Extraction des règles d'association basée sur la mesure classique d'intensité d'implication: application en didactique des mathématiques

Fidy Heritiana Andrianarivony, Anne Cortella, Jean-Jacques Salone, Viviane Durand-Guerrier, Angelo Raherinirina

ARIMA

épisciences

Extraction des règles d'association basée sur la mesure classique d'intensité d'implication: application en didactique des mathématiques

Fidy ANDRIANARIVONY[1,2*], **Anne CORTELLA**[1], **Jean-Jacques SALONE**[3], **Viviane DURAND-GUERRIER**[1], **Angelo RAHERINIRINA**[2]

[1]Institut Montpelliérain Alexander Grothendieck - UMR 5149 - DEMa, Université de Montpellier, France
[2]Centre de Recherche sur l'Enseignement des Mathématiques (CREM), Université de Fianarantsoa, Madagascar
[3]Contextes, Recherches et Ressources en Education et Formation, Université des Antilles, France

*E-mail : fidy-heritiana.andrianarivony@etu.umontpellier.fr

DOI : 10.46298/arima.12231
Soumis le 5 septembre 2023 - Publié le 1 mars 2024
Volume : **40** - Année : **2024**

Éditeurs : *Mathieu Roche, Nabil Gmati, Clémentin Tayou Djamegni*

Résumé

Cet article propose une méthode pour extraire des connaissances en règles d'association en utilisant la mesure classique de l'intensité d'implication. Nous avons ensuite appliqué notre méthode dans des données issues de travaux en didactique des mathématiques. L'objectif de l'étude en didactique est de connaitre les relations entre les difficultés et les compétences des élèves lorsque ceux-ci démontrent une proposition mathématique formulée en langue française. Le résultat de notre étude nous a démontré que notre méthodologie est efficace pour extraire les règles intéressantes. De plus les résultats d'analyse didactique ont montré la dépendance entre compréhension d'un énoncé mathématique en français, compétence à le traduire formellement et à le prouver.

Mots-Clés

Didactique des mathématiques ; Intensité d'implication ; Langage et logique, Règles d'association ; Raisonnements mathématiques

I INTRODUCTION

Dans le domaine de fouille des données, l'extraction des connaissances par les règles d'association est une méthode beaucoup utilisée par les chercheurs. C'est une science pluridisciplinaire qui concerne à la fois la statistique, l'apprentissage automatique et les bases de données. La première application de cette sciences remonte à Agrawal & Srikant [4] qui ont proposé un algorithme qui permet de connaitre la tendance des achats des clients d'un supermarché. Dans ce sens, ils ont utilisé deux mesures d'intérêt : le support et la confiance.

La recherche sur les règles d'association est appliquée dans des domaines plus vastes, dont la

médecine, le marketing, l'ingénierie, *etc*. Par exemple, dans Ait-Mlouk [21], nous trouvons une application de la recherche des règles d'association dans le domaine de la sécurité routière ; dans la thèse de Pennerath [11], nous trouvons une application à des problèmes de synthèse organique, c'est-à-dire dans le domaine de la science chimique ; et enfin dans Idiri [16], nous trouvons une application dans le domaine de la surveillance maritime.

Outre ces applications dans divers domaines, son application dans le domaine de la recherche en didactique des mathématiques est en plein essor. Nous pouvons citer les travaux de Gras et ses collaborateurs ([9][13][14]) qui cherchent une association entre les variables binaires et non binaires par la mesure de l'intensité d'implication. Dans ce cadre nous citons les travaux de Lahanier-Reuter [10] et Ottaviani & Zannoni [6] qui essaient d'appliquer l'analyse statistique implicative en didactique des mathématiques. La thèse de Ramanantsoa [19] qui cherche les règles d'association valides par la mesure MGK dans le domaine de la didactique des mathématiques et didactique de l'informatique. La thèse de Rakotomalala [22] propose une application dans le domaine de didactique de l'informatique.

Un des problèmes de la recherche sur les règles d'association est celui de la sélection de la mesure d'intérêt utilisée. Pour choisir une mesure, certains critères doivent être vérifiés. Ainsi la thèse de Grissa [15] propose une étude comportementale des mesures d'intérêt d'extraction des règles d'association. Nous pouvons aussi citer Gras & Couturier [13] qui étudient la spécificité de l'analyse statistique implicative par rapport à d'autres mesures d'intérêt, dans laquelle l'intensité d'implication est proposée. Ainsi dans Gras, Kuntz & Briand [14], nous trouvons une prolongement de l'analyse statistique implicative vers la fouille de données. Cependant, les résultats obtenus sont plutôt le *clustering* mais non pas les règles d'association.

Dans cet article, nous avons proposé une méthode qui permet d'utiliser la mesure classique de l'intensité d'implication pour extraire des règles d'association dans le domaine de la didactique des mathématiques. Nous avons ainsi deux objectifs : le premier est de pouvoir utiliser la mesure classique de l'intensité d'implication pour extraire des connaissances en règles d'association, et le deuxième est de faire une application en didactique des mathématiques, en particulier dans l'étude des relations entre les difficultés et compétences des élèves lorsqu'ils font une démonstration d'une proposition mathématique formulée en langue française.

Pour aborder cette étude, nous allons présenter dans un premier temps notre état de l'art, dans lequel nous allons parler brièvement des mesures d'intérêt tels que support, confiance et intensité d'implication. Dans un deuxième temps, nous développons notre méthode pour utiliser la mesure classique de l'intensité d'implication pour extraire des connaissances. La dernière partie de cet article aborde une problématique didactique, dans laquelle nous trouvons une application de notre approche.

II ÉTAT DE L'ART

Dans cette section nous présentons un bref aperçu des concepts de base que nous allons utiliser, dont la méthode classique pour extraire les connaissances en règles d'association et la mesure classique de l'intensité d'implication.

2.1 Règles d'association

Une règle d'association est une technique d'analyse de données largement utilisée en exploration de données et en apprentissage automatique. Elle permet de découvrir des relations intéressantes et souvent cachées entre différents éléments ou variables dans un ensemble de données.

Plus précisément, les règles d'association identifient les relations entre les variables ou items, ou encore attributs, qui sont très souvent binaires, d'un ensemble de variables $\mathcal{I}$ dans une base de données transactionnelle. Elles sont basées sur des mesures d'intérêt, dont les plus fréquentes sont le support et la confiance. Elles sont généralement formulées sous la forme « Si X alors Y » où $X = x_1 \wedge x_2 \wedge \cdots \wedge x_p$ et $Y = y_1 \wedge y_2 \wedge \cdots \wedge y_q$ sont des conjonctions finies de variables de $\mathcal{I}$ et sont disjoints. Ces deux sous-ensembles X et Y de $\mathcal{I}$ s'appellent motifs.

2.1.1 Mesures d'intérêt Support-Confiance

Considérons une population $\mathcal{O}$ de n objets o où on fait une enquête. Ces objets peuvent être des entités ou des individus, selon la population choisie pour faire l'étude. Pour définir la relation entre les n objets de $\mathcal{O}$ et les m variables de $\mathcal{I}$, on note par $\mathcal{R}$ la relation binaire dont $o\mathcal{R}x$ si et seulement si $x(o) = 1$, c'est-à-dire que l'objet o possède la variable x, ou que la variable x est présente chez l'objet o.

On appelle contexte d'extraction des règles d'association, ou simplement contexte d'extraction, le triplet $\mathbb{K} = (\mathcal{O}, \mathcal{I}, \mathcal{R})$ avec $\mathcal{O}$ l'ensemble d'objets où on fait l'étude, $\mathcal{I}$ l'ensemble de variables et $\mathcal{R}$ une relation binaire entre un individu de $\mathcal{O}$ et une variable de $\mathcal{I}$. Ce triplet peut être aussi appelé base de données. Il est représenté sous forme d'un tableau à n lignes et à m colonnes. Dans les lignes sont placées les n objets de $\mathcal{O}$ et dans les colonnes sont placés les variables binaires de $\mathcal{I}$. On trouve 1 dans l'intersection de la i-ième ligne ($1 \leq i \leq n$) et la j-ième colonne ($1 \leq j \leq m$) si l'objet $o_i\mathcal{R}x_j$, et 0 sinon.

Pour évaluer la qualité d'une règle $X \rightarrow Y$, le modélisateur utilise le plus souvent les deux mesures d'intérêt *support* et *confiance*.

Définition 1: Support
On définit ainsi le support entre deux motifs X et Y, la fréquence d'objet o de $\mathcal{O}$ qui possèdent en même temps les variables de X et de Y.

Notons par $X' = \{o \in \mathcal{O} / \forall x \in X, o\mathcal{R}x\}$ (respectivement par $Y' = \{o \in \mathcal{O} / \forall y \in Y, o\mathcal{R}y\}$) l'ensemble des objets o de $\mathcal{O}$ qui possèdent les propriétés de toutes les variables de X (respectivement l'ensemble des objets o de $\mathcal{O}$ qui possèdent les propriétés de toutes les variables de Y), alors

$$Supp(X \rightarrow Y) = \frac{Card(X' \cap Y')}{Card(\mathcal{O})} \tag{1}$$

Définition 2: Confiance
La confiance d'une règle d'association $X \rightarrow Y$ est définie par la formule :

$$Conf(X \rightarrow Y) = \frac{Card(X' \cap Y')}{Card(X')} \tag{2}$$

où X' et Y' sont les ensembles définis ci-dessus. En terme de probabilité, cette formule mesure la probabilité conditionnelle de la réalisation de Y' sachant X'.

2.1.2 Règles d'association intéressantes

Pour générer les règles d'association intéressantes, le modélisateur définit un seuil minimum pour le support et la confiance. C'est-à-dire, une règle $X \rightarrow Y$ est intéressante au sens du support et confiance si et seulement si $Supp(X \rightarrow Y) \geq \alpha$ et $Conf(X \rightarrow Y) \geq \beta$, où α et β sont des seuils minimaux pour les support et confiance choisis par le modélisateur.

Par ailleurs, l'algorithme d'extraction des règles est de complexité exponentielle, puisque si on a m variables, on peut avoir 2^m motifs. La phase la plus complexe est l'étape d'extraction des motifs fréquents, dont le premier algorithme dénommé par *algorithme Apriori* est donné par Agrawal & Srikant [4]. Cet algorithme se base essentiellement sur la propriété d'anti-monotonicité de Support existant entre les motifs [15]. Il permet d'évaluer les règles potentiellement intéressantes, et de ne garder que celles qui satisfont les mesures basées sur les mesures d'intérêt support et confiance.

2.2 La mesure classique de l'intensité d'implication

L'utilisation seule des mesures d'intérêt support et confiance génère généralement un grand nombre des règles dont certaines sont parfois redondantes et qui ne sont pas intéressantes. Ainsi, les modélisateurs cherchent des nouvelles mesures pour n'extraire que les règles intéressantes, faciles à interpréter. Ces nouvelles mesures jouent un rôle très important dans l'évaluation en termes d'intérêt et de pertinence des règles extraites en permettant un filtrage ou un ordonnancement automatique de celles-ci.

Parmi les mesures existantes, on a choisi d'utiliser la mesure classique de l'intensité d'implication entre deux variables binaires x_i et x_j. Ce choix est motivé par la propriété dissymétrique et implicative de l'intensité d'implication [13]. En effet, il existe plusieurs mesures qui permettent de déceler des règles de concomitance et qui sont essentiellement symétriques. Par exemple, le support mesure l'occurrence simultanée des motifs X et Y parmi les n individus enquêtés. Tandis que l'intensité d'implication différencie clairement la prémisse du conséquent.

Pour mesurer l'implication $x_i \rightarrow x_j$, Gras & Régnier [9] compte le nombre de contre-exemples qui infirment la règle, c'est-à-dire, le nombre d'objets dans $\mathcal{O}$, qui sont ici les individus enquêtés, qui vérifient x_i mais pas x_j. La règle est ainsi admissible si le nombre de contre-exemples est petit devant le nombre total d'individus enquêtés.

Pour la modélisation, nous considérons les motifs $\{x_i\}$ et $\{x_j\}$, formés par des singletons. Supposons en plus les deux sous-ensembles X' et Y' quelconques de $\mathcal{O}$ et de même cardinaux respectifs que l'ensemble des objets qui ont le motif $\{x_i\}$ et l'ensemble des objets qui ont le motif $\{x_j\}$. Ces deux événements X' et Y' étant alors indépendants.

Définition 3: Intensité d'implication
On appelle intensité d'implication de la règle $x_i \rightarrow x_j$ le nombre :

$$\varphi(x_i \rightarrow x_j) = \begin{cases} 1 - \mathbb{P}\left[Card(X' \cap \overline{Y'}) \leq n_{x_i \wedge \overline{x_j}}\right] \ si\ n \neq n_{x_j} \\ 0\ si\ n = n_{x_j} \end{cases} \tag{3}$$

où $n_{x_i \wedge \overline{x_j}}$ est le nombre d'individus enquêtés dans $\mathcal{O}$ qui vérifient la propriété de la variable x_i mais pas x_j et $Card(X' \cap \overline{Y'})$ est le nombre aléatoire de contre-exemple de la règle $x_i \rightarrow x_j$. Cette variable aléatoire peut suivre la loi de Poisson, la loi binomiale ou la loi hypergéométrique [9]. Dans la suite de ce texte, on choisi la loi de Poisson pour modéliser l'intensité d'implication.

Cette mesure d'intensité d'implication caractérise la qualité de la règle $x_i \rightarrow x_j$ [9], donc selon nous la probabilité $\mathbb{P}(x_i \rightarrow x_j)$ pour avoir une implication de x_i vers x_j. Dans la suite on ne considère que le cas où $n \neq n_{x_j}$. En appliquant la loi de Poisson de paramètre $\lambda = \frac{n_{x_i} \cdot n_{\overline{x_j}}}{n}$, on a la formule de l'intensité d'implication suivante :

$$\varphi(x_i \rightarrow x_j) = 1 - \sum_{k=0}^{n_{x_i \wedge \overline{x_j}}} \frac{\lambda^k}{k!} e^{-\lambda} \tag{4}$$

La variable aléatoire $Q(X' \cap \overline{Y'}) = \frac{Card(X' \cap \overline{Y'}) - \lambda}{\sqrt{\lambda}}$ est déduite de l'autre variable aléatoire $Card(X' \cap \overline{Y'})$. Une valeur empirique observée de la variable aléatoire $Q(X' \cap \overline{Y'})$ est $q(x_i, x_j)$ définie par : $q(x_i, x_j) = \frac{n_{x_i \wedge \overline{x_j}} - \lambda}{\sqrt{\lambda}}$. Sous une hypothèse de paramètre assez élevé ($\lambda \geq 5$), la variable aléatoire $Q(X' \cap \overline{Y'})$ peut être approximée par une loi Normale centrée-réduite $\mathcal{N}(0, 1)$ [9], d'où l'intensité d'implication s'écrit :

$$\varphi(x_i \rightarrow x_j) = \frac{1}{\sqrt{2\pi}} \int_{q(x_i, x_j)}^{\infty} e^{-\frac{t^2}{2}} dt \tag{5}$$

Dans toute la suite de ce texte, nous utiliserons la formule (4) si $\lambda < 5$ et la formule (5) sinon.

III MÉTHODE D'EXTRACTION DES RÈGLES D'ASSOCIATION BASÉE SUR LA MESURE CLASSIQUE DE L'INTENSITÉ D'IMPLICATION

L'intensité d'implication présentée dans Gras & Réginer [9] fait correspondre deux variables binaires, ou en d'autre terme deux motifs disjoints ne contenant que des singletons. En fouille des données, on fait correspondre deux motifs dont chacun est une conjonction de variables binaires. En outre, dans Gras et al. [14], on trouve un prolongement de l'analyse statistique implicative dans les fouilles de données. Ainsi, Gras et al [14] proposent la nouvelle mesure d'intensité d'implication entropique et calculent à partir de cette mesure la cohésion de classe de variables. Cependant les résultats obtenus par cette méthode est plutôt le clustering. Nous, dans cette étude, proposons une approche qui permet d'utiliser la mesure classique de l'intensité d'implication pour la fouille de données, notamment la recherche des règles d'association intéressantes. Ce choix est légitimé par la clarté de signification implicative de la mesure classique de l'intensité d'implication.

Le but de cette section est alors de proposer une formule générale de l'intensité d'implication de Gras & Réginer [9] pour des implications dont la prémisse et le conséquent sont des conjonctions finies de variables binaires et disjoints. Pour ce faire, nous portons notre contributions ci-dessous en proposant trois étapes de généralisation.

3.1 Première étape de généralisation pour calculer la fiabilité de la règle $X \rightarrow Y$

En fouille de données, les deux motifs X et Y sont des conjonctions de variables binaires, soit $X = x_1 \wedge x_2 \wedge \cdots \wedge x_p$ et $Y = y_1 \wedge y_2 \wedge \cdots \wedge y_q$. Donc notre objectif est de chercher une implication entre ces deux conjonctions de variables.

Dans un premier temps, on suppose que le motif X contient deux éléments et le motif Y est un singleton. Soit $X = x_1 \wedge x_2$ et $Y = y$. La règle logique nous permet d'avoir l'équivalence suivante [2] :

$$(x_1 \wedge x_2 \rightarrow y) \Leftrightarrow [x_1 \rightarrow (x_2 \rightarrow y)] \tag{6}$$

Cette équivalence 6 nous montre la fermeture transitive de l'implication entre les trois variables x_1, x_2 et y. La convention de rédaction des règles logiques stipule que $x_1 \rightarrow (x_2 \rightarrow y)$ est l'interprétation de $x_1 \rightarrow x_2 \rightarrow y$. Ainsi, nous admettons que les deux implications $x_1 \rightarrow x_2$ et $x_2 \rightarrow y$ doivent être en même temps vraies pour que cette règle soit valide.

On définit ainsi la probabilité d'avoir la règle $(x_1 \wedge x_2 \rightarrow y)$ par la formule suivante :

$$\mathbb{P}(x_1 \wedge x_2 \rightarrow y) = \mathbb{P}((x_1 \rightarrow x_2) \wedge (x_2 \rightarrow y)) \tag{7}$$

Dans Gras & Régnier [9], on a pris deux événements indépendants pour calculer l'intensité d'implication entre deux variables binaires. Ici, on a une succession de deux règles implicatives $x_1 \rightarrow x_2$ et $x_2 \rightarrow y$, dont chacune de leurs intensités sera calculée avec la méthode présentée dans Gras & Régnier [9]. On peut supposer alors que les deux événements pour avoir les deux implications $x_1 \rightarrow x_2$ et $x_2 \rightarrow y$ sont indépendants. Ainsi, la formule (7) devient :

$$\mathbb{P}(x_1 \wedge x_2 \rightarrow y) = \mathbb{P}(x_1 \rightarrow x_2) \times \mathbb{P}(x_2 \rightarrow y) \tag{8}$$

C'est-à-dire,

$$\varphi(x_1 \wedge x_2 \rightarrow y) = \varphi(x_1 \rightarrow x_2) \times \varphi(x_2 \rightarrow y) \tag{9}$$

On peut démontrer par récurrence que si X est une conjonction de p variables binaires et Y est un singleton, la règle de l'implication logique nous permet d'écrire la formule suivante :

$$(x_1 \wedge x_2 \wedge \cdots \wedge x_p \rightarrow y) \Longleftrightarrow [x_1 \rightarrow (x_2 \rightarrow (x_3 \rightarrow \ldots (x_p \rightarrow y)))] \tag{10}$$

Ainsi la fiabilité de la règle $(x_1 \wedge x_2 \wedge \cdots \wedge x_p \rightarrow y)$ s'écrit :

$$\mathbb{P}(x_1 \wedge x_2 \wedge \cdots \wedge x_p \rightarrow y) = \prod_{i=1}^{p-1} \mathbb{P}(x_i \rightarrow x_{i+1}) \times \mathbb{P}(x_p \rightarrow y) \tag{11}$$

c'est-à-dire,

$$\varphi(x_1 \wedge x_2 \wedge \cdots \wedge x_p \rightarrow y) = \varphi(x_p \rightarrow y) \times \prod_{i=1}^{p-1} \varphi(x_i \rightarrow x_{i+1}) \tag{12}$$

La mesure de l'intensité d'implication d'une conjonction de variables vers un singleton est obtenue par cette formule (12).

3.2 Deuxième étape de généralisation pour calculer la fiabilité de la règle $X \to Y$

Dans ce deuxième temps on suppose que X est un singleton et Y est une conjonction de q variables binaires. Soit $X = x$ et $Y = y_1 \wedge y_2 \wedge \cdots \wedge y_q$. Notre problématique est de chercher l'intensité d'implication de la règle $x \to y_1 \wedge y_2 \wedge \cdots \wedge y_q$. Pour simplifier notre étude, on suppose d'abord que Y est formé par deux éléments, c'est-à-dire $Y = y_1 \wedge y_2$. La formule logique nous donne l'équivalence suivante [2] :

$$(x \to y_1 \wedge y_2) \Leftrightarrow [(x \to y_1) \wedge (x \to y_2)] \tag{13}$$

En reprenant l'hypothèse d'indépendance des événements prise par Gras & Régnier [9], on a la formule de probabilité de la règle $(x \to y_1 \wedge y_2)$ suivante :

$$\mathbb{P}(x \to y_1 \wedge y_2) = \mathbb{P}(x \to y_1) \times \mathbb{P}(x \to y_2) \tag{14}$$

c'est-à-dire,

$$\varphi(x \to y_1 \wedge y_2) = \varphi(x \to y_1) \times \varphi(x \to y_2) \tag{15}$$

On peut généraliser cette formule (15) par récurrence sur q si on a q variables binaires formant le motif Y et les événements d'avoir $x \to y_i$ sont indépendants pour tout i vérifie $1 \leq i \leq q$.

$$\mathbb{P}(x \to y_1 \wedge y_2 \wedge \cdots \wedge y_q) = \prod_{i=1}^{q} \mathbb{P}(x \to y_i) \tag{16}$$

c'est-à-dire,

$$\varphi(x \to y_1 \wedge y_2 \wedge \cdots \wedge y_q) = \prod_{i=1}^{q} \varphi(x \to y_i) \tag{17}$$

Cette formule (17) permet de calculer l'intensité d'implication de la règle : singleton implique une conjonction de variables.

3.3 Troisième étape de généralisation pour calculer la fiabilité de la règle $X \to Y$

Dans cette dernière étape, supposons que $X = x_1 \wedge x_2 \wedge \cdots \wedge x_p = \bigwedge_{i=1}^{p} x_i$, un motif à p variables et $Y = y_1 \wedge y_2 \wedge \cdots \wedge y_q = \bigwedge_{j=1}^{q} y_j$, un motif à q variables. On va chercher la formule de l'intensité d'implication qui généralise la mesure de la fiabilité de la règle $X \to Y$.

La formule logique nous permet d'écrire :

$$\bigwedge_{i=1}^{p} x_i \to \bigwedge_{j=1}^{q} y_j \Longleftrightarrow \bigwedge_{j=1}^{q} \left(\bigwedge_{i=1}^{p} x_i \to y_j \right) \tag{18}$$

En supposant que chaque événement d'avoir $\bigwedge_{i=1}^{p} x_i \to y_j$ sont indépendants, avec $1 \leq j \leq q$, alors on a :

$$\mathbb{P}\left(\bigwedge_{i=1}^{p} x_i \to \bigwedge_{j=1}^{q} y_j \right) = \prod_{j=1}^{q} \mathbb{P}\left(\bigwedge_{i=1}^{p} x_i \to y_j \right) = \prod_{j=1}^{q} \mathbb{P}\left(x_p \to y_j\right) \left[\prod_{i=1}^{p-1} \mathbb{P}\left(x_i \to x_{i+1}\right) \right]^{q} \tag{19}$$

c'est-à-dire,

$$\varphi\left(\bigwedge_{i=1}^{p} x_i \to \bigwedge_{l=1}^{q} y_l\right) = \prod_{j=1}^{q} \varphi(x_p \to y_j) \left[\prod_{i=1}^{p-1} \varphi(x_i \to x_{i+1})\right]^q \tag{20}$$

Cette formule (20) est l'intensité d'implication généralisée entre motifs.

3.4 Processus algorithmique pour générer les règles d'association intéressantes

Dans cette section nous allons donner les étapes algorithmiques pour chercher les règles d'association $X \to Y$ intéressantes.

Dans Lehn, Guillet & Kuntz [5], un processus algorithmique en deux étapes est proposé pour faire une extraction des règles d'assocation par l'utilisation de la mesure d'intensité d'implication : la découverte des motifs fréquents et la découverte des règles d'association intéressantes par le calcul de support, confiance et l'intensité d'implication de la règle. En s'inspirant de cette méthode qui a été proposée dans Lehn et al. [5], voici le processus algorithmique qui permet de générer les règles intéressantes :

1. La découverte des motifs fréquents. Cette première étape est la sélection des motifs ayant un support supérieur à un seuil α fixé par le modélisateur, c'est-à-dire $Supp(X) \geq \alpha$. Par le principe d'anti-monotonie, on peut éliminer les motifs X_1 dont $X \subset X_1$ et X n'est pas fréquent. Par contre, on peut sélectionner tous les motifs X_2, dont $X_2 \subset X$ et X fréquent. L'algorithme Apriori nous permet de découvrir les motifs fréquents [4].
2. La découverte des règles intéressantes par le calcul de l'intensité d'implication, c'est-à-dire si γ est le seuil minimum pour l'intensité d'implication, alors on sélectionne les règles ayant une intensité supérieure à γ. Dans cette étape, on ne calcule plus la confiance. En fait cette mesure sert à exprimer l'implication de X sur Y qui est aussi l'objet de l'intensité d'implication. Pour nous, l'intensité d'implication suffit pour exprimer cette implication.

A l'issue de ce processus algorithmique, pour chaque motif fréquent X, nous avons tous ses sous-ensembles, qui sont aussi fréquents, pour générer toutes les règles d'association de la forme $X_1 \to (X - X_1)$ avec $X_1 \subset X$. L'ensemble des règles d'association intéressantes $\mathcal{R}_v$ peut être représenté comme suit :

$$\mathcal{R}_v = \{X_1 \to (X - X_1) / Supp(X) \geq \alpha \wedge (X_1 \subset X) \wedge \varphi(X_1 \to (X - X_1)) \geq \gamma\} \tag{21}$$

IV APPLICATION EN DIDACTIQUE DES MATHÉMATIQUES

Dans cette section nous présentons une application de notre approche pour extraire des connaissances à partir de données en didactique des mathématiques. Ceci nous permet de mettre en lumière l'articulation entre une approche didactique et une approche mathématique pour modéliser les phénomènes d'enseignement des mathématiques.

L'objet de notre recherche en didactique est de visualiser les règles qui relient les différentes erreurs et compétences des élèves lorsqu'ils démontrent une proposition mathématique formulée en langue naturelle française. Pour cela, nous allons discuter une partie des résultats de la pré-expérimentation de notre thèse en cours avec 28 élèves en classe de terminale scientifique à Fianarantsoa (Madagascar). La question principale de recherche est *qu'est-ce qui peut favoriser ou défavoriser la compétence des élèves à prouver une proposition mathématique formulée en langue française ?* Nous faisons l'hypothèse que des difficultés de compréhension en langue française peuvent constituer un obstacle résistant pour pouvoir traduire formellement et prouver une proposition mathématique.

4.1 Choix de l'item et analyse a priori

Nous avons fait passer un questionnaire qui contient plusieurs items à ces élèves de terminale. Dans cette article, nous avons choisi d'explorer les résultats du premier item. Voici la formulation de l'énoncé :

> Traduire la proposition P suivante en langage symbolique et
> la démontrer en montrant les étapes de votre raisonnement :
> P : « Le carré d'un nombre pair est pair »

En se référant aux différents travaux existants en didactique ([18][23][17]), notamment ceux qui concernent la relation entre le langage et logique, on peut trouver différentes interprétations du mot « un » de la proposition P à traduire. Ceci pourrait alors une source de difficulté pour les élèves. En effet, les usages courant de ce mot se différent souvent avec son usage en mathématiques. Ainsi, en mathématiques, le mot « un » marque souvent un élément générique, c'est-à-dire un élément qui peut représenter une classe de variable (cf. exemple 1), ou sous-entend une quantification universelle (cf. exemple 2). Tandis que dans le langage courant, ce mot « un » peut dénoter un élément particulier, c'est-à-dire singulier (cf. exemple 3), ou peut aussi signifier la quantification existentielle « il existe au moins un » (cf. exemple 4).

— Exemple 1 : Choisis un nombre entre 10 et 20
— Exemple 2 : Un carré a quatre angles droits
— Exemple 3 : Un des chiffres de la liste $\{2, 3, 6, 8\}$ est impair
— Exemple 4 : Il y a un chiffre impair dans la liste $\{8, 7, 3, 6, 0\}$

On dénomme ces quatre interprétations du mot « un » comme des variables linguistiques.

La traduction qu'attend l'enseignant de la part des élèves est la suivante :

$$\forall n \in \mathbb{N}, \big((\exists k \in \mathbb{N}; n = 2k) \Rightarrow (\exists k' \in \mathbb{N}; n^2 = 2k')\big)$$

Cependant, cette traduction peut sembler difficile pour les élèves, ainsi ils peuvent donner d'autres alternatives, comme la traduction de P en langage semi-symbolique. On définit ainsi les variables relatives à la traduction formelle de P suivantes :

— une phrase en langue naturelle reformulée : cette première situation fait référence à la traduction de la proposition P par d'autres propositions formulées en langue naturelle
— une phrase formulée en langue semi-symbolique : cette deuxième situation signifie une traduction de la proposition P par l'utilisation en même temps d'un ou des symbole(s) mathématique(s) et des expressions de la langue courante
— une phrase formulée en langage symbolique : ceci fait référence à la traduction de P par des symboles mathématiques uniquement
— une traduction correcte : si la traduction traduit la signification de P
— une traduction incorrecte : si la traduction traduit une autre signification que P

Pour la tâche de démonstration de P, la preuve qu'un enseignant aurait pu faire est la suivante :

> Soit $n \in \mathbb{N}$. Supposons que n est pair, c'est-à-dire qu'il existe un entier naturel k tel que $n = 2k$. En faisant au carré membre à membre, on a : $n^2 = 4k^2 = 2 \times 2k^2$. Et en posant par $k' = 2k^2$ qui est aussi un entier naturel, alors on a $n^2 = 2k'$, c'est-à-dire qu'il est aussi un nombre pair. Ainsi, si n est pair alors n^2 est pair. En conclusion, le carré d'un nombre pair est pair.

Dans Balacheff [1] on peut distinguer quatre types de preuves qui attestent le passage d'une preuve pragmatique vers une preuve intellectuelle, à savoir l'empirisme naïf, l'expérience cruciale, l'exemple générique et l'expérience mentale. Les deux premières sont des preuves pragmatiques dont la validation d'une assertion s'ancrent beaucoup dans les faits et dans l'action ; Tandis que les deux dernières sont des preuves intellectuelles dans laquelle la connaissance est l'objet de réflexions et de discours. Le passage des preuves pragmatiques aux preuves intellectuelles repose sur les niveaux de connaissances mobilisés et sur la formulation de preuve. Plus la preuve est intellectuelle, plus le langage utilisé est formel, c'est-à-dire il y a une utilisation plus ou moins dense des symboles logico-mathématiques qui marque à la fois la décontextualisation, la dépersonnalisation et la détemporalisation de la preuve qui a été produite.

En outre, dans Balacheff [20], il y a une reprécision de ces types de preuves. Ainsi, l'auteur explique que l'exemple générique est un type de preuve à la frontière entre les preuves pragmatiques et les preuves intellectuelles. On distingue ainsi cinq niveaux de preuves, à savoir l'empirisme naïf, l'expérience cruciale, l'exemple générique, l'expérience de la pensée et le calcul des énoncés. Dans cet article, on supprime le type expérience de la pensée et on retient les autres. Ci-dessous, on revient sur les définitions de Balacheff [1][20] de ces types de preuves en donnant un exemple de preuve qu'un élève peut faire pour prouver la proposition P.

1. *empirisme naïf :* c'est une méthode proposée par l'élève qui se focalise sur la vérification sur quelques cas. Cette démarche pragmatique constitue une forme résistante de généralisation [1]. Dans cet exemple, l'élève vérifie la validité de la proposition P avec quelques valeurs particulières :

   > 2 est pair et $2^2 = 4$ est aussi pair, 4 est pair et $4^2 = 16$ est aussi pair, 8 est pair et $8^2 = 64$ est aussi pair, etc.. . .

2. *expérience cruciale :* ce type de preuve se distingue de la précédente en ce que l'individu pose explicitement le problème de généralisation et le résout en pariant sur la réalisation d'un cas qu'il reconnaisse pour aussi peu particulier que possible [1]. Dans l'exemple suivant, l'élève essaie de vérifier la validité de la proposition P avec un grand nombre d'exemples et en prenant des valeurs plus ou moins considérables :

   > $24^2 = 576$. Le carré de 24 est pair. On peut essayer avec 126, on a $126^2 = 15876$. Le carré de 126 est aussi pair. On peut prendre un nombre pair que l'on veut mais le carré sera toujours pair.

3. *exemple générique :* ceci consiste en l'explicitation des raisons de la validité d'une assertion par la réalisation d'opérations ou de transformations sur un objet présent non pour lui-même, mais en tant que représentant caractéristique d'une classe. La formulation dégage les propriétés caractéristiques et les structures d'une classe en restant attachée au nom propre et à l'exhibition de l'un de ses représentants [1]. La preuve par exhaustion des cas est un exemple de ce type de preuve. Dans l'exemple suivant, l'élève connait la propriété d'un nombre pair qu'il se termine par les chiffres 0, 2, 4, 6 ou 8. Ainsi, en faisant au carré un nombre pair, il remarque que le dernier chiffre se termine toujours par ces chiffres.

> $6^2 = 36$, $16^2 = 256$, on peut remarquer que la terminaison du carré d'un nombre qui se termine par 6 est toujours 6. De même, on peut remarquer que la terminaison du carré d'un nombre qui se termine par 0 est toujours 0. D'autre part, la terminaison du carré d'un nombre qui se termine par 2 est 4 et celui qui se termine par 8 est 6. Ainsi dans tous les cas les terminaisons des carrés des nombres qui se terminent par 0, 2, 4, 6 ou 8 sont soit 0 soit 4 soit 6. Ainsi, le carré d'un nombre pair est pair.

4. *calcul sur les énoncés :* c'est un type de preuve intellectuelle dans lequel l'élève utilise un langage plus ou moins formel en introduisant les symboles logico-mathématiques. Dans ce type de preuve le langage utilisé est un véritable outil de calcul intellectuel [20]. La preuve que l'enseignant aurait pu faire qu'on a montrée plus haut fait partie de ce type de preuve. Cependant, un élève peu avancé peut choisir une partie de cette démarche. Voici un exemple dont l'élève ignore toutes les quantifications :

   > Soit n pair, donc $n = 2k$. On a $n^2 = 4k^2 = 2 \times 2k^2 = 2k'$ avec $k' = 2k^2$. Donc n^2 est pair.

En plus des difficultés linguistiques qui concernent surtout la compréhension du mot « un », les élèves peuvent se confronter avec des difficultés mathématiques. On peut catégoriser en deux les difficultés relatives à la compétence mathématique, dont les difficultés concernant les concepts mathématiques en jeu et les difficultés concernant les transformations mathématiques. La première est présente chez un élève lorsqu'il confond un certain concept mathématique avec un autre ou définit mal un concept mathématique. La deuxième est présente lorsque l'élève fait une erreur d'opération pendant la démonstration ou a utilisé une règle erronée, par exemple l'utilisation d'un théorème inapproprié, ou l'utilisation d'une fausse hypothèse.

Le tableau (1) nous montre les différentes variables qu'on va regarder dans les copies des élèves. On les considère comme des variables binaires. On va élaborer le contexte d'extraction correspondant à notre étude, dont l'ensemble $\mathcal{O}$ des objets sont les copies des élèves, l'ensemble $\mathcal{I}$ des variables est formé par les quinze (15) variables du tableau (1) et $\mathcal{R}$ est la relation binaire entre o et x dont i$\mathcal{R}v$ si et seulement si x est présente dans o.

4.2 Résultat

Cette section présente les résultats de notre expérimentation. Dans un premier temps nous allons donner une vue globale des résultats, ensuite nous continuerons la présentation avec des analyses plus approfondies, dont les règles d'association retenues par les deux mesures de support et intensité d'implication.

4.2.1 Présentation générale des résultats

Pour avoir une première vue générale des résultats, nous avons regardé les occurrences de chaque variable dans l'ensemble des données. Cette occurrence des variables nous permet de révéler des informations importantes sur la répartition des données et peut être utilisée pour nous éclairer sur les tendances des élèves sur chaque type de variable. La figure (1) illustre cela.

Pour les difficultés linguistiques, la figure (1) nous montre qu'aucun élève n'a associé le mot « un » avec une quantification existentielle. Cependant, il y a trois élèves, c'est-à-dire 11% des élèves, qui ont associé ce mot avec un élément singulier. Par ailleurs, il semble que les élèves comprennent que ce mot « un » s'utilise en mathématiques pour désigner des génériques (soit

Description	Codification
Difficultés linguistiques	
Le mot « un » renvoie un élément générique	$LGen$
Le mot « un » renvoie un élément singulier	$LSin$
Le mot « un » renvoie une quantification universelle	$LUni$
Le mot « un » renvoie une quantification existentielle	$LExi$
Difficultés mathématiques	
Sur les concepts mathématiques	$DConc$
Sur les règles de transformations mathématiques	$DTrans$
Compétences sur la traduction	
Traduction en langue naturelle	$TNat$
Traduction en langage semi-symbolique	$TSemi$
Traduction en langage symbolique	$TSymb$
Traduction correcte	$TCor$
Traduction incorrecte	$TInco$
Compétences sur la démonstration	
Preuve par empirisme naïf	$PNaif$
Preuve par expérience cruciale	$PCruc$
Preuve par exemple générique	$PGene$
Preuve par calcul sur les énoncés	$PCalc$

TABLE 1 – Description des variables

36%) ou la quantification universelle (soit 46%). Il faut noter aussi que quelques élèves n'ont pas traduit la proposition P.

Les difficultés linguistiques ont une étroite liaison avec les compétences sur la traduction de P. Dans ce dernier, il existe cinq variables, dont $TNat$, $TSemi$, $TSymb$, $TCor$, $TInco$. Les trois premières sont des variables catégorielles, c'est-à-dire que leurs valeurs sont exclusives ; et les deux dernières sont des variables supplémentaires et qui sont aussi catégorielles. C'est-à-dire, les traductions des élèves qui sont faites soit en langue naturelle, soit en langage semi-symbolique, soit en langage symbolique, peuvent être correcte ($TCor$) ou incorrecte ($TInco$). Nous soulignons que pour les élèves qui n'ont pas traduit P, nous n'avons pas comptabilisé sa copie suivant les trois premières variables, par contre nous l'avons comptabilisé dans la catégorie de la variable $TInco$.

La figure (1) nous montre que les élèves ont tendance à traduire la proposition en langage semi-symbolique, soit 61% des élèves. Bien que l'énoncé demande de traduire P en langage symbolique, 11% des élèves l'ont traduit en langage naturel. Enfin 21% ont essayé de traduire en langage complètement symbolique. Ces résultats confirment les difficultés de nombreux élèves avec l'usage du symbolisme logico-mathématique. Le rudiment de la logique semble un obstacle pour la compréhension mathématique. On a la proportion 54% pour $TInco$ contre 46% pour $TCor$. Il faut noter que la plupart des traductions qu'on a jugées correctes sont de traduction en langue naturelle ou en langage semi-symbolique.

Les variables dans la catégorie difficultés mathématiques ne sont pas catégorielles, c'est-à-dire que leurs valeurs ne sont pas exclusives. Ainsi, selon la compétence de l'élève, il peut ou non faire les deux variables qui y figurent. La figure (1) nous montre une grande proportion d'élèves

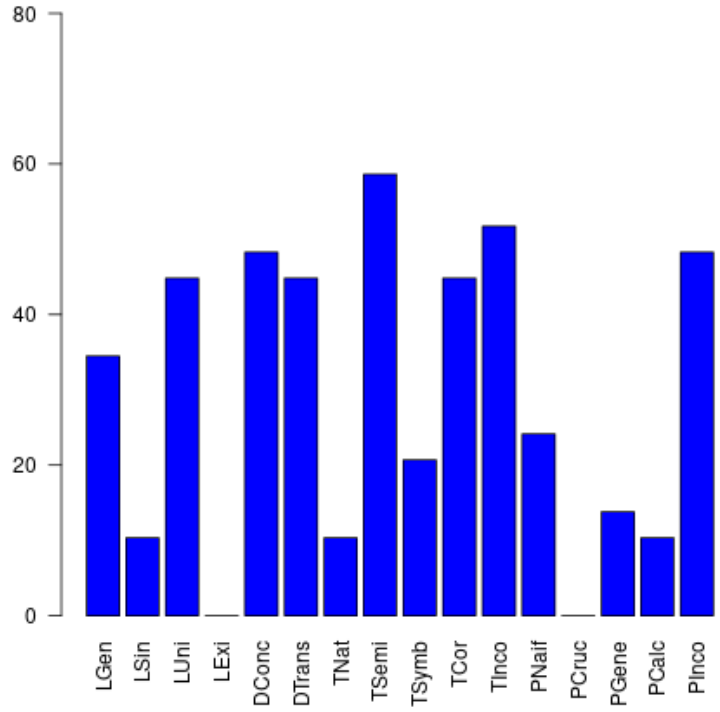

FIGURE 1 – Occurrence de chaque variable (en %).

qui ont commis les deux erreurs sur le concept mathématique ($DConc$) et sur les règles de transformation mathématique ($DTrans$), soit 50% des élèves pour $DConc$ et 46% pour $DTrans$.

Les compétences des élèves à prouver la proposition sont associées à quatre variables, dont $PNaif$, $PCruc$, $PGene$ et $PCalc$. Nous avons remarqué pendant le dépouillement des copies que plusieurs preuves proposées par les élèves ne sont pas compatibles avec P, soit l'élève démontre une autre proposition, soit l'élève démontre par une méthode complètement inadaptée ou incompréhensible. Ainsi, nous avons ajouté la variable $PInco$ qui indique que la preuve est incorrecte.

Le premier constat qu'on a sur la figure (1) est qu'un peu plus de la moitié des élèves ont montré une preuve complètement erronée, soit 54% des élèves. En outre, ce qui ont présenté une preuve compatible à P, les raisonnements des élèves sont en général très naïfs, soit 25% des élèves, c'est-à-dire 50% de ceux qui ont montré une preuve compatible.

Ci-dessous, des transcriptions des copies des élèves sont fournies, car les versions originales ne sont pas très claires. Ces transcriptions permettent une meilleure compréhension du contenu écrit par les élèves.

Transcription de la copie de l'élève E20 :

> 1- Traduisons ces propositions en langage de la logique formelle
> P : $\forall n$ pair $\Rightarrow n = 2k$ ($k \in \mathbb{N}^*$)
> alors : $n^2 = n \Rightarrow n^2 = 2k$

Dans la transcription ci-dessus, le mot « un » est associé avec la quantification universelle. L'élève a compris le sens de la phrase, cependant il a beaucoup de difficulté sur le symbolisme logique, dont l'utilisation de l'implication et des quantifications. En effet, l'élève sait définir un nombre pair, toutefois le caractère existentiel de l'entier k lui pose problème. Il pense que cet entier doit être unique pour tout entier pair.

La traduction dans la transcription suivante, l'élève propose une traduction en langage semi-symbolique. Le sens de la proposition P semble congrue avec la traduction proposée. Cependant, on constate toujours la difficulté sur le symbolisme mathématique. En fait, l'élève ne sait pas utiliser la relation « divise ».

Transcription de la copie de l'élève E6 :

P : « Le carré d'un nombre pair est pair »
1- Traduction en langage de la logique formelle
$\forall n$ pair $\Rightarrow n^2$ pair $\Leftrightarrow n|2 \Rightarrow n^2|2$

La traduction dans la transcription suivante nous montre une association du mot « un » avec un élément singulier.

Transcription de la copie de l'élève E10 :

1- Traduire ces propositions en langage de la logique formelle
soit $n \in \mathbb{N}$, n^2 pair
pour $n = 2$, $2^2 = 4$
4 est pair

Dans la transcription suivante, on voit plusieurs difficultés d'élèves, dont le symbolisme logico-mathématique et la compréhension de la proposition P. Cette traduction signifie plutôt que tout carré d'un entier est pair.

Transcription de la copie de l'élève E2 :

P : « Le carré d'un nombre pair est pair »
1- Traduction en langage de la logique formelle
$\forall n \in \mathbb{N} \Rightarrow n^2 =$ pair $\vee n =$ paire, 2=paire

Dans la transcription suivante on trouve une preuve par calcul sur les énoncés. A part la difficulté sur le symbolisme logique, on constate que les étapes de la démonstration de l'élève sont bien présentées.

Transcription de la copie de l'élève E6 :

2- Démonstration
$\forall n$ pair $\Rightarrow n^2$ pair $\Leftrightarrow n|2 \Rightarrow n^2|2$
montrons que $n^2|2$
$\forall n|2 \Leftrightarrow \exists k \in \mathbb{Z}/n = 2k + r \Rightarrow n^2 = (2k + r)^2$ avec $r = 0$
$n^2 = 4k^2 + 0^2$
$n^2 = 4k^2$
$4k^2|2 \Leftrightarrow \exists 2z/4k = 2 \times 2k^2 + r$ et $r = 0$ donc n^2 pair
conclusion :
$\forall n$ pair $\Rightarrow n^2$ pair $\Leftrightarrow n|2 \Rightarrow n^2|2$

La preuve qu'on trouve dans la transcription suivante n'est pas compatible avec P. En effet, l'élève semble de n'avoir pas compris le sens de n pair. Ainsi, il confond ceci avec le concept de fonction paire.

Transcription de la copie de l'élève E2 :

2- Démonstration
$n^2 =$ paire
$2 \in \mathbb{N} \Leftrightarrow f(n) = f(-n)$
$f(n) = n^2$
$f(-n) = -n^2$
$n^2 = -n^2$
$= n^2 - n^2$
$f(n) = -2n^2 = f(-n)$
donc $n^2 =$ paire si $n \in \mathbb{N}$ et $2 \in \mathbb{N}$

Dans la transcription suivante, la preuve proposée par l'élève montre un empirisme naïf. La validité de P est vérifiée avec des éléments singuliers.

Transcription de la copie de l'élève E3 :

2- Démonstration :
Un nombre est pair s'il est divisible par 2
exemple : $4 \Rightarrow \frac{4}{2} = 2$
4 est divisible par 2 $\Rightarrow$ 4 est un nombre pair
D'après la proposition (A1) :
« le carré d'un nombre pair est pair »vrai
Exemple : On prendra 4 (nombre pair)
$4^2 = 16$ pair car le nombre est divisible par 2
$\frac{16}{2} = 8$
$\Rightarrow$ le carré d'un nombre pair est pair

Nous avons constaté pendant l'exposition des résultats que : *les raisonnements des élèves testés dans le cadre de ce questionnaire pour prouver une proposition mathématique formulée en langue française sont naïfs, voire incompatibles avec le contexte évoqué dans la proposition. Ce niveau est en étroite relation avec la compréhension de P qui se manifeste dans la compétence à la traduire formellement*. La sous-section suivante nous permet d'affirmer ou de faire une réajustement à ce résultat.

4.2.2 Exploration des règles d'association

Nous allons maintenant explorer les règles d'association selon les deux mesures de support et d'intensité d'implication. Dans cette étude, nous avons fixé à 0.25 le seuil minimum pour le support et à 0.70 le seuil minimum pour l'intensité d'implication. Cette choix nous semble raisonnable, car du point de vue du support, cela implique que deux motifs X et Y doivent apparaitre en même temps pour le quart de l'effectif total pour être élu.

En appliquant notre processus algorithmique, nous avons neuf règles (R1 à R9) intéressantes. Le tableau (2) nous montre ces règles.

	Règles	Supports	Intensités d'implication
R1	$DTrans \rightarrow DConc$	0.31	0.78
R2	$DConc \rightarrow PInco$	0.41	0.97
R3	$PInco \rightarrow DConc$	0.41	0.97
R4	$DTrans \rightarrow PInco$	0.37	0.96
R5	$TCor \rightarrow TSemi$	0.34	0.75
R6	$PInco \rightarrow TInco$	0.34	0.78
R7	$DTrans \wedge PInco \rightarrow DConc$	0.31	0.93
R8	$DConc \wedge PInco \rightarrow TInco$	0.31	0.75
R9	$PInco \rightarrow TInco \wedge DConc$	0.31	0.75

TABLE 2 – Les règles d'association intéressantes

Nous pouvons mieux visualiser la lecture des règles entre motifs ne contenant que des singletons par un graphe implicatif. En fait, une règle implicative peut être représentée par une flèche dont les noeuds sont la prémisse et le conséquent. En faisant ainsi, nous avons un graphe orienté qui met en relation les différentes variables. Nous pouvons utiliser pour cela le logiciel CHIC [8], qui est un logiciel dédié pour le traitement des règles implicatives dans le cadre de l'analyse statistique implicative.

Une possibilité qu'offre le logiciel CHIC est la suppression des variables qui ne nous intéressent pas. Ainsi, en supprimant les variables qui n'appartiennent pas à l'ensemble des motifs fréquents, on a le graphe du droite (figure 2). Et pour montrer la différence avec la méthode classique en analyse statistique implicative, qui ne calcule que l'intensité d'implication, on a le graphe de gauche (figure 2).

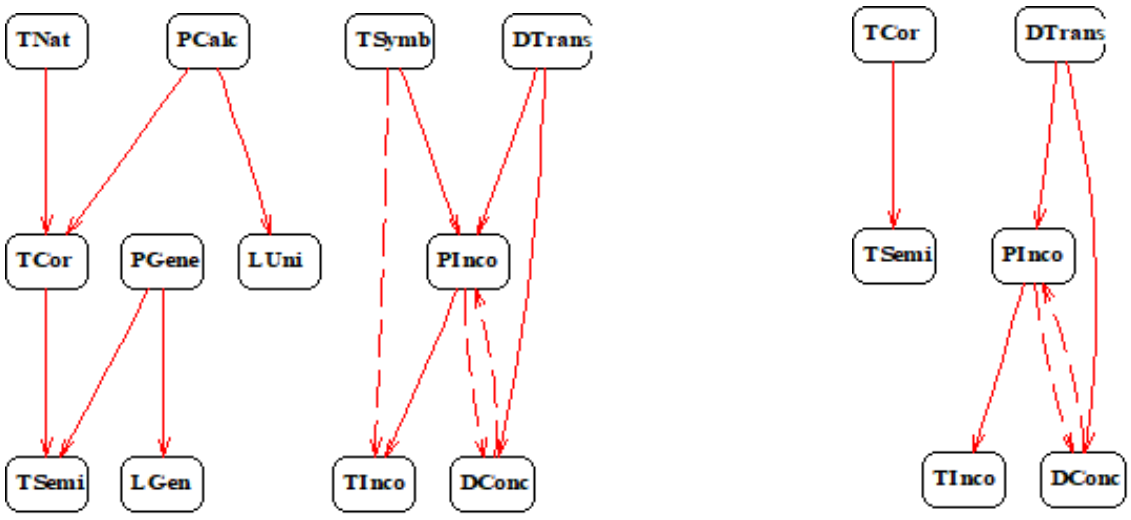

FIGURE 2 – Graphes implicatifs.

En observant le graphe implicatif situé à droite (figure 2), nous remarquons une diminution du nombre de règles obtenues. Ceci s'explique par le fait que notre modèle sélectionne spécifiquement les règles les plus pertinentes. En conséquence, notre approche permet de filtrer et de retenir uniquement les règles considérées comme intéressantes. Dans la section suivante, nous procéderons à une interprétation approfondie du graphe implicatif de droite.

Nous avons l'implication $TCor \rightarrow TSemi$, c'est-à-dire que si la traduction est correcte, alors la traduction est écrite en semi-symbolique. Le résultat du dépouillement nous a montré que 77% des traductions que nous avons jugé correcte sont écrites en langage semi-symbolique.

La figure (2) nous montre aussi que la difficulté sur les règles de transformation mathématique est la source de beaucoup de difficultés sur la compréhension mathématique. En effet, ce sont les élèves qui ont utilisé un théorème faux ou inapproprié, ou ont pris une autre hypothèse, ou ont fait des erreurs sur les règles de calcul qui régissent la procédure de raisonnement pour démontrer la proposition P, qui ont produit une preuve et une traduction erronées, ainsi que des erreurs sur les concepts mathématiques en jeu.

D'autre part les règles entre motifs qui ne sont plus des singletons nous révèlent des informations complémentaires aux résultats précédents. La règle R7 nous informe que lorsqu'on observe dans une copie une difficulté sur les règles de transformation mathématique et une preuve incorrecte, alors il y a des erreurs sur les concepts en jeu dans la démonstration. La règle R8 témoigne que la combinaison de la difficulté sur les concepts mathématiques en jeu et la preuve incorrecte entraine une traduction fausse. Enfin la règle R9, qui vient du même motif que R8, informe que la preuve incorrecte entraine la combinaison des difficultés sur la traduction et sur les concepts mathématiques en jeu.

Ces analyses complètent les résultats constatés dans la sous-section précédente :

— *les élèves sont plus habiles sur l'utilisation de langage semi-symbolique. Ceci marque leurs difficultés sur la formalisation logique des concepts logico-mathématiques comme l'implication, la quantification, la parité d'un nombre, etc.*... ;

— *les compétences à prouver une proposition, à la traduire formellement et les difficultés mathématiques se combinent, s'alimentent entre elles et entraînent des raisonnements complètement erronés.*

4.3 Discussion

4.3.1 La méthode d'extraction des règles d'association

Dans le cadre de cet article, notre premier objectif réside dans la proposition d'une méthode d'extraction des règles d'association en mettant en œuvre la mesure classique d'intensité d'implication. L'originalité de cette approche se manifeste par la capacité à générer un nombre limité de règles, tout en privilégiant celles qui revêtent un intérêt particulier. Cette méthodologie se distingue par sa pertinence, axée sur la qualité plutôt que la quantité des règles obtenues.

Cependant, afin d'évaluer exhaustivement l'efficacité de cette méthode, une étude comparative s'avère nécessaire. Cette comparaison sera réalisée avec la méthode traditionnelle de l'algorithme Apriori d'Agrawal & Srikant [4], largement reconnue dans le domaine de l'extraction de règles d'association. L'objectif de cette démarche comparative est d'analyser les performances respectives de chaque méthode en termes de pertinence, d'efficacité et de capacité à extraire des règles significatives. Une telle étude comparative contribuera à éclairer davantage la valeur ajoutée de la nouvelle méthode proposée par rapport aux approches plus classiques.

En fixant les seuils de support et de confiance respectivement à 0.25 et 0.70, l'algorithme Apriori [4] produit un ensemble de treize règles. Le tableau 3 expose en détail ces règles obtenues :

	Règles	Confiances
r1	$LGen \rightarrow TSemi$	0.70
r2	$TCor \rightarrow TSemi$	0.76
r3	$PInco \rightarrow DConc$	0.85
r4	$DConc \rightarrow PInco$	0.85
r5	$DTrans \rightarrow PInco$	0.84
r6	$TInco \wedge PInco \rightarrow DTrans$	0.70
r7	$TInco \wedge PInco \rightarrow DConc$	0.90
r8	$DTrans \wedge PInco \rightarrow DConc$	0.81
r9	$DTrans \wedge TInco \rightarrow DConc$	0.85
r10	$TSemi \wedge PInco \rightarrow DConc$	0.85
r11	$TSemi \wedge PInco \rightarrow DTrans$	0.85
r12	$DTrans \wedge TInco \wedge PInco \rightarrow DConc$	0.85
r13	$DTrans \wedge TInco \rightarrow PInco \wedge DConc$	0.85

TABLE 3 – Les règles d'association obtenues par l'algorithme Apriori

Lors d'une comparaison entre les tableaux 2 et 3, il est observé que six règles présentent une similarité remarquable. Plus précisément, les règles r2, r3, r4, r5 et r8 du tableau 3 correspondent respectivement aux règles R5, R3, R2, R4 et R7 du tableau 2. Cette similitude met en évidence une correspondance significative entre l'algorithme Apriori [4] et notre méthode d'extraction.

De plus, il est important de noter que les règles R1, R6, R8 et R9 du tableau 2 ne sont pas présentes dans le tableau 3. Cette observation souligne un avantage supplémentaire de notre méthode par rapport à l'algorithme Apriori. Par ailleurs, une analyse approfondie du tableau 3 révèle que plusieurs de ces règles sont redondantes, ne fournissant aucune information additionnelle. En réalité, seule la règle r12 semble apporter les informations fournies par les règles r6, r7, r8, r9 et r13. Cette redondance de règles peut compliquer le travail des spécialistes chargés de les interpréter.

Toutefois, il est également à noter la présence de règles dans le tableau 3 qui ne sont pas répertoriées dans le tableau 2, notamment les règles r1, r6, r9, r10, r11, r12 et r13. Parmi ces règles, r1, r10 et r11 apportent des informations complémentaires. Cela suggère que certaines informations ne peuvent pas être générées par notre méthode d'extraction seule. Par conséquent, l'utilisation conjointe des deux méthodes d'extraction pourrait être complémentaire, permettant ainsi de tirer parti des forces respectives de l'algorithme Apriori [4] et de notre approche pour une analyse plus approfondie et complète des règles d'association.

4.3.2 Analyse exploratoire des règles

Les résultats de notre étude en didactique mettent en évidence des aspects spécifiques concernant la hiérarchisation des difficultés rencontrées par les élèves en mathématiques. Nous avons observé que les élèves démontraient une plus grande habileté dans l'utilisation du langage semi-symbolique, suggérant ainsi une certaine aisance pour exprimer informellement les concepts mathématiques. Cependant, cette difficulté linguistique a mis en lumière leurs difficultés dans la formalisation logique des concepts tels que l'implication, la quantification, la parité d'un nombre, et d'autres notions complexes. Cette hiérarchisation des difficultés souligne l'importance de cibler spécifiquement ces points faibles dans le développement de stratégies pédagogiques appropriées.

De plus, nos résultats ont révélé une corrélation significative entre les compétences en résolution de problèmes mathématiques et les obstacles rencontrés dans la formalisation logique des propositions. Ces deux aspects semblent s'alimenter mutuellement, créant ainsi des raisonnements complètement erronés. Ce constat a été déjà fait par Chellougui [7] qui affirme en particulier que les énoncés avec un conditionnel implicitement quantifié peuvent amener les élèves dans des difficultés. Ceci souligne l'importance d'une approche holistique pour aborder ces problématiques, en prenant en compte à la fois les compétences linguistiques et logiques dans l'apprentissage des mathématiques.

Notre recherche a conforté l'hypothèse selon laquelle l'incompréhension de la langue française peut constituer un obstacle résistant à la formalisation des propositions mathématiques. Lorsque les concepts mathématiques font appel à des terminologies et à des expressions spécifiques, les élèves pourraient rencontrer des difficultés supplémentaires dans leur capacité à traduire ces concepts en langage formel.

En nous appuyant sur les travaux d'eduscol [23], nous soulignons que les objets mathématiques sont abstraits et que leurs définitions, leurs propriétés, ainsi que les preuves de ces propriétés, ont une forte dimension formelle. La formalisation logique des concepts mathématiques est donc une étape cruciale dans l'apprentissage des mathématiques, et les difficultés que nous avons identifiées chez les élèves dans ce domaine méritent une attention particulière.

Les recherches de Duval [3] soulignent également l'importance pour les élèves d'être capables d'appréhender un objet mathématique dans plusieurs registres et de coordonner ces registres. Cette compétence est un enjeu essentiel dans l'apprentissage des mathématiques, car elle permet de mieux saisir les concepts et de les formaliser de manière logique.

Par ailleurs, les travaux de Fabert & Grenier [12] mettent en évidence le fait que les connaissances des élèves en matière de raisonnement mathématique et de logique sont souvent peu stables voire absentes, ce qui peut constituer des obstacles supplémentaires dans leur apprentissage des mathématiques.

V CONCLUSION

Dans cette étude, nous avons présenté une méthodologie innovante pour utiliser la mesure classique de l'intensité d'implication proposée par Gras & Régnier [9] afin d'extraire des connaissances dans le domaine de la didactique des mathématiques. Notre méthode s'est avérée pertinente en permettant de ne retenir que les règles les plus pertinentes, facilitant ainsi l'identification de relations importantes entre les compétences et les difficultés des élèves dans la démonstration de propositions mathématiques formulées en langue française.

Notre recherche présente cependant certaines limites importantes qu'il convient de mentionner. Tout d'abord, nous n'avons pas pu étudier les impacts de la langue malagasy, qui est utilisée par les enseignants pour expliquer les leçons à l'oral. La langue malagasy peut avoir une influence sur la compréhension des concepts mathématiques chez les élèves, et son rôle potentiel dans les difficultés de formalisation logique aurait pu apporter des éclairages supplémentaires. Des recherches futures devraient donc envisager d'inclure cette dimension linguistique pour une compréhension plus approfondie de l'impact des langues d'enseignement sur l'apprentissage des mathématiques.

De plus, la taille de la population étudiée constitue une faiblesse majeure de notre recherche. La petitesse de l'échantillon limite la généralisation des résultats à une population plus vaste

d'élèves. Malgré cette limitation, nos résultats fournissent des informations importantes pour la réflexion sur l'enseignement des mathématiques. En prenant en compte la hiérarchisation des difficultés rencontrées par les élèves et les interactions complexes entre compétences linguistiques et logiques, nous sommes mieux préparés pour élaborer des stratégies pédagogiques plus ciblées et plus efficaces, visant à améliorer la compréhension et la formalisation logique des concepts mathématiques.

RÉFÉRENCES

[1] N. BALACHEFF. « Processus de preuve et situations de validation ». In : *Educational studies in mathmatics* 18 (1987), pages 147-176.

[2] J. PICHON. *Théorie des ensembles. Logique des entiers*. Ellipses, 1989, ISBN 2729889418, 1989.

[3] R. DUVAL. « Registres de représentation sémiotique et fonctionnement cognitif de la penssée ». In : *Annales de didactique et de sciences cognitives* 5 (1993), pages 37-65.

[4] R. AGRAWAL et S. SRIKANT. « Fast algorithm for mining association rules in large databases ». In : *20th Intl. Conf. on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann, 2000, pages 478-499.

[5] R. LEHN, F. GUILLET et P. KUNTZ. « Felix : un outil interactif d'aide à la fouille de connaissances s'appuyant sur l'intensité d'implication ». In : *Researchgate* (2000), récupéré dans https://www.researchgate.net/publication/255621413.

[6] M. G. OTTAVIANI et S. ZANNONI. « Implication statistique et recherche en didactique. Utilisation d'un outil non symétrique d'analyse de données pour l'interprétation des résultats d'un test d'évaluation ». In : *Mathématiques et sciences humaines* 154-155 (2001), pages 61-79.

[7] F. CHELLOUGUI. « L'utilisation des quantificateurs dans l'enseignement secondaire tunisien ». In : *L'ouvert* 108 (2003), pages 1-8.

[8] R. COUTURIER et R. GRAS. « CHIC : traitement de données avec l'analyse implicatve ». In : *Reseachgate* (2005), récupéré dans https://www.researchgate.net/publication/220786956.

[9] R. GRAS et J. C. RÉGNIER. *Fondements théoriques de l'analyse statistique implicative, partie 1*. Université de Nantes, Université de Lyon : Tome E-16 du RNTI, 2009.

[10] LAHANIER-REUTER. « Analyse statistique implicative et didactique des mathématiques ». In : *Tome E-16 du RNTI, Cépaduès-Editions* (2009).

[11] F. PENNERATH. *Méthodes d'extraction de connaissances à partir de données modélisables par des graphes. Application à des problèmes de synthèse organique*. Nancy 1 : Thèse de doctorat de l'Université Henri Poincaré, 2009.

[12] C. FABERT et D. GRENIER. « Une étude didactique de quelques éléments de raisonnement mathématique et de logique ». In : *Petit x* 87 (2011), pages 31-52.

[13] R. GRAS et R. COUTURIER. « Spécificités de l'analyse statistique implicative par rapport à d'autres mesures de qualité de règles d'association ». In : *Educ. Matem. Pesq.* 15.2 (2013), pages 249-291.

[14] R. GRAS, P. KUNTZ et H. BRIAND. « Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données ». In : *Mathématiques et sciences humaines* 154-155 (2013), pages 9-29.

[15] D. GRISSA. *Etude comprtementale des mesures d'intérêt d'extraction de connaissances*. Thèse de doctorat de l'Université Blaise Pascal et de l'Université Tunis El Manar, 2013.

[16] B. IDIRI. *Méthodologie d'extraction de connaissances spatio-temporelles par fouille de données pour l'analyse de comportements à risque. Application à la surveillance maritime*. Thèse de doctorat Paris-Tech, 2013.

[17] J. NJOMGANG-NGANSOP. *Enseigner les concepts de logique dans l'espace mathématique francophone : aspect épistemologique, didactique et langagier. Une étude de cas au Cameroun*. Thèse de doctorat de l'Université Claude Bernard Lyon 1 et de l'Université de Yaoundé 1, 2013.

[18] D. GRENIER. « Logique et raisonnements mathématiques pour l'enseignement au collège et au lycée ». In : *Atelier : Problèmes pour apprendre la logique et le raisonnement*. Limoges, 2016.

[19] H. RAMANANTSOA. *Contributions à l'amélioration de génération des bases des règles d'association MGK-valides et applications en didactique des mathéquatiques*. Madagascar : Thèse de doctorat de l'Université d'Antananarivo, 2016.

[20] N. BALACHEFF. « Contrôle, preuve et détermination. Trois régimes de la validation ». In : *Séminaire DDM*. Paris - France, 2017, pages 1-34.

[21] A. AIT-MLOUK. *Fouille de données et analyse de qualité des règles d'association dans les bases de données massives : Application dans le domaine de la sécurité routière*. Maroc : Thèse de doctorat de l'Université de Marrakech, 2018.

[22] H. F. RAKOTOMALALA. *Classification hiérarchique implicative et cohésitive selon la mesure MGK : Application en didactique de l'informatique*. Madagascar : Thèse de doctorat de l'Université d'Antananarivo, 2019.

[23] *Mathématiques et maîtrise de la langue*. eduscol.education.fr/ressources-2016.