



**HAL**  
open science

## Une mesure de cohésion basée sur la mesure de qualité des règles d'association M\_GK

Hery Frédéric Rakotomalala, André Totohasina, Jean Diatta

### ► To cite this version:

Hery Frédéric Rakotomalala, André Totohasina, Jean Diatta. Une mesure de cohésion basée sur la mesure de qualité des règles d'association M\_GK. SFC 2017 - XXIV èmes Rencontres de la Société Francophone de Classification., Jun 2017, Lyon, France. hal-02486446

**HAL Id: hal-02486446**

**<https://hal.science/hal-02486446>**

Submitted on 21 Feb 2020

**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.

# Une mesure de cohésion basée sur la mesure de qualité des règles d'association $M_{GK}$

Hery Frédéric Rakotomalala\*, André Totohasina\*  
Jean Diatta\*\*

\*Ecole Normale Supérieure pour l'Enseignement Technique  
BP. 0, Université d'Antsiranana, 201 - Antsiranana, Madagascar  
fredericrakotomalala@yahoo.fr  
andre.totohasina@gmail.com  
<http://www.univ-antsiranana.edu.mg>

\*\*Laboratoire d'Informatique et de Mathématiques  
Université de La Réunion, France  
jean.diatta@univ-reunion.fr  
<http://www.lim.univ-reunion.fr>

**Résumé.** Cet article est le prolongement de l'extraction des règles d'association (RA) *Support*-valides à base de  $M_{GK}$ . L'objectif est de donner un éclairage aux valeurs des RA supports-valides et de faire émerger certaines RA à faibles *Supports*, mais peuvent avoir une très forte *Confiance*. Cela présente un réel intérêt. Le modèle créé nous permet de construire un nouvel indice de cohésion, pièce maîtresse d'une Classification Hiérarchique Implicative et Cohésitive.

## 1 Introduction

L'analyse statistique implicative introduite par Régis Gras (Gras et Guillet, 2013) a connu ces derniers temps un succès dans le cadre de la fouille des RA comme outil d'analyse de données. L'analyse statistique implicative est une méthode qui permet de découvrir les RA pertinentes à partir d'un test d'hypothèses sur des données variées, et donne une représentation hiérarchique des métarègles et une analyse des contributions des attributs et individus aux différentes associations. Il a été ainsi conçu un logiciel de la classification hiérarchique, implicative et cohésitive (dénommé CHIC) analysant les données par la méthode probabiliste de l'implication statistique entre variables. Cet outil fournit l'arbre cohésitif, qui associe les variables en classes selon cohésion. Il se traduit par une hiérarchie descendante qui emboîte les classes. La cohésion a été modélisée à partir de la mesure d'intérêt *Intensité d'Implication* basée sur le modèle poissonnien. Par rapport aux caractéristiques de la mesure de la qualité des RA  $M_{GK}$  qui est implicative et non symétrique, il nous inspire à construire un nouvel indice de cohésion permettant de construire une nouvelle classification hiérarchique implicative et cohésitive. Notre stratégie consiste à créer un nouvel indice de cohésion basée sur  $M_{GK}$ . Pour ce faire, nous procédons au processus de la normalisation de la mesure  $supp_{M_{GK}}$  (section 2). Ensuite, nous construisons le nouvel indice de cohésion à partir de la normalisée de la  $supp_{M_{GK}}$  (section 3). Enfin, nous terminons par la conclusion et perspectives (section 4).

Une mesure de cohésion basée sur  $M_{GK}$

## 2 Normalisation de la mesure de qualité $supp_{M_{GK}}$

Dans l'article précédent intitulé "Extraction des RA  $M_{GK}$ -valides avec contribution du *Support*", nous avons défini la relation entre le *Support* d'une RA et sa valeur  $M_{GK}$  par les définitions ci-après.

**Définition 2.1.** Soit  $X$  et  $Y$  deux motifs d'un contexte de fouille de données. On définit  $M_{GK}$  par :  $M_{GK}(X \rightarrow Y) = \begin{cases} \frac{P(Y'|X')-P(Y')}{1-P(Y')}, \text{ si } X \text{ favorise } Y, (P(Y'|X') > P(Y')); \\ \frac{P(Y'|X')-P(Y')}{P(Y')}, \text{ si } X \text{ défavorise } Y, (P(Y'|X') < P(Y')). \end{cases}$  (1)

**Définition 2.2.** Soit une RA  $v_i \rightarrow v_j$ . On appelle support d'une RA à base de  $M_{GK}$ , telle que  $v_i$  favorise  $v_j$ , la quantité notée  $supp_{M_{GK}}(v_i \rightarrow v_j)$  définie par :

$$supp_{M_{GK}}(v_i \rightarrow v_j) = supp(v_i) \left[ (1 - supp(v_j)) M_{GK}^f(v_i \rightarrow v_j) + supp(v_j) \right] \quad (2)$$

Dans (Totohasina et al., 2004), on a montré l'existence d'une mesure normée et centrée appelée *ION* à l'époque, qui n'est autre que  $M_{GK}$ , et qui permet de présenter une vue unificatrice des différentes mesures de qualité. L'objectif de la normalisation, c'est de ramener les valeurs de la mesure de la qualité  $supp_{M_{GK}}$  sur l'intervalle  $[-1, 1]$  tout en reflétant les situations de référence telles que l'incompatibilité, la dépendance négative, l'indépendance, la dépendance positive et l'implication logique entre la prémisse et le conséquent d'une RA.

Processus de la normalisation (Armand et Feno, 2016). Soit  $x_f$  (resp.  $y_f$ ) le coefficient de multiplication (resp. de centrage) de  $supp_{M_{GK}}$  dans le cas où  $v_i$  favorise  $v_j$ . De même,  $x_d$  (resp.  $y_d$ ) dans le cas où  $v_i$  défavorise  $v_j$ . Désignons par  $supp_{(n)M_{GK}}$  la mesure normalisée associée à  $supp_{M_{GK}}$ .

On a alors,  $supp_{(n)M_{GK}}(v_i \rightarrow v_j) = \begin{cases} x_f \cdot supp_{M_{GK}}(v_i \rightarrow v_j) + y_f, & \text{si } v_i \text{ favorise } v_j; \\ x_d \cdot supp_{M_{GK}}(v_i \rightarrow v_j) + y_d, & \text{si } v_i \text{ défavorise } v_j. \end{cases}$

Les coefficients  $x_f$ ,  $y_f$ ,  $x_d$  et  $y_d$  se déterminent par passage aux limites dans les situations de référence, du fait de la continuité de l'évolution dans les deux zones : attraction (dépendance positive) et répulsion (dépendance négative). Posons  $supp_{M_{GKimp}}(v_i \rightarrow v_j)$  la valeur de  $supp_{M_{GK}}(v_i \rightarrow v_j)$  à l'implication,  $supp_{M_{GKind}}(v_i \rightarrow v_j)$  celle de  $supp_{M_{GK}}(v_i \rightarrow v_j)$  à l'indépendance et  $supp_{M_{GKinc}}(v_i \rightarrow v_j)$  la valeur de  $supp_{M_{GK}}(v_i \rightarrow v_j)$  à l'incompatibilité.

Cela donne le système d'équations linéaires suivant :

$$\begin{cases} x_f supp_{M_{GKimp}}(v_i \rightarrow v_j) + y_f = 1, & \text{implication logique} \\ x_f supp_{M_{GKind}}(v_i \rightarrow v_j) + y_f = 0, & \text{indépendance à droite} \\ x_d supp_{M_{GKind}}(v_i \rightarrow v_j) + y_d = 0, & \text{indépendance à gauche} \\ x_d supp_{M_{GKinc}}(v_i \rightarrow v_j) + y_d = -1, & \text{incompatibilité} \end{cases}$$

Après avoir résolu le système d'équations et avoir calculé les coefficients de multiplication et de centrage,

$$\text{on a : } supp_{(n)M_{GK}}(v_i \rightarrow v_j) = \begin{cases} \frac{supp_{M_{GK}}(v_i \rightarrow v_j) - P(v_i)P(v_j)}{P(v_i)(1-P(v_j))}, & \text{si } v_i \text{ favorise } v_j; \\ \frac{supp_{M_{GK}}(v_i \rightarrow v_j) - P(v_i)P(v_j)}{P(v_i)P(v_j)}, & \text{si } v_i \text{ défavorise } v_j. \end{cases}$$

La composante favorisante de la normalisée du support basé sur  $M_{GK}$  est donnée par :

$supp_{(n)M_{GK}}^f(v_i \rightarrow v_j) = \frac{supp_{M_{GK}}(v_i \rightarrow v_j) - P(v_i)P(v_j)}{P(v_i)(1-P(v_j))} = M_{GK}^f(v_i \rightarrow v_j)$ . Nous constatons que  $supp_{(n)M_{GK}}^f$  contraste la valeur de  $supp_{M_{GK}}$ , en même temps permet de faire émerger les RA à forte *Confiance* (les pépites de connaissance : Cf. tableau 1. en gras).  $supp_{(n)M_{GK}}^f$  nous permet de créer un nouvel indice de cohésion à base de  $M_{GK}$ .

### 3 Cohésion implicative du couple de variables selon $M_{GK}$

La cohésion est un indicateur d'ordre implicatif au sein d'une classe de variables. C'est une opposition au « désordre », elle nous amène à se référer à l'entropie d'une expérience aléatoire. L'entropie est déjà utilisée par Shannon en théorie de l'information apparaissant naturellement dans son article en 1948 (Shannon, 1948). Prenons le cas de deux variables qui forment une classe  $(v_i, v_j)$ . Soit  $X$  la variable aléatoire indicatrice de l'événement, pour chaque couple de variables  $(v_i, v_j)$ ,  $p$  représente la portée ou la fiabilité de la RA formée par ledit couple :  $Pr(X = 1) = \text{supp}_{M_{GK}}(v_i \rightarrow v_j) = p$  et  $Pr(X = 0) = 1 - \text{supp}_{M_{GK}}(v_i \rightarrow v_j) = 1 - p$ . L'entropie de cette expérience est  $H = -p \log_2 p - (1 - p) \log_2 (1 - p)$ .

Si  $\text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) = 1$ , alors  $H = 0$  en convenant  $p \log_2 p = 0$ .

Si  $\text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) = 0,5$ , alors  $H = 1$  (entropie maximale). Par analogie à la définition d'une cohésion implicative de deux variables (Ratsimba-Rajohn, 1992), on choisit « le carré de l'entropie  $\text{supp}_{M_{GK}}(v_i \rightarrow v_j)$  » pour renforcer un contraste dans  $[0, 1]$  et « la racine carrée de son complément à 1 » pour donner à la cohésion la dimension de l'entropie et pour accroître sa valeur numérique (pour  $x \in [0, 1], \sqrt{1 - x^2} \geq 1 - x$ ). Lorsque l'implication est stricte, on prend la cohésion de  $(v_i, v_j)$  égale à 1.

**Définition 3.1.** L'indice de Cohésion du couple de variables  $(v_i, v_j)$  à base de la mesure normalisée,  $\text{supp}_{(n)M_{GK}} \in [0,5, 1]$  tel que  $v_i \leq v_j$ , et notée  $\text{coh}_{\text{supp}_{(n)M_{GK}}}$ , est défini par :

$$\text{coh}_{\text{supp}_{(n)M_{GK}}} = \begin{cases} \sqrt{1 - (\text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j))^2} & , \text{ si } \text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) > 0,5 \\ 0 & , \text{ si } \text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) \leq 0,5 \\ 1 & , \text{ si } \text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) = 1 \end{cases}$$

Le tableau 2 nous donne les cohésions des couples de variables calculées à base de la normalisée des supports des règles valides à base de  $M_{GK}$ .

**Proposition 3.1.** L'indice de cohésion à base de  $\text{supp}_{(n)M_{GK}}$  est favorablement implicative :

Si  $v_i$  favorise  $v_j$ ,  $\text{coh}_{\text{supp}_{(n)M_{GK}}}(v_i, v_j)$  est équivalent à sa contraposée, c'est-à-dire :

$$\text{coh}_{\text{supp}_{(n)M_{GK}}}(\bar{v}_j, \bar{v}_i) = \text{coh}_{\text{supp}_{(n)M_{GK}}}(v_i, v_j).$$

**Démonstration :** Nous avons su que :  $\text{supp}_{(n)M_{GK}}^f(v_i \rightarrow v_j) = M_{GK}^f(v_i \rightarrow v_j)$ .

Alors,  $\text{supp}_{(n)M_{GK}}^f(\bar{v}_j \rightarrow \bar{v}_i) = M_{GK}^f(\bar{v}_j \rightarrow \bar{v}_i)$ .

$$\text{coh}_{\text{supp}_{(n)M_{GK}}}(\bar{v}_j, \bar{v}_i) = \sqrt{1 - (\text{supp}_{(n)M_{GK}}^f(\bar{v}_j \rightarrow \bar{v}_i))^2} = \sqrt{1 - (M_{GK}^f(\bar{v}_j \rightarrow \bar{v}_i))^2},$$

or il a été déjà démontré que  $M_{GK}$  est favorablement implicative c'est-à-dire :

$$M_{GK}^f(\bar{v}_j \rightarrow \bar{v}_i) = M_{GK}^f(v_i \rightarrow v_j),$$

$$\text{on a : } \text{coh}_{\text{supp}_{(n)M_{GK}}}(\bar{v}_j, \bar{v}_i) = \sqrt{1 - (M_{GK}^f(v_i \rightarrow v_j))^2} = \sqrt{1 - (\text{supp}_{(n)M_{GK}}^f(v_i, v_j))^2}.$$

Et cela nous montre aussi que :  $\text{coh}_{\text{supp}_{(n)M_{GK}}}(\bar{v}_j, \bar{v}_i) = \text{coh}_{\text{supp}_{(n)M_{GK}}}(v_i, v_j)$ .

**Proposition 3.2.** L'indice de cohésion à base de  $\text{supp}_{(n)M_{GK}}$  n'est pas symétrique :

si  $v_i$  favorise  $v_j$ ,  $\text{coh}_{\text{supp}_{(n)M_{GK}}}(v_j, v_i) \neq \text{coh}_{\text{supp}_{(n)M_{GK}}}(v_i, v_j)$ .

**Démonstration :** Comme  $M_{GK}$  n'est pas symétrique au cas où  $v_i$  favorise  $v_j$ , c'est-à-dire :

$$M_{GK}^f(v_j \rightarrow v_i) \neq M_{GK}^f(v_i \rightarrow v_j) \text{ alors } \text{coh}_{\text{supp}_{(n)M_{GK}}}(v_j, v_i) \neq \text{coh}_{\text{supp}_{(n)M_{GK}}}(v_i, v_j).$$

Une mesure de cohésion basée sur  $M_{GK}$

$supp_{(n)M_{GK}}$	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1			0,206		0,647			0,853		
V2							0,684		0,426	
V3	0,538				1		1	1	0,727	
V4										
V5	0,169		0,1					0,2		
V6										0,571
V7		0,236	0,1							
V8	0,744		0,333		0,667					
V9		0,426	0,211							
V10						0,25				

TAB. 1:  $supp_{(n)M_{GK}}$ -valides de couples de variables.

$coh_{supp_{(n)M_{GK}}}$	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
V1			0		0,35			0,798		
V2							0,436		0	
V3	0,091				1		1	1	0,534	
V4										
V5	0		0						0	
V6										0,17
V7		0	0							
V8	0,571		0		0,397					
V9		0	0							
V10						0				

TAB. 2: cohésions des couples de variables à base de  $supp_{(n)M_{GK}}$ .

## 4 Conclusion et perspectives

Ce travail avait pour objectif de donner un éclairage aux RA  $M_{GK}$ -valides avec la contribution de supports. D'abord, le processus de normalisation des  $supp_{M_{GK}}^f$ -valides a été entrepris pour contraster les valeurs de celles-ci et faire apparaître les RA à forte *Confiance* qui risquent d'être élaguées, mais semblent être pertinentes. Après, nous avons défini un nouvel indice de cohésion des RA à partir de  $supp_{(n)M_{GK}}^f$  afin d'en pouvoir utiliser en Classification Hiérarchique Implicative et Cohésive.

En perspective, nous envisageons d'appliquer cette théorie à l'analyse de données relevant de divers domaines tels diagnostic médical, énergie renouvelable, didactique de disciplines, etc..

## Références

- Armand, Totohasina, A. et D. Feno (2016). Nouvelle vision unificatrice des mesures d'intérêt : une normalisation par homographie. In *Proc. of AAFD & SFC*, Marrakech, pp. 287–292.
- Gras, Régis, R. J. M. C. et F. Guillet (2013). Analyse statistique implicative. méthode exploratoire et confirmatoire à la recherche de causalités. pp. 522. Cépaduès Edition.
- Ratsimba-Rajohn, H. (1992). *Contribution à l'étude de la hiérarchie implicative : Application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradictions*. Ph. D. thesis, Université de RENNES I - U.F.R. de Mathématiques, 1992. p. 210. 232.
- Shannon, C. (1948). A mathematical theory of communication. *s.l. : American Telephone and Telegraph Co. Vol. 27*, 379–423;623–656.
- Totohasina, A., H. Ralambondrainy, et J. Diatta (2004). Une vision unificatrice des mesures de la qualité des règles d'association booléennes et un algorithme efficace d'extraction des règles d'association implicative. In : *Proc. of CARI'04*, Hammamet Tunisie, pp. 511–518.

## Summary

This article is an extension of the *Support-valid*  $M_{GK}$  support extraction. The objective is to clarify the values of the rules-supports and validate some low support rules, but possibly having a very strong *Confidence* and which is of real interest. The created model allows us to construct a new cohesion index that will be useful in Classification Hierarchical Implicative and Cohesive.