UNE TECHNIQUE D'EXTRACTION DES RÈGLES D'ASSOCIATION QUANTITATIVES BASÉE SUR LE COEFFICIENT DE VARIATION

Josoa Michel Tovohery, André Totohasina, Daniel Rajaonasy Feno

UNE TECHNIQUE D'EXTRACTION DES RÈGLES D'ASSOCIATION QUANTITATIVES BASÉE SUR LE COEFFICIENT DE VARIATION

Josoa Michel Tovohery[1], André Totohasina[2], Daniel Rajaonasy Feno[3]

A METHOD BASED ON THE COEFFICIENT OF VARIATION FOR MINING QUANTITATIVE ASSOCIATION RULES

RÉSUMÉ

Notre papier propose d'extraire des règles d'association quantitatives de type $(X = \overline{X}) \Rightarrow (Y = \overline{Y})$, à partir d'une grande quantité de données, où $\overline{X}$ et $\overline{Y}$ sont respectivement les moyennes arithmétiques des variables quantitatives X et Y. Pour ce faire, nous présentons une technique basée sur l'usage des mesures purement statistiques où le coefficient de variation de Sharma joue le rôle de « Support » dans le cas qualitatif, et le rapport de corrélation comme mesures d'intérêt des règles d'association quantitatives. Sans faire de partitionnement des valeurs en plusieurs intervalles et en ne considérant que les valeurs brutes, notre méthode minimise ainsi la perte d'information.

Mots-clés : *Coefficient de variation, rapport de corrélation, règles d'association quantitatives.*

ABSTRACT

Our paper proposes to extract quantitative association rules of type $(X = \overline{X}) \Rightarrow (Y = \overline{Y})$, from a large amount of data, where $\overline{X}$ and $\overline{Y}$ are respectively the arithmetic means of quantitative variables X and Y, respectively. To do this, we present a technique based on the use of purely statistical measures where the Sharma coefficient of variation plays the role of "Support" in the qualitative case, and the correlation ratio as the measures of interest of the quantitative association rules. Without partitioning the values into several intervals and by considering only the gross values, our method minimizes the loss of information.

Keywords : *Coefficient of variation, correlation ratio, Quantitative association rule.*

1 Introduction

La fouille des règles d'association qualitative est un concept introduit par Agrawal *et al.* (1993). Ces chercheurs pionniers ont proposé d'étudier une base de données contenant des transactions d'un grand supermarché américain, afin d'aider le gérant de ce dernier à étudier les comportements de ses clients et à ranger la présentation des articles sur les étagères du supermarché pour avoir le maximum de profit en un temps record.

Le problème de la fouille des règles d'association 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 A l'ensemble de tous les articles dans le supermarché et X une partie de A : 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 dans X. Selon la définition donnée par Agrawal et ses collaborateurs, on appelle règle d'association une implication partielle de type $X \Rightarrow Y$,

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

[2] ENSET- Université d'Antsiranana, andre.totohasina@gmail.com

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

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 : on écrit $s = Supp(X \Rightarrow Y)$. La Confiance « c » de la règle $X \Rightarrow Y$ est le support de la règle $X \Rightarrow Y$ divisé par le support de X : en fait, c'est la probabilité conditionnelle sachant X de Y. On dit alors : « si X, alors Y », avec une confiance de $c \times 100\%$. Le Support et la Confiance sont souvent exprimés en pourcentage et concernent essentiellement l'analyse d'information contenue dans un contexte binaire ou qualitatif.

L'extraction des règles d'association quantitatives a été introduite pour la première fois par Gregory Piatetsky-Shapiro en 1991 selon Srikant et Agrawal (1996). L'auteur se propose implicitement d'extraire dans un grand volume de données une règle d'association quantitative de deux types : Tout d'abord, si $X = a$ alors $Y = b$, qui est notée comme $(X = a) \Rightarrow (Y = b)$, où a et b sont respectivement des valeurs particulières des attributs quantitatives X et Y. On peut prendre, par exemple, la règle fictive $(\text{Âge} = 30) \Rightarrow (\text{Nombre d}'\text{enfants} = 2)$. Ensuite, G. Piatetsky-Shapiro a aussi proposé d'extraire des règles de type $(X \in [a, b]) \Rightarrow (Y \in [c, d])$ pour les attributs dont les valeurs sont très nombreuses et éparpillées dans un intervalle (type continu). Le travail de Sujatha et Naveen (2011) illustre un exemple des travaux de Piatetsky-Shapiro pour le cas $(X \in [a, b]) \Rightarrow (Y \in [c, d])$, sans combiner les intervalles adjacents.

Plus tard, cette dernière approche a été reprise par Srikant et Agrawal (1996). Ils ont élaboré une démarche très sophistiquée afin de combiner les intervalles adjacents pour résoudre les deux problèmes que rencontrent la méthode de G. Piatetsky-Shapiro : problème de MinSupp et le problème de MinConf. C'est-à-dire, plus on a trop d'intervalles, plus les intervalles contiennent moins des valeurs et la plupart des règles ont des supports très faibles. Ensuite, plus on a peu d'intervalles, plus la confiance des règles est faible. Cette approche est combinée avec leur fameux algorithme « Apriori » découvert en 1994 dans Agrawal et Srikant (1994).

Aumann et Lindell (1999) présentent deux nouveaux types de règles d'association quantitatives qui sont plutôt basés sur une description d'un sous-ensemble de la population. Ces auteurs sont les premiers à utiliser les mesures de la statistique descriptive comme moyenne, variance et médiane dans le processus d'extraction des règles d'association. Ils ont proposé d'extraire dans un grand volume de données des règles de type $\text{Sous} - \text{ensemble de la population X} \Rightarrow \text{valeur moyenne de Y}$ en deux manières : premièrement, ils définissent une règle de type $X\,(\text{ensemble d}'\text{attributs qualitatifs}) \Rightarrow (Y = \overline{Y})$, par exemple : (Non-fumeur ET buveur du vin) ⇒ (Espérance de Vie = 85). Deuxièmement, ils définissent une règle de type $(X \in [a, b]) \Rightarrow (Y = \overline{Y})$. Citons, par exemple, les deux règles suivantes qu'ils ont trouvés : $(\text{Education} \in [2\,;13]\ \text{ans}) \Rightarrow (\text{Revenu moyen} = 7.32\ \$\ \text{par heure})$, tandis que $(\text{Education} \in [14\,;18]\ \text{ans}) \Rightarrow (\text{Revenu moyen} = 11.64\ \$\ \text{par heure})$. La notion de règle maximale a été inventée afin de trouver l'intervalle [a ; b] maximisant la moyenne $\overline{Y}$ de l'attribut quantitatif Y. Geoffrey I. Webb a essayé d'améliorer ce dernier type de règle proposé par Aumann et Lindell (1999) dans Webb (2001) en utilisant son algorithme OPUS et en introduisant une autre mesure de sélection des k règles plus intéressantes.

En tenant compte de ces quatre travaux pionniers, nous proposons d'extraire dans un grand volume de données quantitatives des règles d'association de type $(X = \overline{X}) \Rightarrow (Y = \overline{Y})$, qu'on interprète comme suit : « En général, si les clients achètent une quantité $\overline{X}$ de X, alors ils achèteraient aussi une quantité $\overline{Y}$ de Y ». On a une information générale sur

les comportements des clients, car une valeur moyenne $\overline{X}$ d'une variable quantitative X s'interprète comme une valeur générale représentant X, si X est homogène. Donc, afin de sélectionner des règle de tel type, il est évident qu'on doit se poser d'abord une contrainte de dispersion des valeurs de tous les attributs quantitatifs étudiés. Pour ce faire, nous proposons d'utiliser la mesure de dispersion normalisée et symétrique de Sharma *et al.*(2011), notée CVS, et on laisse à l'utilisateur de spécifier son seuil MaxCVS afin d'élaguer les attributs quantitatifs dont les valeurs sont assez dispersées (donc non homogènes). Puis, nous utilisons le rapport de corrélation, afin de quantifier la dépendance fonctionnelle et statistique entre deux attributs quantitatifs à la place des mesures de qualité des règles d'association.

Le reste de cet article est organisé comme suit : la section 2 présentera le détail de notre problème, notre technique d'extraction des règles d'association quantitative, les justifications de notre approche et les méthodologies des expérimentations effectuées ; la section 3 partagera les résultats trouvés lors des traitements des données réelles effectués ; la section 4 discutera les avantages, les limites de notre approche et les liaisons de ce présent travail avec les autres résultats déjà publiés dans la littérature ; la section 5 termine notre exposé par une conclusion.

2 Méthodes

Dans cette section, nous présentons notre approche et les données que nous allons traiter à titre d'expérimentation. Dans cette sous-section, nous allons présenter notre problème, notre approche lors de la résolution et sa justification.

2.1 Position de problème

Une fois qu'un client vienne dans un supermarché, il achète certains articles avec leurs unités de mesures respectives, que ceux – ci soient des nombres ou masses ou longueurs ou volumes, etc. Afin de bien gérer le calcul de bénéfice et le nombre de stock restant, il est important de mémoriser les transactions effectuées. Dans ce cas, on est obligé de stocker des nombres réels. À titre indicatif, nous voyons dans le Tableau 1 un exemple pratique de cet enregistrement (Tableau 1).

N° de la transaction	Huiles (litres)	Sucres (kg)	Savons (unité)	Tomates (unité)	Oeufs (unité)
1	1	0	0	3	4
2	0.5	0.5	0	5	0
3	0	1	3	0	1
4	1.5	1	2	2	1
5	2	2	0	3	3
6	0	2	0	3	2
7	1	1.5	1	0	0
8	0.75	0	2	1	3
9	1	0	3	0	2
10	1	0	1	3	2

Tableau 1 – Exemple de tableau d'enregistrement des transactions (Données fictives)

Afin d'avoir un maximum de profit, le gérant du supermarché s'intéresse, à la fois, à la découverte des articles les plus achetés par les clients (problème de la recherche d'itemsets fréquents), à la consommation moyenne des clients pour chaque article (problème de la statistique descriptive), à l'homogénéité des consommations des clients, à la connaissance des groupes d'articles que les clients achètent souvent ensemble afin de mieux ranger

l'emplacement des articles sur les étagères, à la connaissance de tous les articles courant un risque lorsqu'on cesse de vendre un article X, et surtout à la connaissance de la quantité moyenne u des articles Y vendus par un client sachant qu'il a déjà acheté une quantité de v unités de l'article X, etc.

Dans notre étude, nous proposons d'extraire des règles de type $(X = \overline{\mathrm{X}}) \Rightarrow (\mathrm{Y} = \overline{\mathrm{Y}})$ afin de répondre à ces besoins. Par exemple, supposons qu'on ait la règle d'association R1 = $(\mathrm{Tomate} = 3\ \mathrm{unités}) \Rightarrow (\mathrm{œufs} = 2\ \mathrm{unités})$. On peut dire directement que, en générale, si les clients achètent 3 Tomates, alors ils achèteraient 2 œufs. Ainsi, on a tout de suite les consommations moyennes des articles présents dans une règle valide. D'une part selon notre approche, si X et Y apparaissent dans une règle, alors ce sont tous des articles que les clients consomment souvent. De plus, un motif ne serait pas sélectionné si ces valeurs ne satisfait pas à la contrainte de dispersion MaxCVS, donc chaque motif d'une règle valide possède des valeurs homogènes selon la spécification de l'utilisateur. Ensuite, cette règle nous informe qu'il faut mettre l'étagère contenant les « Tomates » à côté de celui des « œufs », car il devient aussi plus probable que le client qui achète des tomates, achètera aussi une certaine quantité des œufs. Enfin, nous connaissons à partir de cette règle que si l'on arrête de vendre des « tomates », alors la vente « d'œufs » risquerait un problème, car les clients ont tendance d'acheter ailleurs, là où il y a tout pour économiser le temps d'achat et les dépenses en déplacement. D'où l'intérêt d'extraire des règles d'association quantitatives de type $(X = \overline{\mathrm{X}}) \Rightarrow (\mathrm{Y} = \overline{\mathrm{Y}})$.

2.2 Notre approche

Dans cette section, nous allons donner la présentation formelle de notre problème et les étapes algorithmiques de résolution. De plus, nous allons nous servir du Tableau 1 pour présenter un exemple pratique.

Soit D un tableau des données numériques, tel que les lignes représentent les achats que fait un client, et les colonnes représentent les articles vendus par le supermarché. Le Tableau 2 illustre la forme du tableau D (Tableau 2).

N° de la transaction	Article 1 : X_1	Article 2 : X_2	⋯	Article j : X_j	⋯	Article m : X_m
t_1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1m}
t_2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2m}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
t_i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{im}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
t_n	x_{n1}	x_{n2}	⋯	x_{nj}	⋯	x_{nm}

Tableau 2 – Le tableau des données numériques ***D***

La valeur x_{ij} représente ici la quantité d'article X_j achetée lors de la transaction t_i. Les étapes suivantes montrent comment faire pour traiter ce tableau numérique D afin de découvrir les comportements généraux des clients du supermarché propriétaire de ces données.

2.2.1 Construction d'une matrice intermédiaire

C'est une construction d'une matrice booléenne afin d'élaguer dans notre étude les articles qui n'intéressent pas trop les clients, c'est-à-dire les articles qui se vendent plutôt rarement, car un tel article intéresserait seulement un groupe particulier des clients, mais non pas la généralité. Pour ce faire, on utilise la fonction booléenne suivante :

$$\forall\, x_{ij} \in D,\ Bool(x_{ij}) = \begin{cases} 0, si\ x_{ij} = 0\ ; \\ 1, sinon. \end{cases} \qquad (1)$$

Après cette transformation intermédiaire, on a le tableau D', tel que si $x_{ij} \in D'$, alors $x_{ij} = 0$ ou $x_{ij} = 1$ (Tableau 3).

N° de la transaction	Huiles (litres)	Sucres (kg)	Savons (unité)	Tomates (kg)	Oeufs (unité)
1	1	0	0	1	1
2	1	1	0	1	0
3	0	1	1	0	1
4	1	1	1	1	1
5	1	1	0	1	1
6	0	1	0	1	1
7	1	1	1	0	0
8	1	0	1	1	1
9	1	0	1	0	1
10	1	0	1	1	1

Tableau 3 – Tableau booléen obtenu à partir du Tableau 1

2.2.2 Usage de la matrice intermédiaire *D'*

Nous faisons usage de la matrice intermédiaire D' pour le calcul de $support$ de X_j , $j = \overline{1, m}$. À partir de D', on a :

$$Support\ (X_j) = \frac{\sum_{i=1}^{n} bool(x_{ij})}{n}. \qquad (2)$$

Le Tableau 4 illustre les supports des articles présentés dans le Tableau 1.

Article	Huiles	Sucres	Savons	Tomates	Œufs
$Support$	0.8	0.6	0.6	0.7	0.8

Tableau 4 – Support des articles figurant dans le Tableau 1

2.2.3 Premier élagage :

Nous procédons à l'élagage des articles vendus rarement ($itemests$ non fréquents) à partir de la contrainte de $MinSup$. Ici, on garde seulement les articles (1-itemset) dont la fréquence de vente est plus grande que le paramètre $MinSup$ fixé par l'utilisateur. Donc, l'article X_j est retenu si $Support(X_j) > MinSup$.

Soit $DF1$ l'ensemble des articles X_j dont la fréquence de vente $Support(X)$ est plus grand que $MinSup$. Exemple, pour $MinSup = 0.6$, les articles « Sucres » et « Savons » sont élagués. Nous revenons à D mais seulement avec les éléments de $DF1$. Nous procédons au calcul des coefficients de variation de Sharma CVS des $1 - itemsets\ X_j$ fréquents. Soient $\forall\, X_j \in DF1,\ X'_j = \{x_{ij} \in X_j \,|x_{ij} \neq 0\}$ et $n_j = Card(X'_j)$. On note par x'_{ij} les éléments de X'_j. Le coefficient de variation de Sharma *et al.* (2011) d'une variable quantitative X_j est :

$$CVS(X_j) = \begin{cases} \dfrac{\sigma(X'_j)}{\sqrt{[Max(X_{j'}) - \overline{X_{j'}}]\,[\overline{X_{j'}} - Min(X_{j'})]}}, si\ Max(X'_j) \neq \overline{X'_j}\ et\ Min(X'_j) \neq \overline{X'_j}; \\ 0, sinon. \end{cases} \qquad (3)$$

avec $\overline{X_j'} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}{}'$ et $\sigma(X_j') = \sqrt{\frac{1}{n_j}\sum_{i=1}^{n_j}(x'_{ij} - \overline{X_j'})^2}$.

On rappelle que, dans notre approche, le calcul de $CVS(X_j)$ et la moyenne $\overline{X_j'}$ est appliqué seulement sur les valeurs x_{ij} non nulles de X_j. Le Tableau 5 montre les CVS des articles satisfaisants le $MinSupp = 0.6$ (Tableau 5).

Items fréquents	Huiles	Tomates	Œufs
CVS	0.5887635	0.5638694	0.6546537

Tableau 5 – CVS des articles les plus consommés par les clients.

2.2.4 Deuxième élagage

Nous procédons à deuxième élagage des $itemsets$ fréquents dont les valeurs sont très dispersées en utilisant le paramètre $MaxCVS$. L'article X_j est retenu si $CVS(X_j)$ est plus grand que $MaxCVS$. Soit $DF2$ l'ensemble des articles fréquents et dont les consommations des clients sont homogènes. Exemple : pour $MaxCVS = 0.7$, on a : $DF2 = \{Huiles ; Tomates ; Oeufs\}$.

2.2.5 Retour dans *D* et troisième élagage.

Calcul des rapports de corrélation $\eta_X(Y)$ et $\eta_Y(X)$ pour tout X,$Y \in DF2$, tels que $X \cap Y = \emptyset$. En même temps, retenir la règle : $X \Rightarrow Y$ si $\eta_X(Y) > \eta_Y(X)$. Si $\eta_X(Y) = \eta_Y(X)$, alors on a une implication de type $X \Leftrightarrow Y$. Nous allons approfondir cette étape avec une théorie formelle et un exemple de calcul détaillé lors de la justification de ce choix de mesure.

Remarque : Lors du calcul de $\eta_Y(X)$ et $\eta_X(Y)$, on utilise toutes les valeurs de X et Y, sans exception (contrairement au calcul de CVS à l'étape 4).

2.2.6 Quatrième élagage

On peut faire deux types d'élagage selon le critère choisi par l'utilisateur :

(a) Elagage des règles $X \Rightarrow Y$ dont le rapport corrélation $\eta_X(Y) < MinRapCor$, avec $MinRapCor$ est le seuil minimum de rapport de corrélation (voir page 11 et 12).
(b) Retenir les r tops règles d'association intéressantes, c'est-à-dire les r premières règles possédant le rapport de corrélation le plus élevé.

On procède en suite à l'affichage du résultat en tenant compte de la forme $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$, et Fin.

La Figure 1 illustre les règles d'association de type $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$ obtenues par les données du Tableau 1 pour $MinSupp = 0.6$, $MaxCVS = 0.7$ et $MinRapCor = 0.4$.

```
> Resultat
                 Premisse Implication          Consequence Rapport.Correlation
1    Tomates : moyenne = 2          => Oeufs : moyenne = 1.8          0.6955128
2 Huiles : moyenne = 0.875          => Tomates : moyenne = 2          0.4807692
3 Huiles : moyenne = 0.875          => Oeufs : moyenne = 1.8          0.4551282
>
```

Figure 1 – Résultat du traitement des données du Tableau 1 pour $MinSupp = 0.6$, $MaxCVS = 0.7$ et $MinRapCor = 0.4$.

2.3 Justifications de notre approche

Tout d'abord, nous avons décidé de traiter les valeurs réelles car la transformation booléenne fait perdre des informations, alors qu'en statistique on s'intéresse plutôt au maximum d'information.

Quantité d'information — La quantité d'information I contenue dans le tableau de contingence de X et Y est égale à la somme de la variance de X et de la variance de Y (trace de la matrice de variance – covariance) (Totohasina (2008)). On a :

$$I = V(X) + V(Y) \qquad (4)$$

En effet, pour deux variables aléatoires réelles X et Y prenant respectivement des valeurs positives ou nulles dans $X(\Omega) = \{x_1, \dots, x_n\}$ et $Y(\Omega) = \{y_1, \dots, y_m\}$; $n_{i\,.}$ et $n_{.j}$ étant les effectifs marginaux respectifs, on a la quantité d'information :

$$I_{v.réelle} = \frac{1}{N}\left(\sum_{i=1}^{n} n_{i\,.}.x_i^2 + \sum_{i=1}^{m} n_{.j}.y_j^2\right) - \frac{1}{N^2}\left[\left(\sum_{i=1}^{n} n_{i\,.}.x_i\right)^2 + \left(\sum_{i=1}^{m} n_{.j}.y_j\right)^2\right] \qquad (5)$$

où $N = \sum_i^n n_{i\,.} = \sum_j^m n_{.j}$.

Toutefois, si X et Y sont binaires, alors la quantité d'information se réduit à :

$$I_{v.binaire} = \frac{n_X + n_Y}{N} - \frac{n_X^2 + n_Y^2}{N^2} \qquad (6)$$

Cependant, on a : $n_X \leq \sum_{i=1}^{n} n_{i\,.}.{x_i}^2$ et $n_X^2 \leq \left(\sum_{i=1}^{n} n_{i\,.}.x_i\right)^2$. Donc, on a : $\frac{n_X - n_X^2}{N} \leq \frac{1}{N}\sum_{i=1}^{n} n_{i\,.}.{x_i}^2 - \left(\frac{1}{N}\sum_{i=1}^{n} n_{i\,.}.x_i\right)^2$. C'est-à-dire, $V(X_{binaire}) \leq V(X_{réelle})$. D'où, on a : $I_{v.binaire} \leq I_{v.réelle}$.

Lemme — On perd généralement de l'information en réduisant deux variables effectivement à valeurs réelles positives ou nulles, de moyenne marginales strictement supérieures à 1, en deux variables binaires (Totohasina (2008)).

Ensuite, on a choisi le coefficient de variation à la place du support, car si l'on considère le cas binaire où le support $Supp(X) = f_X = \frac{n_X}{n}$, alors le coefficient de variation de Pearson est : $CV(X) = \frac{\sqrt{f_X(1-f_X)}}{f_X}$. On a donc la relation suivante :

$$CV(X) < \alpha \Leftrightarrow f_X > \frac{1}{1+\alpha^2} \qquad (7)$$

pour tout nombre réel positif α.

Par exemple, si $CV(X) < 1$, alors $f_X > 0.5$. Ainsi, la contrainte de maximum de coefficient de variation représente la contrainte de $MinSupp$ dans le cas binaire.

Par la suite, nous avons choisi de remplacer le coefficient de variation de Pearson par le coefficient de variation de Sharma, car ce dernier est la meilleure mesure de dispersion (Sharma *et al.* (2011)). En effet, à l'instar des travaux de Lallich et Teytaud (2004) et les travaux de Grissa (2013) sur les critères d'appréciation des mesures de qualité des règles d'association, nous proposons d'énumérer d'abord les critères que les chercheurs souhaitent avoir sur une mesure de dispersion, puis nous démontrons que la mesure CVS est la meilleure mesure de dispersion de la littérature.

2.3.1 Quelques critères d'appréciation d'une mesure de dispersion :

Mesure possédant un sens concret — Selon Lallich et Teytaud (2004), il est important de travailler avec une mesure possédant un sens parlant pour l'utilisateur. On a : pour toute

mesure de dispersion M, $C_1(M) = 1$ si la mesure M possède un sens concret ; sinon $C_1(M) = 0$.

Mesure dépourvue d'unité — Afin que la comparaison des dispersions des deux variables aléatoire X et Y par la mesure M ait du sens, il faut que la valeur fournie par M soit dépourvue d'unité. Ainsi, on adopte la règle suivante : $C_2(M) = 1$, si $M(X)$ n'admet pas d'unité ; sinon, $C_2(M) = 0$ (Sørensen (2002)).

Mesure normalisée — Une mesure de dispersion est dite normalisée si sa valeur est comprise entre 0 et 1. Totohasina (2008) a affirmé que la normalisation des mesures des règles d'association consiste à ramener les valeurs d'une mesure de qualité sur l'intervalle $[-1\ ;\ 1]$. Cependant, toutes les mesures de dispersion existantes dans la littérature sont des mesures positives. Ainsi, la normalisation d'une mesure de dispersion consiste à ramener la valeur de cette dernière sur l'intervalle $[0\ ;\ 1]$. Par conséquent, on a : pour toute mesure M, $C_3(M) = 1$ si pour toute variable aléatoire réelle X, $0 \leq M(X) \leq 1$; sinon, $C_3(M) = 0$.

Mesure donnant une valeur nulle pour une variable aléatoire à valeurs constantes — Une mesure de dispersion doit être capable de détecter au moins le cas d'une parfaite stabilité. Par conséquent, $C_4(M) = 1$, si pour toute variable aléatoire réelle constante X, $M(X) = 0$; sinon $C_4(M) = 0$. C'est-à-dire, on souhaite que $M(X) = 0$, si pour tout échantillon de n valeurs de X, on a : $x_1 = x_2 = \cdots = x_n$.

Mesure stable par une translation linéaire — Une mesure de dispersion doit évaluer par une même valeur la variable X et la variable $Y = X + a$, où a est une constante réelle. Dès lors, on a : $C_5(M) = 1$, si $\forall\ a \in R, M(X + a) = M(X)$; sinon, $C_5(M) = 0$ Sørensen (2002).

Mesure stable par une multiplication avec un scalaire — Dès fois, en démographie, on est obligé de changer les unités des valeurs obtenues. Citons par exemple, la conversion des unités des monnaies en dollar. Cependant, on veut que la conversion des unités n'affecte pas la dispersion des valeurs. Ainsi, on veut que la mesure de dispersion soit stable par une multiplication par un scalaire : $C_6(M) = 1, si\ \forall\ k \in R^*, M(kX) = M(X)$; sinon $C_6(X) = 0$ (Sørensen (2002)).

Mesure retournant deux valeurs identiques pour deux séries à valeurs symétriques — Selon Sharma *et al.* (2011), une mesure de dispersion ne doit pas favoriser les séries à faibles valeurs ou celles qui ont des valeurs plus élevées. Ainsi, pour évaluer ce critère, on teste la mesure avec plusieurs couples de séries à valeurs symétriques. Prenons, par exemple, le cas des deux séries proposées par Sharma suivantes : $S_1 = \{1\ ;3\ ;11\}$ et $S_2 = \{1\ ;9\ ;11\}$. Ces deux séries ont des valeurs symétriques par rapport à $x = 6$. Une bonne mesure de dispersion doit renvoie une même valeur pour chacune de ces séries. Par conséquent, on a : $C_7(M) = 1$, si M renvoie deux valeurs identiques pour deux séries à valeurs symétriques ; sinon, $C_7(M) = 0$.

Mesure pouvant être utilisée avec des valeurs négatives et positives — Une telle mesure peut être utilisée pour évaluer la variation de toutes les variables aléatoires réelles. En effet, bien qu'une mesure soit jugée performante, le fait qu'elle pose de restriction sur le signe des valeurs engendre un grand problème sur son usage. Ainsi, on souhaite travailler avec une mesure de dispersion définie pour tous les signes des valeurs. On a donc : $C_8(M) = 1$, si la mesure M est définie pour des valeurs négatives et positives ; sinon, $C_8(M) = 0$.

Actuellement, on compte déjà onze mesures de dispersion. Le tableau 6 suivant illustre leurs appellations avec leurs expressions.

N°	Mesures de dispersion	Expressions
1	Variance	$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$
2	Ecart-type	$S(X) = \sqrt{V(X)}$
3	Étendu ou Empan	$Empan(X) = \max_{1\leq i\leq n} x_i - \min_{1\leq i\leq n} x_i$
4	Ecart moyen	$E.M(X) = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - \bar{X}\rvert$
5	Ecart médian	$E.med(X) = \frac{1}{n}\sum_{i=1}^{n}\lvert x_i - Med(X)\rvert$
6	Ecart géométrique $E.géom$	$\log[E.géom(X)] = \frac{1}{n}\sum_{i=1}^{n}(\log x_i - \log[G(X)])^2$ avec $G(X) = \sqrt[n]{\prod_{i=1}^{n} x_i}$
7	Intervalle interquartile	$IQ(X) = Q_3 - Q_1$ où Q_1 et Q_3 sont respectivement les quartiles 25% et 75%.
8	Différence moyenne	$d(X) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lvert x_i - x_j\rvert$
9	Coefficient de variation relative	$Q_r(X) = \frac{Q_3 - Q_1}{Med(X)}$ (1) ou $Q_r(X) = \frac{Q_3 - Q_1}{Q_3 + Q_1}$ (2)
10	Coefficient de variation de variation (de Pearson)	$CV(X) = \frac{S(X)}{\bar{X}}$
11	Coefficient de variation de Sharma	$CVS(X) = \begin{cases} \frac{s(X)}{\sqrt{[Max(X) - \bar{X}\,][\bar{X} - Min(X)]}} \\ si\ Max(X) \neq \bar{X} et\ Min(X) \neq \bar{X}\ ; \\ 0, \quad sinon. \end{cases}$

Tableau 6 – Liste des mesures de dispersion existantes dans la littérature

Le tableau 7 suivant montre les propriétés des onze mesures de dispersion recensées dans la littérature. Nous avons exclu le critère 1, car cela est jugé trop subjectif.

	C_2	C_3	C_4	C_5	C_6	C_7	C_8	Total
$V(X)$	0	0	1	1	0	1	1	4
$S(X)$	0	0	1	1	0	1	1	4
$Empan(X)$	0	0	1	1	0	1	1	4
$E.M(X)$	0	0	1	1	0	1	1	4
$E.med(X)$	0	0	1	1	0	1	1	4
$E.géom(X)$	0	0	1	0	1	0	0	2
$IQ(X)$	0	0	1	1	0	1	1	4
$d(X)$	0	0	1	1	0	1	1	4
$Q_r(X)$ (1)	1	0	1	0	1	0	1	4
$Q_r(X)$ (2)	1	1	0	1	1	0	1	5
$CV(X)$	1	0	1	0	1	0	0	3
$CVS(X)$	1	1	1	1	1	1	1	7

Tableau 7 – évaluation des 11 mesures de dispersion existantes dans la littérature.

Ce tableau montre que le coefficient de variation de Sharma est la meilleure mesure de dispersion, car il est le seul vérifiant tous les critères.

Remarques — La mesure $CVS(X)$ prend exactement ses valeurs entre $[0\,;1]$. $CVS(X)$ prend une valeur nulle pour une variable aléatoire constante X. Elle prend une valeur égale à 1 pour une série contenant seulement deux valeurs différentes x et y, mais de même effectif. Prenons, par exemple, le cas de la série $X = \{1\,;\,1\,;\,1\,;\,5\,;\,5\}$. Dans ce cas, on a : $CVS(X) = 1$. La deuxième expression du coefficient de variation relative prend une valeur égale à 1, si et seulement si $Q_1 \neq Q_3$ et $Q_1 = 0$.

Enfin, nous allons énumérer les belles propriétés de la mesure rapport de corrélation afin de justifier le choix de notre mesure d'intérêt d'une règle quantitative valide.

Rapport de corrélation — On considère deux variables aléatoires X et Y. Soient $X = \{x_1\,;\,x_2\,;\ldots\,;\,x_r\}$ et $Y = \{y_1\,;\,y_2\,;\ldots\,;\,y_s\}$. Ainsi, on a le tableau de contingence des effectifs illustrés par le tableau 8 suivant :

Y / X	y_1	y_2	…	y_j	…	y_s	TOTAL
x_1	$n_{1,1}$	$n_{1,2}$	…	$n_{1,j}$	…	$n_{1,s}$	$n_{1,.}$
x_2	$n_{2,1}$	$n_{2,2}$	…	$n_{2,j}$	…	$n_{2,s}$	$n_{2,.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
x_i	$n_{i,1}$	$n_{i,2}$	…	$n_{i,j}$	…	$n_{i,s}$	$n_{i,.}$
⋮	⋮	⋮	…	⋮	…	⋮	⋮
x_r	$n_{r,1}$	$n_{r,2}$	…	$n_{r,j}$	…	$n_{r,s}$	$n_{r,.}$
TOTAL	$n_{.,1}$	$n_{.,2}$	…	$n_{.,j}$	…	$n_{.,s}$	N

Tableau 8 – Tableau de contingence d'effectifs de X et Y

(a) On définit les moyennes marginales de X et Y par :

$$\mu(X) = \bar{X} = \frac{1}{N}\sum_{i=1}^{r} n_{i,.}\, x_i \quad et \quad \mu(Y) = \bar{Y} = \frac{1}{N}\sum_{j=1}^{s} n_{.,j}\, y_j. \qquad (8)$$

(b) On définit la variance marginale de X et Y par :

$$\sigma^2(X) = \frac{1}{N}\sum_{i=1}^{r} n_{i,.}\,(x_i - \bar{X})^2 \quad et \quad \sigma^2(Y) = \frac{1}{N}\sum_{i=1}^{s} n_{.,j}\left(y_j - \bar{Y}\right)^2. \qquad (9)$$

(c) On définit la moyenne conditionnelle sachant $(X = x_i)$ de Y par :

$$\mu_{(X=x_i)}(Y) = \sum_{j=1}^{s} \frac{n_{i,j}}{n_{i,.}} y_j. \qquad (10)$$

(d) On définit la variance conditionnelle sachant $(X = x_i)$ de Y par :

$$\sigma^2{}_{(X=x_i)}(Y) = \sum_{j=1}^{s} \frac{n_{i,j}}{n_{i,.}} \left(y_j - \mu_{(X=x_i)}(Y)\right)^2. \qquad (11)$$

Remarques — $\mu_X(Y)$ et $\sigma^2{}_X(Y)$ sont des variables aléatoires sur $X(\Omega) = \{x_1\,;\ldots\,;\,x_r\}$. On a :

$$\mu_{X(\Omega)}(Y) = \left\{\mu_{(X=x_1)}(Y)\,;\ldots;\,\mu_{(X=x_r)}(Y)\right\} \text{ et } \sigma^2{}_{X(\Omega)}(Y) = \left\{\sigma^2{}_{(X=x_1)}(Y)\,;\ldots\,;\,\sigma^2{}_{(X=x_r)}(Y)\right\}$$

De plus, l'application $\mu_X(Y) : (X = x_k) \mapsto \mu_{(X=x_k)}(Y) = h(x_k)$ est la régression de Y en X. On estime Y par : $Y \simeq h(X)$. Dans ce cas, h n'est pas nécessairement linéaire.

(e) **Théorème de la variance totale —** Soient X et Y deux variables aléatoires réelles, telles que $\sigma^2(X)$ et $\sigma^2(Y)$ existent. On a alors:
$\sigma^2(Y) = \mu\left(\sigma^2{}_X(Y)\right) + \sigma^2\left(\mu_X(Y)\right)$.

Remarques — Si $\sigma^2(Y)$ est non nul, alors on a :

$$\frac{\mu\left(\sigma^2{}_X(Y)\right)}{\sigma^2(Y)} + \frac{\sigma^2(\mu_X(Y))}{\sigma^2(Y)} = 1. \qquad (12)$$

(f) On définit le rapport de corrélation de Y en X par le nombre réel positif ou nul $\eta_X(Y)$ défini par son carré :

$$\eta^2{}_X(Y) = \frac{\sigma^2(\mu_X(Y))}{\sigma^2(Y)}. \qquad (13)$$

$\eta^2{}_X(Y) \times 100\%$ est le pourcentage de la variance totale expliquée par l'éventuelle régression de Y en X.

(g) On définit le coefficient de liberté de Y par rapport à X par le nombre réel positif ou nul $L_X(Y)$ défini par son carré :

$$L^2{}_X(Y) = \frac{\mu\left(\sigma^2{}_X(Y)\right)}{\sigma^2(Y)}. \qquad (14)$$

$L^2{}_X(Y)$ exprime ainsi la proportion de la variance marginale de Y non expliquée par l'éventuelle régression de Y en X.

Propositions — Voici quelques propriétés évidentes reprises dans les travaux de Totohasina (2008) pages 80-81 :

1. $0 \leq \eta_X(Y) \leq 1$ et $0 \leq L_X(Y) \leq 1$ (mesure normalisée).

2. En générale, on a : $\eta_X(Y) \neq \eta_Y(X)$ et $L_X(Y) \neq L_Y(X)$. C'est-à-dire, ces deux mesures ne sont pas symétriques dans les éventuelles dépendances ou indépendances.

3. $\eta_X(Y) = 1$ $(ou\ L_X(Y) = 0) \Leftrightarrow$ (Il existe une dépendance fonctionnelle continue de Y relativement à X). Dans ce cas, on dit que Y est totalement dépendante de X. Cette relation de dépendance n'est pas nécessairement linéaire.

4. $L_X(Y) = 1$ $(ou\ \eta_X(Y) = 0)$ n'entraîne pas nécessairement que Y soit indépendante de X. On dit simplement que Y est totalement libre de X. C'est-à-dire, la connaissance de X n'apporte aucune information sur Y.

5. En général, le rapport de corrélation de Y en X est toujours supérieur ou égal à la valeur absolue du coefficient de corrélation linéaire de Y et X : $\eta_X(Y) \geq |corrélation(X,Y)|$.

6. Si $\eta_X(Y) = |corrélation(X,Y)|$ ou $\eta_X(Y) \approx |corrélation(X,Y)|$, alors il y a régression linéaire ou quasi-linéaire de Y en X.

7. S'il existe deux réels $a\ et\ b$, tels que $Y = aX + b$ et a non nul, alors $\eta_X(Y) = |corrélation(X,Y)| = 1$.

Selon ces propriétés, le rapport de corrélation, qui n'est pas symétrique, est meilleur que le coefficient de corrélation, car il détecte non seulement la dépendance linéaire, mais toute autre dépendance fonctionnelle possible. D'où le choix de la mesure rapport de

corrélation lors de l'extraction des règles d'association quantitatives (sans transformation en valeur booléenne).

Le Tableau 9 montre les valeurs des $\eta_{(X=x_i)}(Y)$ et $\eta_{(Y=y_i)}(X)$ pour $X =$ »$Tomates$» et $Y =$ »$Oeufs$».

$Y = Oeuf$ / $X = Tomates$	$y_1 = 0$	$y_2 = 1$	$y_3 = 2$	$y_4 = 3$	$y_5 = 4$	Total	$\eta_{(X=x_i)}(Y)$
$x_1 = 0$	1	1	1	0	0	3	1
$x_2 = 1$	0	0	0	1	0	1	3
$x_3 = 2$	0	1	0	0	0	1	1
$x_4 = 3$	0	0	2	1	1	4	2.75
$x_5 = 5$	1	0	0	0	0	1	0
Total	2	2	3	2	1	10	-
$\eta_{(Y=y_i)}(X)$	2.5	1	2	2	3	-	-

Tableau 9 – Valeur des $\eta_{(X=x_i)}(Y)$ et $\eta_{(Y=y_i)}(X)$ pour X = »Tomates» et Y = »œufs».

D'après la ligne de $(X = 0)$ du Tableau 8, on a :

$$\eta_{(X=0)}(Y) = \frac{(1 \times 0) + (1 \times 1) + (1 \times 2) + (0 \times 3) + (0 \times 4)}{1 + 1 + 1 + 0 + 0} = \frac{3}{3} = 1$$

D'après la colonne de $(Y = 2)$ du Tableau 8, on a :

$$\eta_{(Y=2)}(X) = \frac{(1 \times 0) + (0 \times 1) + (0 \times 2) + (2 \times 3) + (0 \times 5)}{1 + 0 + 0 + 2 + 0} = \frac{6}{3} = 2$$

Calcul de $\mu\left(\eta_{(X=x_i)}(Y)\right)$:

$$\mu\left(\eta_{(X=x_i)}(Y)\right) = \frac{3 \times \eta_{(X=0)}(Y) + 1 \times \eta_{(X=1)}(Y) + 1 \times \eta_{(X=2)}(Y) + 4 \times \eta_{(X=3)}(Y) + 1 \times \eta_{(X=5)}(Y)}{3 + 1 + 1 + 4 + 1}$$

On a : $\mu\left(\eta_{(X=x_i)}(Y)\right) = 1.8$, qui est vraiment égale à la consommation moyenne des œufs, car $\bar{Y} = 1.8$. Les pondération 1, 3 et 4 sont les effectifs marginaux des x_i.

Calcul de la variance $\sigma^2\left(\eta_{(X=x_i)}(Y)\right)$:

$$\sigma^2\left(\eta_{(X=x_i)}(Y)\right) = \frac{1}{N}\sum_{i=1}^{r} n_{i,.}\eta^2{}_{(X=x_i)}(Y) - \bar{Y}^2. \qquad (15)$$

Donc, on a : $\sigma^2\left(\eta_{(X=x_i)}(Y)\right) = \frac{(3\times1^2)+(1\times3^2)+(1\times1^2)+(4\times2.75^2)+(1\times0^2)}{10} - 1.8^2 = 1.085.$
Cependant, la variance de Y est : $\sigma^2(Y) = 1.56$ (calculer à partir de toutes les valeurs de Y sans exception).

Ainsi, on a le rapport de corrélation :

$$\eta_X(Y) = \frac{\sigma^2\left(\eta_{(X=x_i)}(Y)\right)}{\sigma^2(Y)} = \frac{1.085}{1.56} = 0.6955128$$

De la même façon, on a : $\eta_Y(X) = 0.1346154$.

Comme $\eta_Y(X) < \eta_X(Y)$, alors la règle $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$ est éliminée à l'étape 6 pour éviter la redondance des règles. De plus, $\eta_X(Y) > MinRapCor = 0.4$, alors la règle $\bar{X} \Rightarrow \bar{Y}$ est retenue. Ainsi, nous affichons au résultat la règle :

$(Tomate = 2) \Rightarrow (Oeufs = 1.8)$ $avec$ $\eta_X(Y) = 0.695$, valeur très élevée.

Interprétation — En générale, si les clients achètent 3 tomates, alors ils achèteront 2 œufs.

3 Matériels

Dans cet article, nous avons utilisé le logiciel R version 4.0.4 pour la programmation et les traitements de nos jeux de données. Ensuite, nous avons expérimenté notre méthode avec les notes des 2351 élèves série A2 de la région DIANA – Madagascar, candidats au baccalauréat de l'année 2014. Notre jeu de données contient la moyenne générale et les notes des matières suivantes :

- Langue : Malagasy, Français ;
- Langue vivante au choix : Anglais ou Allemand ou Russe ou Espagnol ;
- Histoire Géographie ;
- Philosophie ;
- Matières scientifiques : Mathématiques générales, Physique – chimie et Science de la vie et de la terre ;
- Education Physique et Sportive (E.P.S).

On rappelle que, selon le système éducatif de Madagascar de l'année 2014, les élèves candidats au baccalauréat sont répartis en 04 séries :

- Série A1 : pour les élèves purement littéraires (mathématiques coefficient 01, les deux autres matières SVT et Physique chimie sont facultatives) ;
- Série A2 : pour les élèves littéraires mais qui s'intéressent aux mathématiques (mathématiques coefficient 03, les deux autres matières SVT et Physique chimie sont facultatives) ;
- Série C : pour les élèves purement scientifiques, forts en mathématiques et en physique chimie.
- Série D : pour les élèves scientifiques qui maîtrisent le plus la matière SVT, et qui maîtrise peu les mathématiques et la physique – chimie.

4 Résultats

La Figure 2 illustre un résultat pour $MinSupp = 0.99$, $MaxCVS = 0.6$ et $MinRapCor = 0.2$. Tout d'abord, nous voyons que les neufs premières règles disposent comme prémisse la moyenne générale et comme conséquences les neuf matières de l'examen. Ces neufs règles forment une description générale du résultat de l'examen du baccalauréat session 2014 : en générale, les élèves ont eu une moyenne générale de $\frac{8.67}{20}$. Si l'élève a obtenu une moyenne générale de $\frac{8.67}{20}$, alors il aurait une note de 5 /20 en mathématiques, 9/20 en Histoire géographie, 11/20 en Malagasy, 10/20 en Philosophie, 6/20 en physique – chimie, 7/20 en Français, 5/20 en SVT, 5/20 en Langue vivante (que ce soit Anglais ou Espagnol ou Allemand ou Russe) et 11/20 en EPS. De plus, bien que ces élèves soient littéraires, on remarque également que leurs notes dans les matières littéraires sont faibles, à l'exception de la langue malagasy qui est la langue maternelle. Cela montre que l'élève n'est pas fort dans sa spécialité.

```
> Resultat
                        Premisse Implication                  Consequence Rapport.Correlation
1     Moyenne.G : moyenne = 8.67          =>  Mathematiques : moyenne = 5.4              0.68
2     Moyenne.G : moyenne = 8.67          =>  Histoire.geo : moyenne = 9.04              0.67
3     Moyenne.G : moyenne = 8.67          =>     Malagasy : moyenne = 11.48              0.57
4     Moyenne.G : moyenne = 8.67          =>   Philosophie : moyenne = 9.71              0.57
5     Moyenne.G : moyenne = 8.67          =>     Physique.Chi : moyenne = 6              0.53
6     Moyenne.G : moyenne = 8.67          =>      Francais : moyenne = 7.39              0.50
7     Moyenne.G : moyenne = 8.67          =>           SVT : moyenne = 4.59              0.47
8     Moyenne.G : moyenne = 8.67          => Langue.Vivante : moyenne = 4.98             0.40
9     Moyenne.G : moyenne = 8.67          =>          EPS : moyenne = 11.27              0.32
10 Mathematiques : moyenne = 5.4          =>     Physique.Chi : moyenne = 6              0.31
11 Mathematiques : moyenne = 5.4          =>           SVT : moyenne = 4.59              0.23
```

Figure 2 –. Résultat pour MinSup = 0.99, MaxCVS = 0.6 et MinRapCor = 0.2

Ensuite, on note que la meilleure règle valide est la suivante : (Moyenne générale=8.67)⇒(Mathématiques=5.4). Cette règle nous incite à penser que la réussite des élèves série A2 dépend de leurs notes des mathématiques. Cela est évident, car la série A2 de Madagascar est faite pour les élèves littéraires qui s'intéressent aux mathématiques (coefficient 03). C'est pourquoi la faiblesse de note de mathématiques entraîne un échec pour ces candidats de la série A2.

Les deux dernières règles montrent l'existence d'une liaison fonctionnelle entre les notes des matières scientifiques. Ce résultat peut être interprété comme suit : en général, si l'élève série A2 de la région DIANA de Madagascar a une note égale à 05/20 en mathématiques, alors il aurait une note de 05/20 en Science de la Vie et de la Terre, et une note de 06/20 en Physique Chimie. Ce niveau démontre clairement que les élèves de série A (série littéraire) ne maîtrisent pas les matières scientifiques. On peut aussi admettre que, en tant que littéraire, ces élèves négligent les matières scientifiques.

Ces informations pourraient aider les décideurs en éducation de Madagascar de décider :

- d'améliorer la qualité d'enseignement des matières littéraires ;
- de diminuer les coefficients des matières scientifiques dans la série A, car ces matières n'intéressent pas les élèves de la série A ;
- d'éliminer peut – être la série A2, car leurs notes des mathématiques sont très faibles : 05/20. Toutefois, la série A2 est faite pour les élèves série A qui s'intéressent beaucoup aux mathématiques (coefficient 03 selon le système de l'année 2014).

D'où, la justification du nouveau programme scolaire du Plan Sectoriel de l'Education (PSE) actuel pour les séries littéraires. Les décideurs ont décidé d'éliminer la série A2, en fixant seulement la série L. En série L, les matières littéraires sont toutes dotées de coefficients 04 et 05, tandis que les matières scientifiques sont toutes de coefficient 01. Ainsi, les élèves littéraires peuvent se spécialiser dans leurs domaines.

5 Discussion

Dans cette section, nous allons discuter et comparer théoriquement les résultats de notre méthode avec les résultats obtenus par d'autres théories existantes dans la littérature.

5.1 Comparaison avec la théorie de Aumann et Lindell

Selon la théorie de Aumann et Lindell (1999), une règle $(X \in [a\,;b]) \Rightarrow Y(M_Y = \overline{Y})$ est dite irréductible si pour toute constante $c \in [a\,;b]$, on a : $\overline{Y}_{X\in[a,c]} > \overline{Y}$ et $\overline{Y}_{X\in[c,b]} > \overline{Y}$, où $\overline{Y}$ est la moyenne arithmétique de Y ; $\overline{Y}_{X\in[a,c]}$ et $\overline{Y}_{X\in[c,b]}$ correspondent respectivement à la moyenne arithmétique de Y sur les individus vérifiant $X \in [a, c]$ et $X \in [c, b]$. Ensuite, une règle $(X \in [a\,;b]) \Rightarrow Y(M_Y = \overline{Y})$ est dite maximale si l'on ne peut plus élargir l'intervalle $[a, b]$ de la prémisse pour que cette règle soit irréductible. Une règle maximale est la règle irréductible qui contient l'intervalle la plus longue $[a, b]$. Alors, le fait qu'une règle $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$ soit validée avec un seuil $MaxCVS$ faible (inférieur à 0.3) et un seuil $MinRapCor$ élevé (supérieur à 0.7) ne nous permet pas de conclure que la règle $(X \in [\min(X)\,, \max(X)]) \Rightarrow (Y = \bar{Y})$ est quasi-maximale selon la théorie de Aumann et Lindell (1999).

En effet, si une règle $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$ est valide avec un seuil $MinRapCor$ élevé, alors il existe une relation de dépendance continue, non constante et non nécessairement linéaire f entre les deux variables aléatoires X et Y, telle que $Y = f(X)$. Alors, il suffit simplement de considérer deux variables aléatoires X et Y linéairement et positivement dépendantes : $Y = aX + b$, avec $a > 0$. Dans ce cas, on peut trouver une constante $c \in [\min(X)\,, \max(X)]$ et $c \leq \bar{X}$, telle que $\bar{Y}_{[\min(X),c]} \leq \bar{Y}$. Ainsi, la règle $(X \in [\min(X)\,, \max(X)]) \Rightarrow (Y = \bar{Y})$ n'est pas maximale selon la théorie de Aumann et Lindell (1999).

5.2 Comparaison selon la théorie de Agrawal et al.

Tout d'abord, selon Chafaï et Zitt (2016), le théorème de Bienaymé – Tchebycheff affirme que pour une variable aléatoire X ayant une moyenne E(X) et une variance $(X \in L^2)$, on a :

$$\forall\, r > 0, P(|X - E(X)| \geq r) \leq \frac{\sigma^2(X)}{r^2}. \qquad (16)$$

C'est une majoration de la probabilité pour que l'écart entre les valeurs de X et sa valeur moyenne $E(X)$ soit plus grand que r. Ainsi, plus $\sigma^2(X)$ est faible, plus $E(X)$ représente vraiment une valeur typique de X. Dans ce cas, on dit que la moyenne $E(X)$ est significative, car les valeurs de X sont homogènes ou regroupées autour de la valeur moyenne $E(X)$. Par conséquent, si $\sigma^2(X)$ est faible, alors il existe un intervalle $[a\,;b]$ contenant $E(X)$, et un petit nombre réel $\varepsilon\, \epsilon\, [0\,;1]$, tel que $P(X \notin [a\,;b]) < \varepsilon$ et $[a\,;b] \subset [Min(X)\,;Max(X)]$. En passant à l'événement contraire, il existe un intervalle $[a\,;b]$ contenant $E(X)$, et un petit nombre réel $\varepsilon\, \epsilon\, [0\,;1]$, tel que $P(X \in [a\,;b]) \geq 1 - \varepsilon$ et $[a\,;b] \subset [Min(X)\,;Max(X)]$.

Toutefois, on a : $CVS(X) = a \times \sigma(X)$, avec $a = 0$, si $Max(X) = \bar{X}$ ou $Min(X) = \bar{X}$; $a = \frac{1}{\sqrt{[Max(X)-\bar{X}\,]\,[\,\bar{X}-Min(X)]}}$, si $Max(X) \neq \bar{X}$ et $Min(X) \neq \bar{X}$. Comme $a \geq 0$, alors si $\sigma(X)$ est faible, alors $CVS(X)$ l'est aussi. Par conséquent, si $CVS(X)$ est faible, alors il existe un intervalle $[a\,;b]$ contenant $E(X)$, et un petit nombre réel $\varepsilon\, \epsilon\, [0\,;1]$, tel que $P(X \in [a\,;b]) \geq 1 - \varepsilon$ et $[a\,;b] \subset [Min(X)\,;Max(X)]$. Cependant, le fait que $CVS(X)$ soit faible nous permet de faire l'approximation $[a\,;b] \approx [Min(X)\,;Max(X)]$. Ensuite, la règle $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$ peut être considérée comme $(X \in [\min(X)\,, \max(X)]) \Rightarrow (Y \in [\min(Y)\,, \max(Y)])$, car les moyennes $\bar{X}$ et $\bar{Y}$ représentent respectivement les valeurs générales de X et Y dans les intervalles $[\min(X)\,, \max(X)]$ et $[\min(Y)\,, \max(Y)]$. Par conséquent, si CVS(X) et CVS(Y) sont

faibles, alors les deux intervalles [min(X) , max(X)]) et [min(Y) , max(Y)]) peuvent être considérés respectivement comme les « meilleures partitions »pour les attributs quantitatifs X et Y. Ainsi, si CVS(X) et CVS(Y) sont faibles, alors on peut avoir tout de suite les « meilleures partitions » nécessaires pour engendrer une règle quantitative comme dans le travail de Srikant et Agrawal (1996). Le seul problème qui ne nous permet pas d'avoir une même partition que celle de Srikant et Agrawal (1996) est la contrainte de MaxSup qu'ils ont intégré dans leur algorithme afin d'éviter le problème de MinConf. Plus nous utilisons un MaxSupp élevé et un MaxCVS faible, plus on retrouve des partitions similaires. Cependant, les règles obtenues ne seront pas toujours les mêmes, car la liaison entre les mesures confiance et le rapport de corrélation est floue. Notre approche minimise la perte d'informations, car on ne fait point de partitionnement. Toutefois, les résultats sont limités à des règles comportant seulement $1-$ itemset dans la prémisse et dans la conséquence. Ainsi, il reste encore à trouver une autre technique permettant de réappliquer cette méthode afin d'extraire des règles avec $k-$ itemsets.

6 Conclusion

Nous avons introduit une nouvelle technique d'extraction des règles d'association quantitative utilisant la mesure de dispersion « coefficient de variation de Sharma CVS » et la mesure de dépendance fonctionnelle « rapport de corrélation ». En général, le traitement se fait en trois étapes : (1) sélection des items fréquents satisfaisant la contrainte $MinSupp$ fixé par l'utilisateur, en passant par une matrice booléenne intermédiaire ; (2) sélection des items fréquents satisfaisant la contrainte $MaxCVS$ spécifié par l'utilisateur, afin de travailler seulement avec les items possédant des valeurs homogènes ; (3) formation des règles quantitatives et sélection des règles satisfaisant la contrainte minimum de rapport de corrélation donnée par l'utilisateur, notée $MinRapCor$. Le résultat obtenu est un ensemble des règles d'association quantitatives de type $(X = \bar{X}) \Rightarrow (Y = \bar{Y})$, telle que $\bar{X}$ et $\bar{Y}$ sont les moyennes des valeurs des attributs quantitatifs X et Y. Une telle règle est interprétée comme un comportement général du phénomène étudié. Les résultats des expérimentations avec les données réelles de l'éducation de la région nord de Madagascar présentés dans ce papier justifient les reformes en éducation effectuées par les décideurs de Madagascar sur le Plan Sectoriel de l'Education dans la section littéraire. Ainsi, ce papier offre un outil d'aide à la prise de décision scientifique à tous les gérants, économistes et hauts responsables d'un pays. L'extension de notre technique à des règles d'association dont la prémisse et le conséquent sont de $k-itemsets$ est la suite naturelle de notre travail. En perspective, nous pensons que la justification de l'emploi du *rapport de corrélation* pourrait se faire aussi à partir de la considération de l'intensité d'implication de Gras.

7 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

[1] Agrawal R., Imielinski T. et Swami A. (1993), *Mining association rules between sets of items in large databases*, Proceedings of the ACM SIGMOD international conference on Management of data. 207–216.

[2] Agrawal R. et Srikant R. (1994), *Fast algorithms for mining association rules*. Proceedings of the 20th International Conference on Very Large Data Bases, 487 – 499.

[3] Aumann Y. et Lindell Y. (1999). *A statistical theory for quantitative association rules*. Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. August, 261–270.

[4] Chafaï D. et Zitt P. A. (2016), Probabilités – Préparation à l'agrégation interne, CreatSpace Independent Publishing, https://doi.org/978-1537566542.

[5] Sharma R., Shandil R.G. et Kapoor G. (2011), A note on Karl Pearson's coefficient of dispersion. Himachal Pradesh University Journal, https://pdfs.semanticscholar.org/229c/6e64c38e0b39773b03b82f458d8779e1f638.pdf or https://studylib.net/doc/8762538/a-note-on-karl-pearson-s-coefficient-of-dispersion

[6] Sørensen B. J. (2002), *The Use and Misuse of the Coefficient of Variation in Organizational Demography Research,* Sociological Methods & Research, 30(4), 475–491.

[7] Srikant R. et Agrawal R. (1996), *Mining quantitative association rules in large relational databases*, Proceedings of the 1996 ACM SIGMOD international conference on Management of data, 1–12.

[8] Sujatha D. et Naveen C. H. (2011), *Quantitative Association Rule Mining on Weighted Transactional Data.* International Journal of Information and Education Technology, Vol. 1, No. 3.

[9] Reh W. and Scheffier B. (1996), Significance Tests and Confidence Intervals for Coefficients of Variation. The statistical software newsletter, 449 – 452.

[10] 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*. Thèse de HDR spécialité Mathématiques et informatique, Université de Madagascar.

[11] Webb G.I. (2001), *Discovering associations with numeric variables*, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 383–388.