Actes EGC'06, pp. 569-590, Cepadues RNTI-E-6.

Utilisation des réseaux bayésiens dans le cadre de l'extraction de règles d'association

Clément Fauré[*,**], Sylvie Delprat[*]
Alain Mille[***], Jean-François Boulicaut[**]

[*]EADS CCR, Centreda 1, F-31700 Blagnac
{clement.faure, sylvie.delprat}@eads.net
[**]LIRIS UMR 5205, INSA Lyon, Bâtiment Blaise Pascal,
F-69621 Villeurbanne cedex
[***]LIRIS UMR 5205, Université Lyon 1, Nautibus,
F-69622 Villeurbanne cedex
{amille, jboulica}@liris.cnrs.fr

Résumé. Cet article aborde le problème de l'utilisation d'un modèle de connaissance dans un contexte de fouille de données. L'approche méthodologique proposée montre l'intérêt de la mise en œuvre de réseaux bayésiens couplée à l'extraction de règles d'association dites delta-fortes (membre gauche minimal, fréquence minimale et niveau de confiance contrôlé). La découverte de règles potentiellement utiles est alors facilitée par l'exploitation des connaissances décrites par l'expert et représentées dans le réseau bayésien. Cette approche est validée sur un cas d'application concernant la fouille de données d'interruptions opérationnelles dans l'industrie aéronautique.

1 Introduction

Un des objectifs de l'extraction de connaissances à partir de données consiste à fournir des énoncés valides et utiles aux utilisateurs propriétaires de ces données. L'utilité de ces énoncés est d'autant plus grande qu'ils décrivent une réalité du domaine non encore explicitée jusqu'ici, autrement dit, une nouvelle connaissance.
Nous nous intéressons à l'extraction de connaissances au moyen de règles descriptives comme les règles d'association (Agrawal et al., 1993). Les problèmes posés par l'extraction de telles règles ont été étudiés intensivement ces dix dernières années. Bien que l'extraction de toutes les règles fréquentes et valides soit difficile dans de grands jeux de données, des dizaines d'algorithmes efficaces ont été proposés (Goethals et Zaki, 2003, par exemple). Un second problème concerne le nombre considérable de règles qui peuvent être fréquentes et valides et donc extraites. Une première solution consiste à rechercher des couvertures des ensembles de règles, ou si l'on préfère, à éliminer des règles redondantes. Des travaux importants dans cette direction concernent l'exploitation de représentations condensées des ensembles fréquents comme les ensembles fermés (Pasquier et al., 1999; Boulicaut et al., 2000) ou bien les ensembles δ-libres (Boulicaut et al., 2003). (Jeudy, 2002) est une étude assez complète de ces propositions.

Par exemple, les règles dites δ-fortes, car construites à partir d'ensembles fréquents δ-libres, ont des propriétés intéressantes : membre gauche minimal, fréquence minimale mais aussi niveau de confiance controlé par le nombre d'exceptions toléré (le paramètre δ) pour la règle (voir Becquet et al., 2002, pour une application en biologie moléculaire).
Cependant, l'élimination des redondances indépendamment du domaine d'application montre clairement ses limites. Pour éviter de présenter aux utilisateurs experts des milliers de règles fréquentes, valides et non redondantes, il faut alors travailler soit avec d'autres mesures d'intérêt objectives (i.e., au delà des seules mesures de fréquence et de confiance), soit assister la prise en compte de l'intérêt subjectif de l'analyste. La première direction de travail a donné lieu à de multiples propositions (voir par exemple Azé, 2003, pour une synthèse récente). Une seconde direction de travail consiste à assister le post-traitement des collections de règles extraites pour la prise en compte de la connaissance du domaine et ainsi éviter de présenter des informations triviales et/ou attendues. Notre hypothèse de travail est que les règles dites « intéressantes » sont celles qui non seulement satisfont certaines contraintes sur des mesures d'intérêt objectives (e.g., fréquence minimale, confiance suffisante) mais aussi sortent du cadre des connaissances existantes pour l'utilisateur. Ainsi, il est nécessaire de s'intéresser à la modélisation et l'exploitation des connaissances de l'expert dans un contexte d'extraction de règles d'association. Les travaux de Padmanabhan et Tuzhilin (1998) ont montré l'utilisation des connaissances de l'utilisateur par la définition de règles. Cette approche a ensuite été formalisée par un réseau de croyances (Padmanabhan et Tuzhilin, 2000) permettant d'extraire l'ensemble minimum des règles d'association en fonction des connaissances du domaine. Ce type d'approche présente cependant une limitation. En effet, une règle est jugée intéressante si elle diffère des règles définies dans le réseau de croyance, et non pas par rapport à ce que l'on pourrait inférer de ces croyances. Cette notion d'inférence a été développée dans Jaroszewicz et Simovici (2004); Jaroszewicz et Scheffer (2005). Ces auteurs décrivent l'utilisation d'un réseau bayésien pour calculer l'intérêt d'ensembles d'attributs extraits à l'aide d'un algorithme de type Apriori (Agrawal et al., 1996). La différence entre le support estimé sur les données et le support inféré à partir du réseau bayésien est calculée pour chaque ensemble d'attributs. Les motifs les plus intéressants sont ceux pour lesquels la divergence entre les connaissances de l'utilisateur (i.e., l'évaluation au moyen du réseau) et ce qui est observé dans les données réelles est la plus forte. Ces ensembles d'attributs sont ensuite soumis à l'utilisateur pour une éventuelle mise à jour de la structure ou des paramètres du réseau bayésien.
Dans cet article, nous proposons également une approche méthodologique pour exploiter des connaissances du domaine dans le cadre de la découverte de motifs locaux intéressants, typiquement des règles d'association. Nous validerons cette approche sur un cas d'application réel en aéronautique. La section 2 présente l'approche envisagée. La section 3 introduit les notations et détaille la solution proposée. La section 4 décrit le cas d'application utilisé et les expérimentations qui ont été menées. Enfin la dernière section est une brève conclusion.

2 Approche envisagée

Nous décrivons l'approche méthodologique envisagée pour faciliter le processus de découverte de connaissances à bases de motifs fréquents. On peut considérer quatre étapes importantes :

– explicitation et modélisation des connaissances a priori de l'expert,

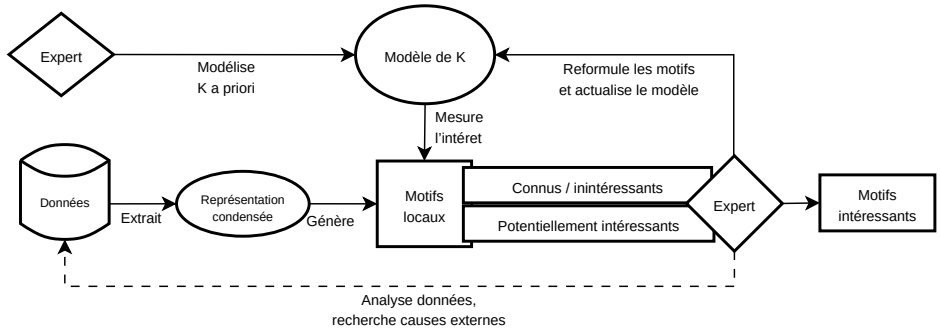

FIG. 1 – *Aperçu du processus de découverte de connaissances étudié*

– extraction d'une représentation condensée des motifs fréquents dans les données,
– utilisation du modèle de connaissance pour faciliter la lecture des motifs extraits,
– analyse et interprétation des motifs extraits.

La figure 1 reprend les différentes étapes de ce processus de découverte de connaissances et en montre une vue globale.

2.1 Explicitation et modélisation des connaissances a priori

Cette phase de modélisation a pour principal objectif de représenter les connaissances dont l'expert dispose par rapport au domaine d'application. Il est possible, au départ, de modéliser uniquement les connaissances les plus évidentes. Puis, au fur et à mesure de l'exploration des différents résultats d'extractions, l'expert peut souhaiter améliorer son modèle pour faciliter la découverte de motifs toujours plus intéressants en éliminant ceux qui apparaissent triviaux au regard des connaissances déjà connues. Pour ce faire, il faut disposer d'un formalisme de représentation adapté.

L'approche par réseau bayésien pour la modélisation des connaissances apparaît comme particulièrement adaptée. En effet, les réseaux bayésiens permettent de considérer dans un formalisme commun les modèles de causalité[1] et les probabilités. Ils ont aussi été utilisés de manière intensive pour des applications de modélisation de la connaissance (Naïm et al., 2004, chapitre 8) et il existe de nombreux outils pour les construire et les exploiter.

Dans notre cadre d'application, il n'est pas souhaitable de réaliser un apprentissage automatique de la structure et des paramètres du réseau bayésien. En effet, on s'intéresse aux réseaux bayésiens pour leur capacité à représenter de manière compacte et intelligible la connaissance d'un expert plutôt qu'à une utilisation « directe » (prédiction, aide à la décision, etc.). Ainsi, la structure du réseau est définie, puis mise à jour par l'utilisateur expert du domaine. L'apprentissage automatique des paramètres du réseau est possible (Heckerman, 1997, notamment), mais cette solution n'a pour l'instant pas été implémentée.

[1]La notion de causalité est à prendre ici au sens *intuitif* du terme, en ce sens qu'elle est plus facilement compréhensible par un expert que la notion de corrélation statistique.

2.2 Extraction d'une représentation condensée des ensembles fréquents

Cette phase concerne l'utilisation d'un algorithme d'extractions de motifs, en l'occurrence des règles d'association. Les algorithmes de type Apriori permettent d'extraire toutes les règles d'association au dessus d'un certain seuil de fréquence et de confiance (spécifiés par l'utilisateur). Un premier reproche classique vis à vis des algorithmes de ce type est qu'ils ne sont pas utilisables sur des volumes denses et/ou fortement corrélées, tout du moins pour des seuils de fréquences qui paraissent pertinents aux experts. Un second problème vient du fait que toutes les règles qui satisfont les contraintes de fréquence et de confiance sont extraites. La question de la redondance de ces collections, quel que soit le domaine d'application a été très étudiée (voir les nombreuses propositions de couvertures de collections de règles). Par contre, il y a encore peu de travaux pour l'élimination de motifs redondants au regard des connaissances déjà acquises par l'expert.
Pour résoudre le premier problème, on utilisera un algorithme (Boulicaut et al., 2000) capable d'extraire une représentation condensées des ensembles fréquents, les ensembles dits δ-libres fréquents. Cet algorithme permet également, en calculant la δ-fermetures de tels ensembles, de produire une collection concise de règles d'association à forte confiance appelées règles δ-fortes. En effet, le paramètre δ détermine le nombre d'exceptions toléré pour les règles et sa valeur est supposée être petite au regard du seuil de fréquence utilisé. Pasquier (2000) a d'ailleurs étudié les propriétés de ces collections lorsque $\delta = 0$.
Pour résoudre le second problème, il faut intégrer la connaissance de l'expert au calcul de l'intérêt des règles d'association extraites. A partir de la proposition décrite dans Jaroszewicz et Simovici (2004), nous définirons une mesure d'intérêt des règles δ-fortes. Cependant, l'objectif de notre démarche ne consiste pas seulement à définir une nouvelle mesure de similarité, mais plutôt à mettre en place un cadre méthodologique permettant l'exploitation de la connaissance expert et s'appuyant sur des algorithmes d'extraction efficaces. Certains liens importants entre les réseaux bayésiens et les règles d'association seront aussi explicités.

2.3 Utilisation du modèle de connaissance pour faciliter la lecture des motifs extraits

En utilisant le formalisme des réseaux bayésiens, l'expert explicite les connaissances qui vont lui être utiles pour le processus de fouille de données. Ce modèle de connaissance a pour but de faciliter la découverte de motifs pertinents, c'est-à-dire ceux qui ne sont pas pris en compte par le modèle de connaissance, ou ceux qui le contredisent.
Plus précisément, ce modèle permet à l'utilisateur de définir -a priori- des dépendances entre des attributs qu'il ne souhaite pas retrouver dans les résultats de l'extraction. Pour cela, on utilise les capacités d'inférence du réseau bayésien pour mesurer l'intérêt des règles extraites, en comparant les dépendances déduites du réseau bayésien (construit à partir des connaissances du domaine) et les règles d'association (extraites à partir des données réelles). Une divergence forte indique un motif potentiellement intéressant ; inversement, une convergence entre les données réelles et l'estimation effectuée à partir du réseau bayésien indique des motifs déjà connus.

2.4 Analyse et interprétation des motifs extraits

Il est peu probable que les premières itérations du modèle parviennent à éliminer correctement les motifs non intéressants. Par contre, nous pensons que des mises à jour successives du modèle de connaissance vont pouvoir s'appuyer sur les extractions réalisées. Ainsi, à chaque itération du processus, deux possibilités se présentent :

- *Les résultats de l'extraction font apparaître des motifs connus* ; ce cas de figure nécessite de la part de l'expert une reformulation des motifs découverts, de manière à pouvoir intégrer de nouvelles dépendances dans le modèle de connaissance. Le but est à la fois de pouvoir éviter par la suite la présentation de ce type de motifs, mais aussi de capitaliser une certaine connaissance du domaine sous la forme de dépendances quantitatives et qualitatives entres les variables du domaine.
- *Les résultats de l'extraction présentent des motifs potentiellement intéressants* (du point de vue de l'expert) ; cela implique généralement un travail d'analyse et de recherche dans l'ensemble des données relatives au domaine d'application. L'expert peut ensuite déterminer si les motifs sont effectivement porteurs de nouvelles connaissances ou si ils révèlent une insuffisance (en terme de représentativité ou d'exhaustivité des attributs pris en compte) des données disponibles. L'expert peut, par exemple, décider d'enrichir la base de données en intégrant de nouvelles variables (attributs), avant de procéder à une nouvelle itération du processus.

Ainsi, l'expert doit reformuler les connaissances induites par les motifs découverts afin de pouvoir les intégrer progressivement à son modèle. Nous allons ainsi pouvoir éliminer progressivement les règles triviales ou connues et faciliter l'émergence de règles plus intéressantes. Notons que, dans ce contexte, le réseau bayésien représente un modèle partiel (et dégradé) des connaissances de l'expert. Il n'est défini et utilisé que pour faciliter la lecture et l'interprétation de règles d'association extraites (post-traitement des règles).

3 Réseaux bayésiens et motifs fréquents

3.1 Définitions et notations

Soit BD une base de données booléenne, et $H = \{A_1, A_2, \ldots, A_n\}$ l'ensemble de ses attributs booléens. H est défini sur $D_H = D_{A_1} \times D_{A_2} \times \ldots \times D_{A_n}$. $P_I^{BD}(i)$ dénote la probabilité pour que l'ensemble d'attributs $I \subseteq H$ prenne comme valeur le vecteur i. Un itemset est représenté par la paire (I, i) avec $I \in H$ ensemble d'attributs fini non vide et i ensemble des valeurs des attributs de I. Lorsque cela n'est pas strictement nécessaire, l'itemset (I, i) sera désigné simplement par I.

Un réseau bayésien RB est un graphe dirigé acyclique défini par un ensemble de noeuds correspondants aux attributs de H et par $E \subset H \times H$ l'ensemble des arcs du graphe. A chaque nœud on associe une distribution de probabilité conditionnelle $P_{A_i|\Pi_{A_i}}$, où $\Pi_{A_i} = \{A_j|(V_{A_j}, V_{A_i}) \in E\}$ représente les parents du nœud A_i. Pour une discussion détaillée sur les réseaux bayésiens consulter Pearl (1988). Une des propriétés du réseau bayésien est de définir de manière unique la distribution de probabilité jointe de H :

$$P_H^{RB} = \prod_{i=1}^{n} P_{A_i|\Pi_{A_i}} \tag{1}$$

Une règle d'association R est un motif $X \Rightarrow Y$, où X et Y sont des itemsets tels que $Y \neq \emptyset$ et $X \cap Y = \emptyset$. X est appelé *partie gauche* de la règle, Y la *partie droite*. Soit I un itemset, le support de I dans BD, noté $supp_{BD}(I)$, est l'ensemble des lignes (ou transactions) de BD qui contiennent I.
Ainsi, étant donné une base de donnée BD définie sur un ensemble d'attributs H et un réseau bayésien RB, il est possible d'obtenir la confiance d'une règle d'association $R = X \Rightarrow Y$ (voir Dechter, 1999, pour un exemple d'algorithme d'inférence) :

$$\begin{aligned} \text{conf}_{RB}(X \Rightarrow Y) &= P_{Y|X}, \\ &= \prod_{i=1}^{m} P_{Y_i|\Pi_{Y_i}} \end{aligned} \tag{2}$$

et par estimation sur les données :

$$\text{conf}_{BD}(X \Rightarrow Y) = \frac{\text{supp}_{BD}(X \cup Y)}{\text{supp}_{BD}(X)} \tag{3}$$

On travaille sur une représentation condensée des itemsets fréquents au moyen d'itemsets δ-libres et fréquents. En fait, l'algorithme utilisé produit une collection de couples $(I, \delta - fermeture(I)\backslash I)$. Chaque élément I est un itemset δ-libre fréquent. Sa δ-fermeture est l'ensemble de tous les attributs qui sont vrais pour un enregistrement lorsque ceux de I le sont à δ exceptions près. Il s'agit donc d'une généralisation de la notion classique de cloture au sens de la connection de Galois puisque, lorsque $\delta = 0$, $I \cup \delta - fermeture(I)$ est un ensemble fermé fréquent. L'entier positif δ permet donc de borner le nombre d'exceptions[2] d'une règle dite δ-forte, i.e., une règle R de la forme $I \Rightarrow \delta - fermeture(I)\backslash I$. Il faut comprendre que -par construction- les règles ainsi générées ont une partie gauche minimale et une partie droite maximale, ce qui implique une confiance maximale de 1 sur BD lorsque la règle ne comporte pas d'exceptions (par exemple, lorsque l'on exige $\delta = 0$.
En s'inspirant de Jaroszewicz et Simovici (2004), nous définissons maintenant une mesure de l'intérêt d'une règle d'association. Cette mesure est basée sur la différence entre la confiance de la règle estimée à partir des données et celle inférée par le réseau bayésien. Elle s'exprime de la manière suivante :

$$\text{Int}(R) = |\text{conf}_{BD}(R) - \text{conf}_{RB}(R)| \tag{4}$$

3.2 Exploitation du modèle de connaissance

Nous disposons d'un algorithme qui calcule une collection de règles d'association δ-fortes, d'un formalisme pour modéliser les connaissances a priori de l'expert, ainsi que d'une mesure prenant en compte ces connaissances pour évaluer l'intérêt des règles.
Pour se faire une idée du comportement de la mesure d'intérêt et des liens existants entre les règles extraites et les implications au niveau du modèle de connaissance, nous allons décrire de manière empirique deux cas de découverte de règles. Une règle peut représenter (1) un motif connu de l'expert, mais qui n'est pas encore pris en compte par le réseau bayésien ou (2) un motif connu et pris en compte par le réseau.

[2]Une exception à une règle $X \Rightarrow Y$ est un enregistrement de la base qui contient X mais pas Y

Soit $R = A \Rightarrow B$ le motif extrait ($\text{conf}_{BD}(R)$ est proche de 1 puisque nous ne calculons que des règles δ-fortes). Dans le cas (1), R n'est pas prise en compte par le réseau bayésien. On se place alors sous l'hypothèse d'indépendance entre A et B pour calculer l'intérêt de R selon le réseau bayésien. Ainsi, à partir de l'équation 4, on obtient : $\text{Int}(R) = |\text{conf}_{BD}(R) - P_B|$. Ici, l'intérêt dépend donc principalement de la distribution de probabilité P_B définie dans le réseau bayésien. Par exemple, si l'expert a défini $P_B = 0,95$, la règle R aura un intérêt presque nul car B est un événement très fréquent. Dans ce cas, l'association n'apporte pas de connaissance supplémentaire. Inversement, si P_B est faible, alors l'intérêt de la règle sera élevé, signifiant alors à l'utilisateur la possible existance d'une dépendance entre A et B.
On se place maintenant dans le cas (2) où la règle R est prise en compte par le réseau bayésien. Cela signifie que l'on a défini explicitement un lien de dépendance $A \rightarrow B$ ainsi que la probabilité $P_{B|A}$. On a alors $\text{Int}(R) = |\text{conf}_{BD}(R) - P_{B|A}|$, soit un intérêt dépendant de $P_{B|A}$; ce qui correspond bien au comportement souhaité. En effet, dans le cas où $P_{B|A}$ est proche de 1, l'association n'est pas jugée intéressante car elle est correspond au modèle défini par l'expert. Inversement, si l'on a défini $P_{B|A}$ faible alors que $\text{conf}_{BD}(R)$ est proche de 1, l'intérêt de la règle sera plus important, mettant ainsi en évidence une contradiction entre les données réelles et la modélisation de l'expert.

4 Application à la fouille de données d'interruptions opérationnelles

Dans le domaine aéronautique, une interruption opérationnelle est un retard au départ (décollage) de plus de quinze minutes, une annulation ou une interruption de vol suite à un problème technique (panne ou dysfonctionnement). Un tel événement est aujourd'hui considéré comme important par les compagnies aériennes pour le coût et les mécontentements engendrés.
De ce fait, lors du lancement de nouveaux projets avions, les ingénieurs doivent fournir dès la phase de conception une prédiction la plus réaliste possible de la fréquence des interruptions opérationnelles, qui sera mesurée lors de la future exploitation commerciale des avions. Ces prédictions initient, guident et valident les choix de conception. Pour effectuer cette activité, les ingénieurs utilisent un outil informatique implémentant un modèle mathématique stochastique intégrant les paramètres dont les impacts sur la fréquence des interruptions opérationnelles sont connus. Cet outil est calibré et paramétré par le retour d'expérience obtenu à partir d'avions, de systèmes ou d'équipements en service comparables.
Les besoins de recherche portent sur l'amélioration des modèles de calcul utilisés par cet outil de prédiction. Ainsi, la fouille des données en service est intéressante car elle permet de découvrir de nouveaux facteurs qui pourraient être intégrés à ces modèles pour améliorer la prédiction de la fréquence des interruptions opérationnelles. On se propose d'encadrer ce processus de découverte par l'approche méthodologique présentée dans la section 2. L'analyse des données doit permettre de valider les hypothèses qui ont été prises et d'enrichir le modèle de prédiction. Plus précisément, il s'agit d'aider l'expert à détecter, ou à vérifier, la présence de contributeurs de la fiabilité opérationnelle par la fouille des données en service.

4.1 Expérimentations

La base de données relative aux interruptions opérationnelles regroupe les détails de tous les problèmes techniques. Pour notre étude nous avons pris, en accord avec l'expert, un sous-ensemble de la base de données initiale. Après pré-traitement, on dispose de 23 attributs discrétisés et de plus de 12000 enregistrements décrivant les interruptions opérationnelles.
On se propose, dans un premier temps, de regarder les résultats issus de l'extraction des règles d'association δ-fortes. Ces règles sont présentées à l'expert en fonction de différentes mesures d'intérêt : confiance (Agrawal et al., 1993), J-mesure (Smyth et Goodman, 1992) et moindre contradiction (Azé, 2003). Cette première extraction comporte de nombreuses règles connues de l'expert, ce qui permet de se rendre compte de la nécessité d'expliciter certaines connaissances exactes (taxonomies) mais aussi des croyances fortes de l'expert sur son domaine.
On peut donc identifier plusieurs catégories de connaissances que l'on peut expliciter dans le réseau bayésien :

- *Taxonomie*, la présence d'une telle structure dans les données est intéressante car elle permet de capturer différents niveaux de détails dans les règles d'association. Cependant, l'existence de taxonomies introduit des dépendances exactes qui vont être capturées par un grand nombre de règles, occultant ainsi la lecture de règles potentiellement intéressantes.
- *Valeur d'attribut prépondérante*, lorsqu'un attribut de la base de données a une valeur très dominante (e.g. 95% des problèmes ont eu pour conséquence un retard) alors il est important de pouvoir intégrer cette information au modèle de connaissance afin d'éviter la production de règles triviales.
- *Croyance forte*, elle peut être représentée comme une dépendance entre un ou plusieurs attributs du réseau bayésien et par la définition des tables de probabilités jointes correspondantes. L'outil le plus connu et le plus facile à mettre en œuvre pour décrire les probabilités est l'échelle de probabilité (Druzdzel et van der Gaag, 2000).

Après intégration de ces connaissances dans un réseau bayésien, nous pouvons procéder au calcul de l'intérêt des règles d'association. Les résultats mettent alors en avant d'autres connaissances, plus pertinentes, pour le domaine d'application.

4.2 Résultats obtenus

Considérons d'abord l'extraction sans exploitation d'un réseau bayésien. L'extraction a donné 17760 règles δ-fortes. Le tableau 1 montre des exemples -choisis- de telles règles au moyen de l'algorithme décrit dans (Boulicaut et al., 2000) ($support_{min} = 100$, $\delta = 15$). Sur une configuration PC de bureau, l'extraction a demandé 2 minutes et 55 secondes.

Afin d'éclaircir la lecture de ces résultats, il peut être utile de faire quelques précisions. Les mots-clés `remove`, `ecam`, `mel`, etc. indiquent que l'analyse du texte libre rédigé par un technicien a permis de déceler une action particulière : « pose/dépose » d'un équipement, apparition de messages d'alertes, application d'une procédure spécifique. Lorsqu'un mot-clé est préfixé de `last=` cela signifie qu'il s'agit du dernier mot-clé, correspondant à une action, détecté dans la description du problème. Les nombres de 2, 4 ou 6 chiffres désignent les équipements incriminés dans l'interruption opérationnelle. Ces nombres obéissent à une taxonomie bien précise : la norme ATA 100. Ainsi l'équipement `286322` est un sous-équipement de `2863`,

Index	itemset δ-libre ⇒ fermeture	support	confiance
1	CS DY last=remove ⇒ remove	4046	1,00
2	2863 ⇒ SYSTEM 28 286322(-11) DY(-10)	245	0,96
3	TX last=nff ⇒ SYSTEM(-15) DY(-2) nff	247	0,94
4	15 ⇒ ENGINE 1511(-6) effect=DY(-8)	178	0,96
5	028886 ⇒ SYSTEM 02 0288 DY(-10)	234	0,96
6	ST2 last=none ⇒ OP2(-5) EngineXXA(-9) DY(-7)	238	0,96
7	4345 delay=0.5_1.5 ⇒ SYSTEM 43 434512(-3) DY	107	0,97
8	OP1 EngineXXA CS delay=0.5_1.5 remove ⇒ ST1(-12) DY last=remove(-14)	168	0,92
9	9911 TX delay=0.5_1.5 ⇒ SYSTEM 99 991112(-2) DY	155	0,98
10	8885 month=Aug ⇒ SYSTEM 88 DY(-4)	183	0,95

TAB. 1 – *Exemples choisis de règles extraites.*

etc. Ces équipements se regroupent en trois catégories représentées par les mots-clés SYSTEM, ENGINE et STRUCTURE. Les codes CS, TX, indiquent la phase de vol pendant laquelle le problème est survenu (phase de vérification au sol, décollage, etc.). Les codes DY ou CN représentent la nature de l'interruption provoqué (retard, annulation, etc.). Pour des raisons de confidentialité, les données sur les équipements ont été falsifiées, les sigles des compagnies et des aéroports, ainsi que les numéros de série des avions ont été volontairement rendus anonymes (ST fait référence aux aéroports et OP aux compagnies). Enfin, les chiffres négatifs entre parenthèses indiquent le nombre d'exceptions liées à un attribut de la partie droite de la règle.
Une analyse des résultats du tableau 1 permet de mettre en avant la présence d'associations correspondant :

- à la taxonomie des équipements (règles 2, 4, 5, 7, 9 et 10),
- à des relations triviales entre l'identification du dernier mot-clé et la présence de ce mot clé dans le texte (règles 1, 3, 8),
- à la prépondérance dans les données de certaines valeurs d'attributs comme DY ou CS.
- ou encore à des connaissances plus spécifiques du domaine, telles que les liens entre les compagnies et leur aéroport principal (règle 8), ou encore entre un mois de l'année et l'apparition d'incidents sur un équipement spécifique (e.g., système de conditionnement d'air pendant les périodes d'été, règle 10), etc.

Ainsi, beaucoup de règles contiennent des connaissances bien connues de l'expert. Nous montrons maintenant que l'approche proposée permet d'améliorer la découverte d'informations pertinentes grâce à l'exploitation d'un modèle des connaissances du domaine, de type réseau bayésien.

Taxonomie et associations évidentes L'intégration de ce type de connaissance au réseau bayésien est triviale. L'expert va ajouter un lien de causalité entre les attributs qui correspondent à cette information. Par exemple, on peut définir un lien entre ATA4d et ATA2d, puis un lien entre ATA2d et Category, etc. Les tables de probabilités sont ensuite définies de manière à exclure toute autre relation entre ces attributs, e.g., $P(\texttt{ATA2d = 43}|\texttt{ATA4d = 4345}) = 1.0$; et ainsi de suite pour tous les attributs obéissant à ce type de relation.

Itemsets prépondérants Il est important de définir la probabilité d'apparition des événements fréquents. En effet, la définition des distributions de probabilités des événements fréquents permet de limiter le facteur d'intérêt lié à la présence de ces attributs dans la partie

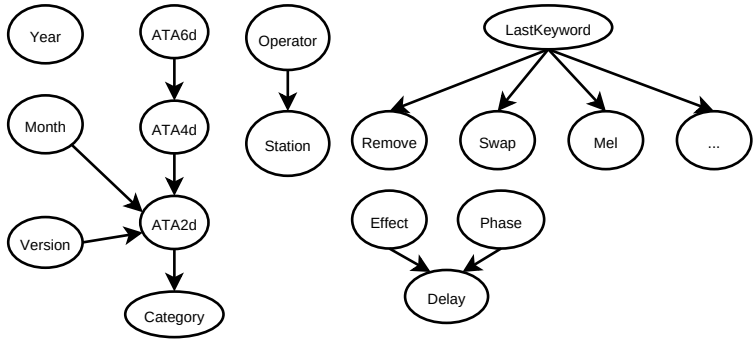

FIG. 2 – *Représentation partielle des connaissances de l'expert par réseau bayésien*

Index	itemset δ-libre $\Rightarrow$ fermeture	support	intérêt
1	CS DY last=remove $\rightarrow$ remove	4046	0,00
2	2863 $\Rightarrow$ SYSTEM 28 286322(-11) DY(-10)	245	0,11
3	TX last=nff $\rightarrow$ SYSTEM(-15) DY(-2) nff	247	0,13
4	15 $\Rightarrow$ ENGINE 1511(-6) effect=DY(-8)	178	0,01
5	028886 $\Rightarrow$ SYSTEM 02 0288 DY(-10)	234	0,01
6	ST2 last=none $\Rightarrow$ OP2(-5) EngineXXA(-9) DY(-7)	238	0,31
7	4345 delay=0.5_1.5 $\Rightarrow$ SYSTEM 43 434512(-3) DY	107	0,40
8	OP1 EngineXXA CS delay=0.5_1.5 remove $\Rightarrow$ ST1(-12) DY last=remove(-14)	168	0,16
9	9911 TX delay=0.5_1.5 $\Rightarrow$ SYSTEM 99 991112(-2) DY	155	0,09
10	8885 month=Aug $\Rightarrow$ SYSTEM 88 DY(-4)	183	0,05

TAB. 2 – *Utilisation du réseau bayésien pour mesurer l'intérêt des règles d'association*

droite d'une règle. Par exemple, on pourra définir $P(\texttt{Effect = DY}) = 0,97$ ou encore $P(\texttt{Remove = true}) = 0,95$. Ainsi la présence de l'attribut DY dans la partie droite d'une règle d'association n'influera pas sur le calcul de l'intérêt de cette règle.

Connaissances plus spécifiques L'expert peut vouloir définir des connaissances fortes du domaine, par exemple, le lien entre une compagnie et sa base principale, ou encore entre un mois de l'année et l'apparition d'un problème sur un équipement particulier. L'expert doit reformuler sa connaissance du domaine pour l'adapter à la modélisation du réseau bayésien : il doit créer des liens de causalité entre un ou plusieurs attributs du réseau puis définir les distributions de probabilités correspondantes. Pour cela, nous pouvons employer la méthode de l'échelle des probabilités pour expliciter, par exemple, qu'il est *probable* que la compagnie `OP1` soit associée à l'aéroport `ST1`, etc.
La figure 2 reprend les différentes informations que l'expert a pu extraire à partir des premiers résultats obtenus et montre une modélisation partielle de ces connaissances, après analyse et reformulation des motifs déjà extraits. Les tables de probabilités jointes ont elles aussi été définies par l'expert mais elles ne sont pas présentées pour des raisons de clarté. Ce réseau est ensuite utilisé pour calculer l'intérêt des règles d'association présentées dans le tableau 1. Le calcul montre (tableau 2) que les différentes règles ont un intérêt faible. Ce résultat est cohérent puisque le réseau bayésien a été défini de manière à pouvoir éliminer la plupart de ces motifs. Néanmoins les règles (6) et (7) ont un intérêt supérieur aux autres et nécessitent une analyse

plus poussée.
La règle (6) montre un lien entre un aéroport et l'absence d'action de maintenance (`last = none`) associés à une compagnie aérienne, ce qui pousse à s'intéresser sur le fonctionnement de cette compagnie. En effet, certaines compagnies préfèrent effectuer les opérations de maintenance « lourdes » dans leur base principale.
La règle (7) met en avant une relation entre une tranche de retard assez importante (`0.5_1.5`) et un équipement particulier (`434512`). Ce motif paraît particulièrement intéressant, mais une analyse plus poussée reste nécessaire pour valider ou non la réalité de cette association.

5 Conclusion

A partir des travaux de Boulicaut et al. (2000) et de Jaroszewicz et Simovici (2004), on a mis en place une méthodologie permettant de faciliter la découverte de règles d'association potentiellement intéressantes. Notre approche met en avant la collaboration entre réseaux bayésiens et collections de règles d'association (et donc ensemble fréquents). Cette méthodologie a été testée sur un cas d'application concret concernant la fouille des données d'interruptions opérationnelles dans l'aéronautique et elle montre des résultats encourageants.

Références

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. In P. Buneman et S. Jajodia (Eds.), *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., pp. 207–216.

Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, et A. I. Verkamo (1996). *Fast discovery of association rules*, Chapter 12, pp. 307–328. Menlo Park, CA, USA : American Association for Artificial Intelligence.

Azé, J. (2003). *Extraction de connaissances à partir de données numériques et textuelles*. thèse de doctorat, Université Paris-Sud.

Becquet, C., S. Blachon, B. Jeudy, J.-F. Boulicaut, et O. Gandrillon (2002). Strong association rule mining for large gene expression data analysis : a case study on human SAGE data. *Genome Biology 12*. See http ://genomebiology.com/2002/3/12/research/0067.

Boulicaut, J.-F., A. Bykowski, et C. Rigotti (2000). Approximation of frequency queries by means of free-sets. In *Proceedings of the 2000 PKDD European Conference on Principles and Practice of Knowledge Discovery in Databases*, pp. 75–85.

Boulicaut, J.-F., A. Bykowski, et C. Rigotti (2003). Free-sets : a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery journal 7*(1), 5–22.

Dechter, R. (1999). Bucket elimination : A unifying framework for reasoning. *Artificial Intelligence 113*(1-2), 41–85.

Druzdzel, M. J. et L. C. van der Gaag (2000). Building probabilistic networks : 'where do the numbers come from ?' guest editors' introduction. *IEEE Transactions on Knowledge and Data Engineering 12*(4), 481–486.

Goethals, B. et M. J. Zaki (2003). *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations FIMI 2003*. Melbourne, USA.

Heckerman, D. (1997). Bayesian networks for data mining. *Data Mining and Knowledge Discovery 1*(1), 79–119.

Jaroszewicz, S. et T. Scheffer (2005). Fast discovery of unexpected patterns in data, relative to a bayesian network. In *Proceedings of the 2005 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA. ACM Press.

Jaroszewicz, S. et D. A. Simovici (2004). Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 178–186. ACM Press.

Jeudy, B. (2002). *Optimisation de requêtes inductives : application à l'extraction sous contraintes de règles d'association*. Ph. D. thesis, Université Lyon I, INSA de Lyon.

Naïm, P., P.-H. Wuilleminn, P. Leray, O. Pourret, et A. Becker (2004). *Réseaux bayésiens*. Eyrolles.

Padmanabhan, B. et A. Tuzhilin (1998). A belief-driven method for discovering unexpected patterns. In *Proceedings of the 1998 KDD International Conference on Knowledge Discovery and Data Mining*, pp. 94–100.

Padmanabhan, B. et A. Tuzhilin (2000). Small is beautiful : discovering the minimal set of unexpected patterns. In *Proceedings of the 2000 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, pp. 54–63. ACM Press.

Pasquier, N. (2000). *Data mining : algorithmes d'extraction et de réduction des règles d'association dans les bases de données.* thèse de doctorat, Université Clermont-Ferrand II, LIMOS, Complexe scientifique des Céseaux, F-63177 Aubière cedex, France.

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24*(1), 25–46.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems : networks of plausible inference*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

Smyth, P. et R. M. Goodman (1992). An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering 4*(4), 301–316.

Summary

This paper deals with the implementation of a knowledge model in a data mining context. The proposed approach relies on coupling bayesian networks with a rule extraction technique, namely delta-strong association rule mining (minimal left member, minimal frequency and high confidence). The discovery of potentially useful rules is enhanced by the use of encoded expert knowledge. For validation purposes, our approach is illustrated with a use case dealing with operational interruptions in the aeronautic industry.