

Les propriétés (a) et (b) donne la définition de la mesure M_{GK} :

$$\forall X, Y \in \Omega, M_{GK}^f(X \Rightarrow Y) = \begin{cases} \frac{P_X(Y) - P(Y)}{1 - P(Y)}, & \text{si } P_X(Y) > P(Y); \\ \frac{P_X(Y) - P(Y)}{P(Y)}, & \text{si } P_X(Y) \leq P(Y). \end{cases}$$

Ainsi, la mesure M_{GK} trouve sa fondation dans la théorie des probabilités et avec un modèle probabiliste de la loi de probabilité discrète uniforme. De plus, par définition, une mesure probabiliste de qualité (MPQ) est une mesure qui se définit entièrement à partir d'un tableau de contingence (Totohasina, 2008, p. 27). Donc, on a : P.19(M_{GK}^f) = 1.

P.20 : M_{GK}^f est une mesure statistique, c'est-à-dire une mesure croissante en fonction de n . En effet, on a : $M_{GK}^f(X \Rightarrow Y) = \frac{P_X(Y) - P(Y)}{1 - P(Y)} = \frac{n_{XY} - n_X \cdot n_Y}{n_X(n - n_Y)}$. On a : $\frac{\partial M_{GK}^f(X \Rightarrow Y)}{\partial n} = \frac{n_Y(n_X - n_{XY})}{n_X(n - n_Y)^2}$. Comme $n_X \geq n_{XY}$, alors on a : $n_X - n_{XY} \geq 0$. Ainsi, $\frac{\partial M_{GK}^f(X \Rightarrow Y)}{\partial n} \geq 0$. Ainsi, on a : P.20(M_{GK}^f) = 1.

Remarque : nous considérons seulement la composante favorisante de M_{GK} , qui est M_{GK}^f car c'est la seule composante implicative et la seule qu'on utilise dans l'extraction des règles d'associations M_{GK} -valides.

Ainsi, on a le tableau rectifié suivant :

Mesure	P.3	P.4	P.5	P.6	P.7	P.8	P.9	P.10	P.11	P.12	P.13	P.14	P.15	P.16	P.17	P.18	P.19	P.20	P.21	TOTAL
MGK	1	1	1	1	1	1	1	1	0	1	1	1	0	0	1	0	1	1	1	15
I.Implication	1	1	1	1	1	1	1	0	0	1	1	2	1	0	0	0	1	1	0	14
TCP	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	1	1	10
Lift	0	1	0	1	1	1	1	0	0	1	1	1	0	0	0	0	0	0	1	9
Confiance	1	1	1	1	0	0	0	1	1	0	0	1	0	0	0	0	0	0	1	8

Tableau 2 : Tableau récapitulatif des comportements des mesures de qualité des règles d'association considérées selon (Grissa 2013) corrigé.

Remarques : Les propriétés de TCP sont étudiées dans (Tovohery, 2017). De plus, nous rappelons encore une fois que la mesure « indice d'implication », qui est le cœur de la mesure « Intensité d'implication » de Gras est une mesure M_{GK} -normalisable selon une fonction de normalisation affine (Armand, 2019, p.121). Ainsi, nous confirmons que M_{GK} est encore présent dans la mesure Intensité d'implication de CHIC.

Nous voyons dans le tableau 2 que le logiciel SDD-GI-MGK-TCP que nous venons de concevoir utilise la meilleure mesure de qualité que celles qui sont utilisées dans CHIC et Tanagra.

2.2 Comparaison des résultats obtenus

Nous remarquons que le logiciel CHIC nous permet de voir toutes les variables avec un graphe simple. Cela n'était pas le cas avec les deux autres logiciels à cause de la contrainte *MinSup*. Ensuite, nous remarquons aussi que son résultat est très résumé, car le logiciel arrive à représenter un motif à faible support ($Supp(EPS \cap MG) = 0.244$) avec ceux qui ont un support très élevé $Supp(Mal) = 0.80$. De plus, on remarque qu'il n'y a pas de redondance dans les résultats. Cela marque que sa version a subi vraiment une grande évolution.

Toutefois, le logiciel Tanagra ne donne pas de graphe implicatif, mais un résultat sous forme de tableau représentant toutes les règles valides selon les trois paramètres offerts par l'utilisateur : *MinSupp*, *MinConf* et *MinLift*. On remarque que le logiciel Tanagra ne fait point de sélection formelle, il donne seulement toutes les règles dépassant les seuils fixés par l'utilisateur.

Le logiciel SDD-GI-MGK-TCP donne à la fois un graphe implicatif et un tableau résumant les implications trouvées sur le graphe. Cela permet à l'utilisateur de voir les valeurs des mesures utilisés sur les implications valides. Une légère filtration est faite : pour deux règles symétriques valides $A \Rightarrow B$ et $B \Rightarrow A$, le logiciel retient seulement $A \Rightarrow B$ si $P(A) < P(B)$. Comme le logiciel CHIC, l'utilisateur peut déplacer et supprimer les motifs ou règles qu'il ne veut pas en cas d'abondance.

3 Conclusion

En somme, nous venons de présenter le nouveau logiciel SDD-GI-MGK-TCP, qui est un logiciel de graphe implicatif permettant d'extraire des connaissances dans un grand volume de données. Ce nouveau logiciel fonctionne avec la meilleure mesure de qualité dénommée M_{GK} , qui est le noyau de plusieurs mesures de qualité des règles d'association. Nous avons démontré via la mesure TCP que la mesure M_{GK} incarne une sémantique du taux de croissance orienté et normalisé. Une comparaison avec le logiciel CHIC (version 6.0) et Tanagra (version 1.4.50) était faite après avoir présenté les spécificités de notre nouveau logiciel. La comparaison était faite au niveau du comportement des mesures de qualité utilisées par chacun de ces logiciels, puis au niveau de la qualité des résultats obtenus. Nous remarquons que le logiciel SDD-GI-MGK-TCP essaie de combiner les belles propriétés trouvées sur CHIC et Tanagra. De plus, nous avons vu aussi que l'indice d'implication de Gras, qui est la base de la mesure utilisée dans CHIC, est M_{GK} -normalisable. D'où l'intérêt de l'usage de notre mesure M_{GK} et le logiciel associé.

La suite logique de notre travail est d'introduire dans SDD-GI-MGK-TCP un algorithme d'extraction des règles d'association quantitatives avec la mesure M_{GK} selon l'approche présentée dans (Totohasina 2008).

Remerciements :

Nous remercions Monsieur le Chef du service du baccalauréat de l'Université d'Antsiranana pour son accord à l'utilisation des données nécessaires à la réalisation de cet article.

Références :

Agrawal R., Imielinski T. & Swami A., 1993. Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD international conference on Management of data. June, 207–216, <https://dl.acm.org/doi/10.1145/170036.170072>

Armand (2019). Exploration des propriétés des homographies : applications en science des données et en didactique des mathématiques. Thèse de doctorat, Université de Toamasina. <https://hal.archives-ouvertes.fr/tel-02120375>

Bemariska P. (2016). Extraction des règles d'association selon le couple Support – MGK : Graphes implicatifs et applications en didactique des mathématiques. Thèse de doctorat, Université d'Antananarivo – Madagascar. <https://hal.archives-ouvertes.fr/tel-01466790>

Bennani Y. (2016). Science des données : Défis mathématiques et algorithmiques. AAFD et SFC à Marrakech – Royaume de Maroc, p.11 – 14.

Couturier R. & Gras R. (2005). CHIC : Traitement de données avec l'analyse implicative. Extraction et Gestion des Connaissances, Volume II, RNTI, Cepadues, Paris, p.679-684, , ISBN 2.85428.683.9. : <https://www.researchgate.net/publication/220786956>

Couturier R., Bodin A. & Gras R. (2003). Classification Hiérarchique Implicative et Cohésitive. Aide du logiciel CHIC version 6.0.

Feno D. R. (2007). Mesures de qualité des règles d'association : normalisation et caractérisation des bases. Thèse de doctorat, Université de La Réunion. <https://tel.archives-ouvertes.fr/tel-00462506>

Gras R. (2004) : L'analyse implicative : ses bases, ses développements, Revue « Educaao Matematica Pesquisa », v.4-n.2, p. 11-48, ISSN 1516-5388, Université PUC, Sao Paulo. <https://hal.archives-ouvertes.fr/hal-00442562>

Grissa D. (2013). Etude comportementale des mesures d'intérêt d'extraction de connaissances. Thèse de doctorat – Université de Tunis – El Manar, Tunisie. <https://tel.archives-ouvertes.fr/tel-01023975>

Guillaume S (2000). Traitement des données volumineuses. Mesures et algorithmes d'extraction des règles d'association et règles ordinales. France, Thèse de PhD, Université de Nantes.

Isoz V. & Rakotomalala R. (2013). Eléments de Data Mining avec Tanagra. <http://www.sciences.ch/dwnldbl/informatique/DataMiningTanagra.pdf>

Mazerolle F. (2005). Mémentos LMD – Statistique descriptive. Guiliiano éditeur. Paris, ISBN :2-84200- 91-X

Rakotomalala H. F.(2019). Classification Hiérarchique Implicative et Cohésitive selon la mesure MGK - Application en didactique de l'informatique. Thèse de doctorat, Université d'Antananarivo – Madagascar. <https://tel.archives-ouvertes.fr/tel-02172222>

Ralahady B. B. & Totohasina A. (2019). ASI-MGK: implicative statistical analysis tool based on MGK. International Journal of Computer Science and Technology (IJCST) Vol.02, No.1, January. DOI: 10.5121/ijcst.2019.01201

Ramanantsoa H. (2016). Contributions à l'amélioration de génération des bases des règles d'association MGK – valides et application en didactique des mathématiques. Thèse de doctorat, Université d'Antsiranana. <https://tel.archives-ouvertes.fr/tel-02114765>

Régnier J.C., Almouloud S.A. & Gras R. (2013). Analyse Statistique Implicative – Cadre théorique et applicatif pour l'exploration sémantique et non symétrique des données. A.S.I.7 (Résumé) São Paulo.

Shortliffe E. H & Buchanan B. G.(1975). A model of inexact reasoning in medicine. Mathematical Biosciences n°23, p. 351-379. [https://doi.org/10.1016/0025-5564\(75\)90047-4](https://doi.org/10.1016/0025-5564(75)90047-4)

Totohasina A., 2008. Contribution à l'étude des mesures de la qualité des règles d'association : normalisation sous cinq contraintes et cas de MGK : propriétés, bases composites des règles et extension en vue d'applications en statistique et en sciences physiques. HDR spécialité Mathématiques et informatique. Université de Madagascar. <https://hal.archives-ouvertes.fr/tel-02481713/document>.

Tovohery J.M. (2017). Elaboration d'un logiciel de fouille des règles d'association implicative suivant TCP et MGK et une didactique innovante de probabilité conditionnelle. Mémoire de Master en Mathématiques et Informatique, ENSET – Université d'Antsiranana, Madagascar.

Wu X. & Kumar V. (2009). The top ten algorithms in data mining. Chapman & Hall/CRC – Taylor & Francis Group, ISBN : 13: 978-1-4200-8964-6

Wu X, Zhang C., & Zhang S (2004). Efficient mining of both positive and negative association rules. ACM Transactions on Information Systems, July. <https://doi.org/10.1145/1010614.1010616>

Un nouveau logiciel de traitement de données basé sur les graphes implicatifs

TOVOHERY Josoa Michel¹, RALAHADY Bruno Bakys², TOTOHASINA André³ & FENO Daniel Rajaonasy⁴

¹Ecole Doctorale Thématische « Science, Culture, Société et Développement » de l'Université de TOAMASINA, josoamichel@tovohery@gmail.com

²ENSET- Université d'Antananarivo, ralahadybr@yahoo.fr

³ENSET- Université d'Antananarivo, andre.totohasina@gmail.com

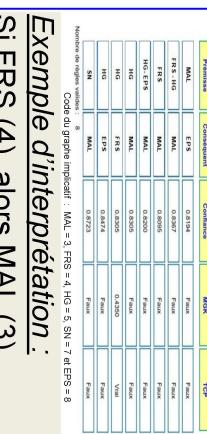
⁴Faculté de Droit, d'Économie, de Gestion, et de Mathématiques, Informatique et Applications Université de TOAMASINA, fenodanie12@yahoo.fr

1. Objectif

Présenter le nouveau logiciel appelé SDD-GI-MGK-TCP, qui est un logiciel de graphe implicatif permettant d'extraire des règles d'association du type « Si X, alors Y » dans un grand volume de données avec une mesure de qualité très performante nommée MGK.

3. Résultats

Graphé implicatif valué



Mesure de qualité utilisée	
$M_{\text{MGK}}(X \Rightarrow Y) = \begin{cases} M_{\text{MGK}}^f(X \Rightarrow Y) = \frac{P_X(Y) - P(Y)}{1 - P(Y)}, & \text{si } P_{X'}(Y') \geq P(Y'); \\ M_{\text{MGK}}^d(X \Rightarrow Y) = \frac{P_{X'}(Y') - P(Y')}{P(Y')}, & \text{si } P_{X'}(Y') < P(Y'). \end{cases}$	Sélection des motifs fréquents
$M_{\text{MGK_Critique}}(X \Rightarrow Y) = \sqrt{\frac{n_Y(n - n_X)}{n \times n_X \times n_Y} \chi^2(\alpha)}$, si $P_{X'}(Y') \geq P(Y')$	par l'algorithme « apriori » de Agrawal & Srikant sous la contrainte de MinSup.
$\text{Si } P_{X'}(Y') < P(Y')$, alors on étudie la règle $X \Rightarrow Y$.	

4. Discussion

Comparaison avec le logiciel CHIC v.6 et TANAGRA v.1.4.50

- 1) Regroupe les deux représentations de CHIC et Tanagra ;
- 2) Utilise une mesure de qualité plus performante que les mesures intensité d'implication, confiance et Lift.

2. Matériels et Méthodes

2.1. Etapes de travail avec SDD-GI-MGK-TCP

Codage des réponses obtenues en binaire :

SDD-GI-MGK-TCP :

C_1	C_2	...	C_k
0	1	...	0
1	1	...	0
⋮	⋮	⋮	⋮
1	0	...	1

$C_{ij}=1$ si la réponse de l'individu i satisfait le souhait S_j sur le caractère C_j (en format .xls).

2.2. Fonctionnement du logiciel SDD-GI-MGK-TCP

Principe des règles MGK—valides :

- 1) Si $M_{\text{MGK}}^f(Y \Rightarrow X) < M_{\text{MGK}}^f(X \Rightarrow Y)$, alors on

retient la règle $X \Rightarrow Y$, respect val. critique

En cas d'égalité, on a la règle $X \Rightarrow Y$.

- 2) Si $M_{\text{MGK_critique}}(X \Rightarrow Y) < M_{\text{MGK}}^f(X \Rightarrow Y)$ au

seuil b, alors la règle $X \Rightarrow Y$ est validée au niveau de confiance (1-b)100%.

- Résultat :**
Graphe Implicatif valué

Procédure d'extraction des règles d'association

Principe des règles MGK—valides :

- 1) Si $M_{\text{MGK}}^f(Y \Rightarrow X) < M_{\text{MGK}}^f(X \Rightarrow Y)$, alors on

retient la règle $X \Rightarrow Y$, respect val. critique

En cas d'égalité, on a la règle $X \Rightarrow Y$.

- 2) Si $M_{\text{MGK_critique}}(X \Rightarrow Y) < M_{\text{MGK}}^f(X \Rightarrow Y)$ au

seuil b, alors la règle $X \Rightarrow Y$ est validée au niveau de confiance (1-b)100%.

6. Références

[1] Couturier R., Traitement de l'analyse statistique dans CHIC.

[2] Totohasina A., 2008. Contribution à l'étude des mesures de la qualité des règles d'association : normalisation sous cinq contraintes et cas de MGK, propriétés, bases composées des règles et extension en vue d'applications en statistique et en sciences physiques. HDR spécialité Mathématiques et informatique. Université de Madagascar, <https://hal.archives-ouvertes.fr/tel-02481713/document>

[3] Feno R. J., 2007. Mesure de qualité des règles d'association : normalisation et caractérisation des bases. Thèse de Doctorat. Université de La Réunion.