ISSN No:-2456-2165

Improved Extension of MGK on Several Premisses Simplified

Chrisant GUYNO RASOLONOMENJANAHARY[1]
[1]Phd Student, Thematic doctoral school: "Science, culture, society and development", TOAMASINA Madagascar, hosting team: Mathematics, Computer Science and Application.

AMBEONDAHY[2]
[2]Institut Supérieure en Science de Téchnologie de Mahajanga (ISSTM), University of Mahajanga, Mahajanga – 401 – Madagascar

André TOTOHASINA[3]
[3]Mention: Education-Learning-Didactics and Engineering in Mathematics and Computer Science, Course: Stochastic Modeling, Ecole Normale Supérieure pour l'Enseignement Technique (ENSET), University of Antsiranana – B.P. 0 -Antsiranana 201 – Madagascar

Abstract:- Extracting knowledge as association rule is one of the important results from data mining. His first appearance was in the domain of medicine when Shortiliff's team had developed the MYCIN an expert system on diseases before Agrawal and his team have focus their research on it. In MYCIN, they have used certainty factor measure to evaluate a rule. By having compared this measure with MGK measure, we have seen that MGK is more efficient and safer than this one. Thanks to this efficiency, we proceeded to its extension on several premises and then simplified on two itemset. As application, we have studied a covid-19 dataset to study the implication: $\mathbf{symptoms} \rightarrow \mathbf{covid-19}$.

***Keywords*:-** *Association Rule, Measure, Probability, Patterns Frequent, Certainty Factor, M_{GK}, Premises/Consequent.*

I. INTRODUCTION

Currently, the amount of data is increasing exponentially through an explosion of computerization in society [1] [2]. On this huge amount of data, we can extract knowledge. Moreover, it is not on the whole data that contains useful information, but only on a significant part. In the case analysis of association rules, probabilistic quality measures are tools that make it possible to reach this significant data and to extract surprising useful knowledge. Before the 70s, we had used the conditional probability to select these rules. In 1975, Shortliffe could see that the use of probabilities is not enough and he came up with a new tool called Certainty Factor. Later, in 1993, Agrawal and his team also found another more interesting way to extract knowledge and since, several measures have appeared; the MGK measure was one of them. The particularity of this measure is that it has well-defined properties that are well founded on mathematics-statistics theories. We will take a closer look at the two measures: Certainty Factor (CF) and the Guillaume MGK measure, then present the extended form of both measure and the improvement that we have bring on it to generalize the extension.

II. MATERIALS AND METHODES

A. *Certainty Factor*

Since 1975, an expert system model has emerged for diagnosing and treating [3] [4]. Since then, this model has become a standard approach to model the uncertainty in the system based on association rules because it is reasoned from cause to effect or vice versa. Before its appearance, researchers in artificial intelligence used conditional probabilities and Bayesian probabil- ities which are mutually exclusive and exhaustive. The expert model was first used in a diagnostic tool in medicine called MYCIN, by Shortliffe and Buchanna, so that the knowledge symptoms or evidence allows the disease in question to be deduced. Thanks to the model they introduced, the MYCIN has become a powerful representation tool. Their intention was to represent the uncertainty by a probability between 0 and 1.

- *Original Version of Certainty Factor*

The Certainity factor (CF) measure is initially defined, with A and e a hypothesis and evidences, by:

$$CF(A,e) = MB(A,e) - MD(A,e) \quad (1)$$

Where

$$MB(A|e) = \begin{cases} 1 & , if\ P(A) = 1 \\ \dfrac{\max(P(A|e), P(A)) - P(A)}{\max(1,0) - P(A)} & , if\ not \end{cases} \quad (2)$$

$$MD(A|e) = \begin{cases} 1 & , if\ P(A) = 0 \\ \frac{\min(P(A|e), P(A)) - P(A)}{\min(1,0) - P(A)} & , if\ not \end{cases} \quad (3)$$

With

- $MB(A, e)$ is called measure of increased belief in the hypothesis A, based on the evidence e
- $MD(A, e)$ is called measure of increased disbelief in the hypothesis A, based on the evidence e
- $CF(A, e)$ reads: « Certainity Factor of hypothesis A based on the evidence e ».

We say that a hypothesis based on an evidence is significant if its absolutevalue is greater than 0.2 (by convention).

➢ *From these Definitions, we can Draw the following Characteristics:*

- CF gives a value between -1 and +1
- If CF is positive, then the hypothesis is validated by the evidences ($CF(A|e) \geq 0.2, P(A|e) \geq P(A)$). So, the higher the CF, the more the evidence confirms the correctness of the hypothesis.
- $CF = 1$, if hypothesis is correct.
- CF is negative ($CF \leq -0.2$), if the evidence reduces the credulity of the hypothesis. Then confirm its negation.
- $CF = -1$, if the evidence totally rejects the hypothesis, i.e $CF(A|e) = -1 \iff CF(\neg A|e) = 1$
- If $CF = 0$, then nothing can be said, so the hypothesis is then assumed to be false.

➢ *Notes:*

- If $P(A|e) = P(A)$, then we are in the case of independence betweenthe hypothesis and the evidence. So, the evidence can neither confirm or reject the hypothesis. Thus, $MD(A|e) = MB(A|e) = 0$.
- As long as $MB(A|e) \neq 0$, then $MD(A|e) = 0$ and vice versa. Because, evidence cannot both increase and reduce gullibility of a hypothesis.
- Beliefs can also be defined by:

$$MD(A, e) = \begin{cases} 0, & if\ MB \neq 0 \\ \frac{P(A) - P(A|e)}{P(A)}, & if\ e\ disfavors\ A\ (P(A|e) \leq P(A)) \end{cases} \quad (4)$$

$$MB(A, e) = \begin{cases} 0, & if\ MD \neq 0 \\ \frac{P(A) - P(A|e)}{1 - P(A)}, & if\ e\ favors\ A\ (P(A|e) \geq P(A)) \end{cases} \quad (5)$$

- The values of MB and MD are in the interval [0, 1]
- $MD(A|e) = 1 \iff MB(\neg A|e) = 1$

✓ *MB and MD properties*

Let e_1 and e_2 be two evidences such that $e = (e_1, e_2)$ and h_1 and h_2 two hypothesis such that $h = (h_1, h_2)$.

✓ *Incrementation of Evidences*

$$MD(h_1, e_1 \wedge e_2) = \begin{cases} 0, & if\ MB(h_1, e_1 \wedge e_2) = 1 \\ MD(h_1, e_1) + MD(h_1, e_2)\big(1 - MD(h_1, e_1)\big), & otherwise \end{cases} \quad (6)$$

$$MB(h_1, e_1 \wedge e_2) = \begin{cases} 0, & if\ MD(h_1, e_1 \wedge e_2) = 1 \\ MB(h_1, e_1) + MB(h_1, e_2)\big(1 - MB(h_1, e_1)\big), & otherwise \end{cases} \quad (7)$$

✓ *Conjunctions of Hypotheses*

$$\begin{cases} MB(h_1 \wedge h_2, e) = \min[\, MB(h_1, e), MB(h_2, e)] \\ MD(h_1 \wedge h_2, e) = \max[\, MD(h_1, e), MD(h_2, e)] \end{cases} \quad (8)$$

- ✓ *Disjonctions of Hypothesis*

$$\begin{cases} MB(h_1 \vee h_2, e) = \min[\,MB(h_1,e), MB(h_2,e)] \\ MD(h_1 \vee h_2, e) = \max[\,MD(h_1,e), MD(h_2,e)] \end{cases} \qquad (9)$$

- ✓ *Remarks*

- ▪ The first property is a point which is not treated by the conditional probability that was used before the appearance of the CF measure
- ▪ The last two properties are only conventions for compiling the program.

➢ *Certainty Factor Improved by Bill Van Melle*

After an analysis, Bill [4] [5] was able to observe two significant flaws in the Buchanna CF measure:

- The Potential for Negative Evidence has Overturned Some Positive Evidence:

A number of evidences that justify the hypothesis can be dismissed through a single negative evidence. If the value of several evidences is 0.999 and one evidences has a value of -0.8 then, $CF = 0.999 - 0.8 = 0.199 < 0.2$, the evidence has no meaning.

- The Memory Capacity Used By MB And MD Is Very Large.

To remedy to these problems, Bill redefined CF as follows [4]:

$$CF = \frac{MB - MD}{1 - \min(MB, MD)} \qquad (10)$$

Then

$$CF_{Combined}(X,Y) = \begin{cases} X + Y(1-X), & if\ X,Y > 0 \\ \dfrac{X+Y}{1-\min(|X|,|Y|)}, & if\ X < 0\ or\ Y < 0 \\ -CF_{Combined}(-X,-Y), & if\ X,Y < 0 \end{cases} \qquad (11)$$

Where X, Y are a CF. This way of reasoning is very logic and that let us to redefine another measure further.

B. M_{GK}Measure

The measure of interest M_{GK} of association rule takes various independences designation, according to researchers and the year of its discovery [6]: inspired by the Loevinger index, M_{GK} (measure of Guillaume-Kenchaff)was independently proposed and named in 2000 by Guillaume, CPIR (Conditionnal Probability Incrementation Ratio) in 2004 by Wu and Zhang, ION (Implication Oriented Normalized) in 2003 by Totohasina, verifying the implicative oriented property of Brin, Motwani, and Silverstaein, in 1997. Due to the expression of a minimum condition and efficiency ratio to extract the non-redundant rules, and applying the Support and the Confidence in the implications of Ferré (2002) shows that this measure is both more precise and understandable. Before talking about the M_{GK} measure, let us first talk about the measure of the quality of association rules in its generality.

➢ *Quality Measure of Association Rules*

The quality measures of the association rules, or measures of interest or probabilistic quality measures, are numerical indicators intended to guide the user to potentially interesting knowledge in the large volumes of rules produced by the mining algorithms of data. These measures evaluate the quality of the rules according to different points of view, and make it possible to order the rules from the best to the worst. They can also play the role of filter, by rejecting the rules below a minimum quality threshold.

➢ *Formal Context*

Let $\mathcal{K} = (\mathcal{T}, \mathcal{A}, \mathcal{R})$ be a binary data mining context, be a binary data mining context, such that $\mathcal{T}$ is a finite set of transaction, A the set of items or variables and $\boldsymbol{\mathcal{R}}$ a binary relation. Let X and Y be two patterns of $\mathcal{A}$, i.e $(X,Y) \in \mathcal{P}(\mathcal{A}) \times \mathcal{P}(\mathcal{A})$, and X' and Y' their respective extension such that $X' = \{t \in \mathcal{T} | \forall x \in X, t\mathcal{R}x\}$ and $Y' = \{t \in \mathcal{T} | \forall y \in X, t\mathcal{R}y$. In all transactions, which we noted $\mathcal{T}$, we can define a discrete probability space $(\mathcal{T}, \mathcal{P}(\mathcal{T}), P)$ where P is a discrete uniform probability [7] [8]. Thus, if $X' \subset \mathcal{T}$ then $P(X) = \frac{|X'|}{|\mathcal{T}|} = \frac{n_X}{n}$. If X' and Y' being two units such that $P(X' \cap Y') \geq \alpha$, where α is a discrete uniform probability (like the value 0.2 indicating the significance of CF). From these significant combinations, we can develop an association rule such as that $X \to Y$ or $Y \to X$.

➢ *A Quality Measure: $\boldsymbol{\mu}$*

Let $X \in \mathcal{P}(\mathcal{A})$ and $Y \in \mathcal{P}(\mathcal{A})$be patterns. A quality measure probabilist is a real function μ of $\mathcal{P}(\mathcal{A}) \times \mathcal{P}(\mathcal{A})$ such that for any rule of association $X \to Y$, $\mu(X \to Y)$ is a real value calculated from the four quantities: $n, P(X'), P(Y')$ and $P(X' \cap Y')$, where P denote the uniform discrete probability over the probability space $(\boldsymbol{\mathcal{I}}, \boldsymbol{\mathcal{P}}(\boldsymbol{\mathcal{I}}))$. It is said symetric (resp perfectly symetric) if $\mu(X \rightarrow Y) = \mu(Y \rightarrow X)$ (resp $\mu(X \rightarrow Y) = \mu(\bar{X} \rightarrow \bar{Y})$). According to Frédéric [9] [2] and André Totohasina [10] [7], the quality of a measurement is measured by satisfying some of the following criteria :

- Understandability of the Measurement for the User;
- Nature of the Rules Targeted by the Measure;
- Direction of Measurement Variation;
- Nature of the Variation: Linear / Non-Linear;
- Impact of the Scarcity of the Consequent;
- Sensitivity to Data Size;
- Discriminant Nature of the Measure;
- Use of a Pruning Threshold;
- Classification Induced by a Measure;
- Behavior in Relation to the Context of the Studied Rules;
- Deviation from Equilibrium;
- Contradiction of the User's a Priori Knowledge;
- Noise Sensitivity.

Given the number of criteria to be satisfied, it is impossible to satisfy all of them. Currently, several measures have been presented in the literature, Grissa [11] and Rakotomalala [2] have drawn up a list of these measures.

➢ *M_{GK} Measure*

The M_{GK} measure is defined by [7]:

$$M_{GK}(X \rightarrow Y) = \begin{cases} \dfrac{P(Y'|X') - P(Y')}{1 - P(Y')}, & if\ P(Y'|X') > P(Y') \\ \dfrac{P(Y'|X') - P(Y')}{P(Y')}, & if\ P(Y'|X') \leq P(Y') \end{cases} \qquad (12)$$

- *Indeed:*

If X favors Y, on the one hand $P(Y'|X') \geq P(Y')$ therefore $P(Y'|X') - P(Y') \geq 0$ and on the other hand $1 \geq P(Y'|X') \geq 0$, therefore

$$1 - P(Y') \geq P(Y'|X') - P(Y') \geq -P(Y') \qquad (13)$$

Therefore

$$1 - P(Y') \geq P(Y'|X') - P(Y') \geq 0 \qquad (14)$$

(because$P(Y'|X') - P(Y') \geq 0$),

Therefore

$$1 \geq \frac{P(Y'|X') - P(Y')}{1 - P(Y')} \geq 0 \qquad (15)$$

If X disfavors Y, on the one hand, $P(Y'|X') \leq P(Y')$ therefore $P(Y'|X') - P(Y') \leq 0$ and on the other hand $1 \geq P(Y'|X') \geq 0$, therefore

$$1 - P(Y') \geq P(Y'|X') - P(Y') \geq -P(Y') \qquad (16)$$

Therefore,

$$0 \geq P(Y'|X') - P(Y') \geq -P(Y') \qquad (17)$$

(because $P(Y'|X') - P(Y') \leq 0$)

Therefore,

$$0 \geq \frac{P(Y'|X') - P(Y')}{P(Y')} \geq -1 \qquad (18)$$

If we are going to denote by

$$M_{GK}^{f}(X \rightarrow Y) = \frac{P(Y'|X') - P(Y')}{1 - P(Y')} \qquad (19)$$

Called Furthering Component, and

$$M_{GK}^{d}(X \to Y) = \frac{P(Y'|X') - P(Y')}{P(Y')} \qquad (20)$$

Called Un-Furthering Component, then the M_{GK} Measure can be Expressed:

$$M_{GK}(X \to Y) = \begin{cases} M_{GK}^{f}(X \to Y), & if\ X\ favor\ Y \\ M_{GK}^{d}(X \to Y), & if\ X\ disfavor\ Y \end{cases} \qquad (21)$$

- *Threshold Validation of M_{GK}*

After calculated $M_{GK}(X \to Y)$, now we should decide if the implication is valid or not. To do it, Totohasina [6] proposed to call to χ^2-Pearson independence test, by combining it with some probability such that:

$$M_{GK}Threshold(\alpha) = \pm\sqrt{\frac{1}{n} \times \frac{n_{\bar{X}}}{n_X} \times \frac{n_Y}{n_{\bar{Y}}} \chi^2(\alpha)} \qquad (22)$$

Where α is a confidence level according to the liaison between pattern X and Y.

- *Reference Situation for M_{GK} Measure*

A definition of the reference situations is a criterion justifying the quality of a measure. The M_{GK} measure presents these situations [7] [9] [8] as following:

- *Incompatibility Situation:* X and Y are incompatible if and only if $M_{GK}(X \to Y) = -1$
- *Disfavor Situation or Negative Dependency*: X disfavor Y if and only if $-1 \leq M_{GK}(X \to Y) \leq 0$
- *Dependency Situation:* X and Y are independent if and only if $M_{GK}(X \to Y) = 0$
- *Favor Situation or Positive Dependency*: X favor Y if and only if $0 < M_{GK}(X \to Y) < 1$
- *Logically Implication Situation*: X logically implies Y if and only if $M_{GK}(X \to Y) = 1$
- *Equilibrium Situation*: in an equilibrium situation, i.e. $|X' \cap Y'| = |X' \cap \overline{Y'}|$, $M_{GK} = \pm 1/2$

III. RESULTS AND DISCUSSIONS

C. *Comparative Study Between $\boldsymbol{CF}$ And $\boldsymbol{M_{GK}}$*

As defined in equations (1), (4), (5) and (12):

$$CF(A, e) == \begin{cases} MB = \dfrac{P(A|e) - P(A)}{1 - P(A)}, & if\ e\ favors\ A \\ -MD = -\dfrac{P(A) - P(A|e)}{P(A)}, & if\ e\ disfavors\ A \end{cases}$$

$$M_{GK}(X \to Y) == \begin{cases} \dfrac{P(Y'|X') - P(Y')}{1 - P(Y')}, & if\ X\ favors\ Y \\ \dfrac{P(Y'|X') - P(Y')}{P(Y')}, & if\ X\ disfavors\ Y \end{cases}$$

By assigning common notations, we have:

$$M_{GK}(X \to Y) == \begin{cases} \dfrac{P(Y'|X') - P(Y')}{1 - P(Y')}, & if\ X\ favors\ Y \\ \dfrac{P(Y'|X') - P(Y')}{P(Y')}, & if\ X\ disfavors\ Y \end{cases}$$

$$CF(Y, X) == \begin{cases} MB = \dfrac{P(Y|X) - P(Y)}{1 - P(Y)}, & if\ X\ favors\ Y \\ -MD = -\dfrac{P(Y) - P(Y|X)}{P(Y)}, & if\ X\ disfavors\ Y \end{cases} \qquad (23)$$

ISSN No:-2456-2165

In both measures, we try to determine the level of consequence X over Y. Since, if we consider the X and Y as two products in sale in a supermarket, they cannot be attributed a probability, but it is their appearances at the cash register. Then instead of writing $P(X)$ we will write $P(X')$, X' is the extension of X [7]. From these facts, we have:

$$CF(Y,X)=\begin{cases} MB=\dfrac{P(Y'|X')-P(Y')}{1-P(Y')}, & if\ X\ favors\ Y\\ -MD=-\dfrac{P(Y')-P(Y'|X'')}{P(Y')}, & if\ X\ disfavors\ Y\end{cases} \qquad (24)$$

$$M_{GK}(X\to Y)=\begin{cases} \dfrac{P(Y'|X')-P(Y')}{1-P(Y')}, & if\ X\ favors\ Y\\ \dfrac{P(Y'|X')-P(Y')}{P(Y')}, & if\ X\ disfavors\ Y\end{cases}$$

Thus in both cases, favoring and disfavoring, $M_{GK}(X\to Y)=CF(Y,X)$. The only difference between the two measures is the choice of the validation threshold of a rule. With the CF measurement, the threshold is set at 0.2, an deterministic value. On the other hand, for M_{GK}- measure, one calls upon the independency test of χ^2 such that $M_{GK}threshold(\alpha)=\pm\sqrt{\frac{1}{n}\times\frac{\overline{X'}}{X'}\times\frac{Y'}{\overline{Y'}}\chi^2(\alpha)}$, where α is the risk level, based on mathematics theory.

D. *Extension of M_{GK} - Measure*

Suppose now that the pattern X in M_{GK} definition, is a combination of several frequent patterns $(X_1, X_2, \dots, X_k)$ which are assumed to be independent or not. By reasoning, for example in medicine field, the premises are the sets of symptoms of the patient and the consequent is disease. These symptoms can be grouped together to form the X_i, for example: those who affirm the presumption of the doctor and those who question it. In a database, several different patterns are associated with an even consequent. Now our object is to check if a pattern X implies the consequent Y $(X\to Y)$. Then in relation to the number of patterns in data, we will use the union to have a new single premise pattern**,** instead intersection, used in our previous research [12] because in practice two units can contain different items/itemset with empty intersection which has a null probability. Now, to measure this implication, it can only be done using frequent patterns in the database, defined by Agrawal [13]. Note that, the set of pattern extension are probabilisable but not the set of pattern, we note X' the extension of pattern X and $X'\cap Y'=(X\wedge Y)'$ (resp. $X'\cup Y'=(X\vee Y)'$) the extension of $X\wedge Y$ (resp. $X\vee Y$).

- *Proposition :*

Let $X=(X_1,X_2,\dots,X_k)$ a vector of pattern and Ya pattern The extension of M_{GK}-measure is given by :

$$M_{GK}\big((X_1,X_2,\dots,X_k)\to Y\big)=\begin{cases} \dfrac{P\big(Y'\big|(X_1'\cup X_2'\cup\dots\cup X_k')\big)-P(Y')}{1-P(Y')}, & if\ (X_i)_{1\le i\le k}\ favors\ Y\\ \dfrac{P\big(Y'\big|(X_1'\cup X_2'\cup\dots\cup X_k')\big)-P(Y')}{P(Y')}, & if\ (X_i)_{1\le i\le k}\ disfavors\ Y\\ \dfrac{M_{GK}^f+M_{GK}^d}{1-\min(|M_{GK}^f|,|M_{GK}^d|)}, & if\ (X_i)_{1\le i\le s}\ favors\ Y\ and\ (X_i)_{s+1\le i\le k}\ disfavors\ Y\end{cases} \qquad (25)$$

Where

$$P\big(Y'\big|(X_1'\cup X_2'\cup\dots\cup X_k')\big)=\frac{\sum_{i=1}^{k}(-1)^{i+1}\sum_{1\le i\le k}P(Y'\cap X_1'\cap\dots\cap X_i')}{\sum_{i=1}^{k}(-1)^{i+1}\sum_{1\le i\le k}P(X_1'\cap\dots\cap X_i')} \qquad (26)$$

Proof In fact,

$$M_{GK}(X\to Y)=\begin{cases} \dfrac{P(Y'|X')-P(Y')}{1-P(Y')}, & if\ X\ favors\ Y\\ \dfrac{P(Y'|X')-P(Y')}{P(Y')}, & if\ X\ disfavors\ Y\end{cases}$$

Replacing X by a pattern vector, we have

$$M_{GK}\big((X_1,X_2,\dots,X_k) \to Y\big) == \begin{cases} \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{1-P(Y')}, & if\ (X_i)_{1\le i\le k}\ favors\ Y \\ \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{P(Y')}, & if\ (X_i)_{1\le i\le k}\ disfavors\ Y \end{cases} \quad (27)$$

The expression of the measure is differentiated on the conditional probability of the consequent knowing the premise(s):

$$P(Y'|X') \Rightarrow P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big)$$

By having a database, the last expression can be calculated from two ways: either by directly calculating the probability $P(X_1' \cup X_2' \cup \dots \cup X_k')$, then deduce the conditional probability, either by decomposing the probability conditional of events. The second way is more or less versatile because it takes into account the different probabilities of patterns. Then,

$$P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) == \frac{P\big(Y' \cap (X_1' \cup X_2' \cup \dots \cup X_k')\big)}{P(X_1' \cup X_2' \cup \dots \cup X_k')}. \quad (28)$$

Using the Poincaré property, on the one hand,

$$P(X_1' \cup X_2' \cup \dots \cup X_k') = \sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(X_1' \cap \dots \cap X_i') \quad (29)$$

On The Other Hand,

$$P\big(Y' \cap (X_1' \cup X_2' \cup \dots \cup X_k')\big) = P(\bigcup_{i=1}^{k} (Y' \cap X_i')) = \sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(Y' \cap X_1' \cap \dots \cap X_i') \quad (30)$$

As we try to identify the implication between the patterns, (X_1, X_2, …,X_k) and Y, on the one hand, one cannot deduce a priori the independence or not, that is to say $P(Y' \cap X_1' \cap X_2' \cap \dots \cap X_k') = P\big(Y' \big| (X_1' \cap X_2' \cap \dots \cap X_k')\big) \times P(X_1' \cap X_2' \cap \dots \cap X_k') = P(Y')P((X_1' \cap X_2' \cap \dots \cap X_k')|Y') \neq P(Y')P(X_1' \cap X_2' \cap \dots \cap X_k')$. Then, introducing (29) and (30) into (28), we have:

$$P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) == \frac{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(Y' \cap X_1' \cap \dots \cap X_i')}{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(X_1' \cap \dots \cap X_i')} \quad (31)$$

In the case where there is independence between the premise X_i and Y, then, one,

$$P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) = P(Y') \qquad (32)$$

One the other hand, if one of the patterns X_i is disjoin with others motifs, $P(X_1' \cap X_2' \cap \dots \cap X_k') = P(\emptyset) = 0$, then $P(Y' \cap X_1' \cap X_2' \cap \dots \cap X_k') = 0$. Therefore, if there is $X_i' \cap X_j' = \emptyset$,

$$P\big(Y' \cap (X_1' \cup X_2' \cup \dots \cup X_k')\big) = \sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k-1} P(Y' \cap X_1' \cap \dots \cap X_i') \qquad (33)$$

By introducing (31) in M_{GK} measure, we have

$$M_{GK}\big((X_1,X_2,\dots,X_k) \to Y\big) = \begin{cases} \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{1-P(Y')}, & if\ (X_i)_{1\le i\le k}\ favors\ Y \\ \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{P(Y')} & if\ (X_i)_{1\le i\le k}\ disfavors\ Y \end{cases}$$

Where

$$P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) == \frac{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(Y' \cap X_1' \cap \dots \cap X_i')}{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(X_1' \cap \dots \cap X_i')}$$

In the case where the intersection of the patterns contains, on the one hand, a pattern which affirms the consequent and, on the other hand, items which lead to the exclusion of the consequent, after calculating the two expressions of the measure in question, it is normal that we must provide a new expression that completes the measure. As the premise is composed of two opposing groups, the expression should then be a combination of M_{GK}^{f} and M_{GK}^{d} which we will denote by $M_{GK}^{f,d}$ such as:

$$M_{GK}^{f,d} = \frac{M_{GK}^{f} + M_{GK}^{d}}{1 - \min(|M_{GK}^{f}|, |M_{GK}^{d}|)}, \quad if\ (X_i)_{1\le i\le s}\ favors\ Y\ and\ (X_i)_{s+1\le i\le k}\ disfavors\ Y \quad (33)$$

Consequently, our measure become

$$M_{GK}\big((X_1, X_2, \dots, X_k) \to Y\big) = \begin{cases} \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{1 - P(Y')}, & if\ (X_i)_{1\le i\le k}\ favors\ Y \\ \dfrac{P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) - P(Y')}{P(Y')}, & if\ (X_i)_{1\le i\le k}\ disfavors\ Y \\ \dfrac{M_{GK}^{f} + M_{GK}^{d}}{1 - \min\big(|M_{GK}^{f}|, |M_{GK}^{d}|\big)}, & if\ (X_i)_{1\le i\le s}\ favors\ Y\ and\ (X_i)_{s+1\le i\le k}\ disfavors\ Y \end{cases} \quad (35)$$

Where

$$P\big(Y' \big| (X_1' \cup X_2' \cup \dots \cup X_k')\big) = \frac{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(Y' \cap X_1' \cap \dots \cap X_i')}{\sum_{i=1}^{k} (-1)^{i+1} \sum_{1\le i\le k} P(X_1' \cap \dots \cap X_i')}$$

With $P(Y' \cap X_1' \cap X_2' \cap \dots \cap X_i') = P(Y')P(X_1'|Y')P(X_2'|(Y' \cap X_1')) \dots P\big(X_i' \big| (Y' \cap X_1' \cap \dots \cap X_{i-1}')\big)$ and $P(X_1' \cap X_2' \cap \dots \cap X_k') = P(X_1')P(X_2'|X_1') \dots P\big(X_k' \big| (X_1' \cap X_2' \cap \dots \cap X_{k-1}')\big)$ for all $i, j \in [\![1, k]\!], i \neq j, X_i' \cap X_j' \neq \emptyset$ and 0 if not. That which was to be proof.

To validate an implication measured by M_{GK}, we need an acceptance threshold, calculated from the test of connections of χ^2. For a level of risk α of being wrong, its expression is given by: $M_{GK}Threshold(\alpha) = \pm\sqrt{\frac{1}{n} \times \frac{n_{\bar{X}}}{n_X} \times \frac{n_Y}{n_{\bar{Y}}} \chi^2(\alpha)}$. Then, $M_{GK}(X \to Y)$ (where $X = (X_1, X_2, \dots, X_k)$) is valid, with a risk α of being wrong, if $M_{GK}^{f}(X \to Y) > |M_{GK}Threshold(\alpha)|$ or $M_{GK}^{f}(X \to Y) < -|M_{GK}Threshold(\alpha)|$.

E. M_{GK} Extended Simplified

As we have seen before, we have extended our measure by considering a large number of itemset, $X = (X_1, X_2, \dots, X_k)$. However, it is possible to reduce the k-itemset into only two itemset, one set for the items which favor the consequent and one other for which disfavor it. Then, instead k-itemset, we have two itemset $X = (X_f, X_d)$. Thus measuring $(X_1, X_2, \dots, X_k) \to Y$ reduced to measuring $(X_f, X_d) \to Y$. Applying our reduction in our measure formula, we have:

$$M_{GK}\Big(\big(X_f, X_d\big) \to Y\Big) = \begin{cases} M_{GK}^{f} = \dfrac{P\big(Y' \big| X_f'\big) - P(Y')}{1 - P(Y')}, & X_f\ \ favors\ Y \\ M_{GK}^{d} = \dfrac{P(Y' | X_d') - P(Y')}{P(Y')}, & X_d\ disfavors\ Y \\ \dfrac{M_{GK}^{f} + M_{GK}^{d}}{1 - \min(|M_{GK}^{f}|, |M_{GK}^{d}|)}, & X_f'\ favors\ Y\ and\ X_d'\ disfavors\ Y \end{cases}$$

In this simplified extension, the measure expression is simplified and the validation threshold still the same.

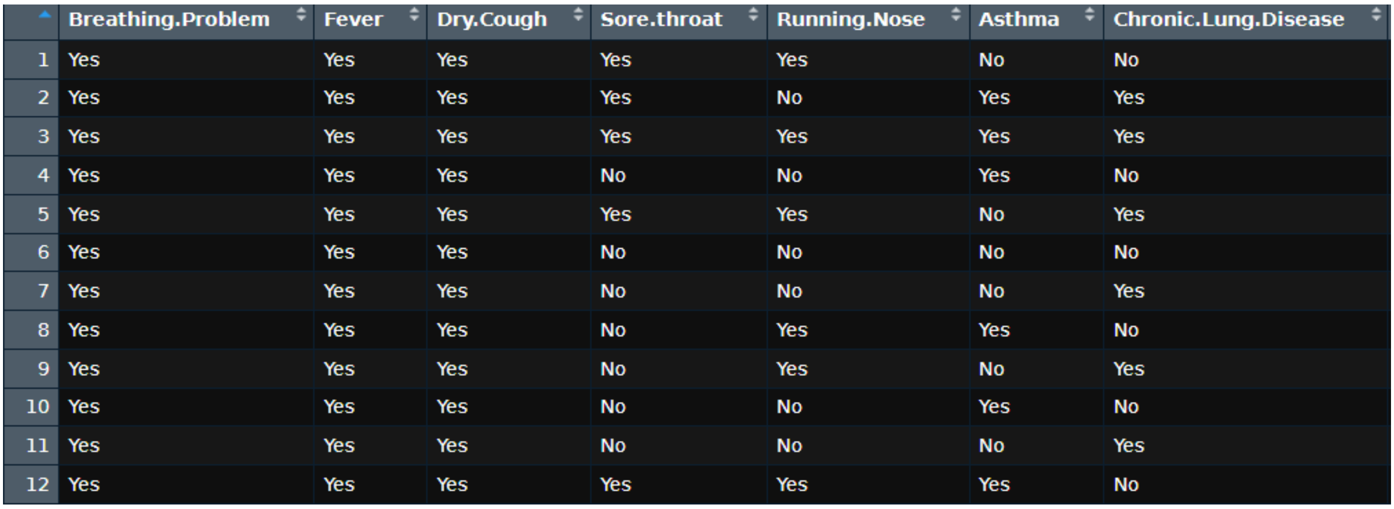
F. Application on the Covid-19 Data

To prove our theory on implication measure, let apply it on data that we have downloaded on kaggle website, where some searcher and engineer leave a data or program code that they have used on their work or research.

In our apply, we focus our study on covid-19 study. We will try to measure the implication of some symptoms observed on some patient to identify if he is ill or not. In this case, then we will identify if the patients with the symptoms are ill of covid19 or not.

- *Data Presentation*

Our data is an observation of twenty-one (21) variables on five thousand four hundred and thirty-four (5434) persons. Those variables are: "Breathing. Problem", "Fever", "Dry.Cough", "Sore.throat", "Running.Nose","Asthma", "Chronic.Lung.Disease", "Headache", "Heart.Disease", "Diabetes", "Hyper.Tension", "Fatigue", "Gastrointestinal","Abroad.travel", "Contact. with. COVID.Patient", "Attended.Large.Gathering", "Visited.Public.Exposed.Places", "Family. working.in. Public. Exposed. Places", "Wearing. Masks", "Sanitization.from.Market", “COVID.19".



	Breathing.Problem	Fever	Dry.Cough	Sore.throat	Running.Nose	Asthma	Chronic.Lung.Disease
1	Yes	Yes	Yes	Yes	Yes	No	No
2	Yes	Yes	Yes	Yes	No	Yes	Yes
3	Yes	Yes	Yes	Yes	Yes	Yes	Yes
4	Yes	Yes	Yes	No	No	Yes	No
5	Yes	Yes	Yes	Yes	Yes	No	Yes
6	Yes	Yes	Yes	No	No	No	No
7	Yes	Yes	Yes	No	No	No	Yes
8	Yes	Yes	Yes	No	Yes	Yes	No
9	Yes	Yes	Yes	No	Yes	No	Yes
10	Yes	Yes	Yes	No	No	Yes	No
11	Yes	Yes	Yes	No	No	No	Yes
12	Yes	Yes	Yes	Yes	Yes	Yes	No

Fig 1 Data Presentation

- *Results*

From our data, the probability to get person with a Covid-19 is equal 0.806. By calculating the conditional probability of Covid-19 knowing each variable (the *support* as defined by Agrawal [14] [11]), we had the following group of variables:

Table 1 Results

Favoring Variable	**Disfavoring Variable**
"Breathing.Problem", "Fever", "Sore.throat", "Asthma","Heart.Disease", "Diabetes", "Hyper.Tension", "Abroad.travel", "Contact.with.COVID.Patient", "Attended.Large.Gathering", "Visited.Public.Exposed.Places", "Family.working.in.Public.Exposed.Places", "Wearing.Masks", "Sanitization.from.Market", “COVID.19"	"Dry.Cough","Running.Nose","Chronic.Lung.Disease", "Gastrointestinal", "Headache", "Fatigue"

Knowing the group of each variable, now we can calculate their MGK. For the favoring group, the MGK measure is null ($M_{GK}(Favoring\ groupe\ \rightarrow Covid19) = 0$), for the other group, disfavoring group, the MGK is equal to -8.85×10^{-4} ($(M_{GK}(disfavoring\ groupe\ \rightarrow Covid19) = -8.85 \times 10^{-4}$). As defined, if we have two group, then we have to use the third component of our measure, thus for this third component, we have

$$M_{GK}^{f,d} = \frac{M_{GK}^{f} + M_{GK}^{d}}{1 - \min(|M_{GK}^{f}|, |M_{GK}^{d}|)} = \frac{0 + (-8.85 \times 10^{-4})}{1 - 0} = -8.85 \times 10^{-4}$$

For the validation threshold, $M_{GK}Threshold(\alpha) = \pm\sqrt{\frac{1}{n} \times \frac{n_{\bar{X}}}{n_X} \times \frac{n_Y}{n_{\bar{Y}}} \chi^2(\alpha)}$. As we have studied all variables, union of variable, in our dataset, $n_{\bar{X}} = 0$, therefore, whatever the value of risk α, the $M_{GK}Threshold(\alpha) = 0$.

- *Result Interpretation*

As $M_{GK} = -8.85 \times 10^{-4}$, if we don’t consider the threshold validation, we can conclude that with all of those symptoms diagnose in the patient, we can’t say that he has a covid-19 disease. However, if we take in consideration the threshold, we can’t conclude anything because this one is null, however the risk that we consider.

IV. CONCLUSION AND PERSPECTIVES

A huge database contains knowledge that we could not imagine. Extracting a relation as $\boldsymbol{X(cause)} \rightarrow \boldsymbol{Y(effect\ or\ consequence)}$, is very important in decision making. By having a consequent, we can therefore look for its causes from several associated frequent patterns $(\boldsymbol{X_1}, \boldsymbol{X_2}, \ldots, \boldsymbol{X_k})$ obtained in the base. The cause $\boldsymbol{X}$ will then be the union of the frequent $\boldsymbol{k}$-patterns. To validate the implication obtained, it is preferable to apply the $\boldsymbol{M_{GK}}$ measure which is a measure of association rules based on probabilistic theories, in particular on the choice of a threshold validation. We have proved that this measure can be extended to several premises patterns and summed in just two pattern combinations measurement. To validate our theory on extension, we have use a Covid-19 dataset from kaggel website fo application. As result, the MGK measure gives $\mathbf{-8.85 \times 10^{-4}}$ as value, which mean that we could not affirm the implication: $\boldsymbol{symptoms \rightarrow covid19}$. However, on use of the threshold, this one gives us a null value whatever the risk considered, so we cannot conclude anything. As continuation of this work, we suggest another experimentation with another real data in other fields and we guess to build a program for it for all kind of binary dataset.

REFERENCES

[1]. William J. Frawley and C. J. Matheus, "Knowledge Discovery in Databases: An overview," AI Magazine, vol. 13, 1992.

[2]. H. F. Rakotomalala, Classification Hiérarchique Implicative et Cohésitive selaon la mesure MGK - application en didactique de l'informatique, 2019.

[3]. D. Heckerman, "The Certainty-Factor Model," https://www.researchgate.net/publication/228798687, 1992.

[4]. B. G. Buchanan and E. H. Shortliffe, RULE-BASED EXPERT SYSTEMS : The MYCIN experiments of the stanford heuristic programming project, B. G. Buchanan and E. H. Shortliffe, Eds., Addison-Wesley Publishing Company, 1984.

[5]. E. H. S. Bruce G. Buchanan, Rule Based expert systems, vol. 753, Addison Wesley Publishing Company.

[6]. H. F. RAKOTOMALALA, "Classification Hiérarchique Implicative et Cohésitiveselon la mesure M_(GK) - Application en didactique de l'informatique," 2019.

[7]. T. André, "Contribution à l'étude des mésures de la qualité des règles d'association: normalisation sous cinq contraintes et cas de M_(GK) : propriétés, bases composites des règles et extention en vue d'applications en statistique et en sciences physiques," 2008.

[8]. D. R. Feno, "Mesures de qualité des règles d'association : normalisation et caractérisation des bases," 2007.

[9]. T. A. e. D. J. Rakotomalala H. F., "Classification des mesures des règles d'association selon CHIC-M_(GK)," Actes des 25^(ème) Rencontres de la Société Francophone et Classification - SFC 2018, 2018.

[10]. T. André, "Note sur l'implication statistique: dépendance positive orientée, valeurs critiques.," Montréal, 1994.

[11]. D. GRISSA, "Etude comportementale des mesures d'intérêt d'extraction de connaissances," 2013.

[12]. C. R. GUYNO, AMBEONDAHY and A. TOTOHASINA, "Extension of certainty factor an MGK measure on several premises," AJSER, vol. 4, p. 109, 2021.

[13]. A. RAKESH, I. Tomasz and S. Arun, "Mining association rules between set of Items in Large Database," researchGate, 1993.

[14]. F. GUILLET, "Qualité, Fouille et Gestion de Connaissances," 2006.