Chapitre 11 : Historique et fonctionnalités de CHIC

Raphaël Couturier* et Saddo Ag Almouloud**

* Laboratoire d'Informatique de l'université de Franche-Comté (LIFC),
IUT de Belfort-Montbéliard, BP 527, 90016 Belfort, France
raphael.couturier@iut-bm.univ-fcomte.fr
** Pontifícia Universidade Católica de São Paulo - PUC/SP
Rua Marquês de Paranaguá, 111, Consolação, São Paulo - SP - Brasil
saddoag@gmail.com
http://www.pucsp.br/pos/edmat

Résumé. CHIC permet d'utiliser la plupart des méthodes définies dans le cadre de l'ASI (Implication Statistique Implicative). Il a pour objectif de découvrir les implications les plus pertinentes entre les variables d'un ensemble de données. Pour cela, il propose d'organiser les implications sous forme d'une hiérarchie cohésitive (orientée) ou un graphe implicatif. De plus, il permet d'obtenir une hiérarchie des similarités (non orientée) basée sur les ressemblances des variables. Ce papier décrit l'historique, les caractéristiques et l'usage de CHIC.

1 Introduction

L'analyse statistique implicative (ASI) a été développée par Régis Gras et ses collaborateurs. Elle permet d'établir des règles d'association à partir d'un ensemble de données croisant sujets et variables. Le but initial de cette méthode a été de répondre à la question : "Si un objet possède une propriété, est ce qu'il en possède une autre ?". Bien entendu lorsque la réponse est totalement affirmative, il est facile de répondre à cette question. Cependant, il est possible que ce ne soit pas le cas, alors on peut constater que des tendances apparaissent. L'ASI a pour objectif de mettre en évidence de telles tendances dans un ensemble de propriétés. Comparée aux autres méthodes statistiques qui permettent de générer des règles d'association, l'ASI se distingue par le fait qu'elle utilise une mesure non linéaire qui satisfait des critères importants. Tout d'abord, cette mesure est basée sur l'intensité d'implication qui mesure le degré de surprise inhérent à une règle. Ainsi, les règles triviales qui sont potentiellement évidentes et connues de l'expert sont supprimées. Cette intensité d'implication peut être renforcée par le degré de validité, défini par rapport à l'entropie de Shannon, si l'utilisateur choisit ce mode de calcul. Dans ce cas, la mesure ne prend pas simplement en compte la validité de la règle, mais aussi sa contraposée. En effet, quand une règle d'association est estimée valide, c'est-à-dire que l'ensemble des items A est fortement proche de l'ensemble des items B, alors il est légitime et intuitif d'attendre que la contraposée soit valide, c'est-à-dire que l'ensemble des items non-B soit fortement

proche de l'ensemble des items non-A. Ces deux mesures originales sont complétées par une mesure classique basée sur la taille du support de la règle. Ainsi, en combinant les trois mesures, on peut définir une mesure pertinente qui possède les qualités des trois mesures (si on considère l'utilisation de la théorie entropique), c'est-à-dire la résistance au bruit comme la contraposée de la règle est prise en compte et le rejet des règles triviales. Pour plus d'information le lecteur intéressé peut consulter Gras et al. (2004). Grâce à cette mesure originale, CHIC permet de calculer les règles d'associations à partir d'un ensemble de données. CHIC et l'ASI ont été utilisés pour un large spectre de domaines de recherche (Couturier et al., 2004; Froissard, 2005; Couturier, 2005; Orus et Gregori, 2005; Ramstein, 2008).

CHIC permet de construire deux types de hiérarchie et un graphe. La hiérarchie la plus connue est la hiérarchie des similarités. L'index de similarité a été défini dans Lerman (1981) et il permet de construire une hiérarchie ascendante. De manière similaire, l'intensité d'implication peut être utilisée afin de construire une hiérarchique orientée. En plus de cela, CHIC offre la possibilité de générer un graphe original, appelé graphe implicatif qui permet à l'utilisateur de sélectionner les règles d'associations et les variables qu'il souhaite voir apparaître.

L'historique de CHIC est décrit dans la section 2. Dans la section 3 nous passons en revue les variables que peut traiter CHIC ainsi que les options qui peuvent aider l'utilisateur. La section 5 présente la hiérarchie des similarités et la hiérarchie cohésitive. La section 6 présente le graphe implicatif. Dans la section 7 nous présentons d'autres possibilités de CHIC. La section 8 donne une illustration du calcul avec variables intervalles et du calcul des typicalités et contributions. Finalement, la section 9 conclut ce chapitre.

2 Historique

2.1 Le cadre théorique

Dans ces paragraphes, nous retraçons la genèse historique et l'évolution du logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive) parallèlement aux développements théoriques de l'ASI. Dans sa thèse d'état, R. Gras (Gras, 1979), à partir de l'idée de l'indice de similarité développé par I. C. Lerman a construit l'indice d'implication statistique entre variables binaires. A partir de nouveaux problèmes réels posés pour analyser d'autres types de variables, R. Gras et son équipe ont élargi l'étude de l'implication en concevant à chaque étape les notions mathématiques en réponse à ces problèmes :

- Entre variables non binaires, c'est-à-dire modales ou fréquentielles ;
- Entre variables-sur-intervalles, variables-intervalles et variables floues ;
- Entre des classes de variables de nature quelconque.

Les applications à l'analyse de problèmes de didactique ont mis en évidence la nécessité d'autres développements théoriques. C'est ainsi que le concept "d'implication-inclusion", les notions de typicalité et de contribution des variables supplémentaires, de nœuds significatifs, ont été développés toujours en écho aux questions posées par les praticiens.

Mais l'émission de ces questions de terrain conduisant aux développements théoriques de l'ASI a été facilitée, rendue possible grâce à l'implémentation de l'outil informatique CHIC sur micro-ordinateur. Ses élaborations successives ont pris en compte le développement théorique, mais aussi les progrès des outils informatiques. C'est donc l'interaction ternaire questions posées - réponses théoriques - réalisations informatiques qui est le moteur du paradigme systémique que serait l'implication statistique.

2.2 Les premières étapes de CHIC

Avant 1990, le logiciel CHIC (il ne portait pas encore ce nom) consistait en une version primitive de R. Gras, en Basic (1984), implémentant les calculs des intensités d'implication et surtout, plus délicat, l'algorithme de construction sur micro-ordinateur (Thomson 05) de la hiérarchie des similarités de I.C.Lerman.

Dès 1990, puis dans le cadre de sa thèse S. Ag Almouloud (Ag Almouloud, 1992), s'est attaqué à l'un des objectifs de la réalisation d'un outil informatique fiable et assez convivial pour permettre de traiter, outre l'analyse de similarité de I. C. Lerman, l'analyse implicative de R. Gras et ses extensions : la hiérarchie implicative de classe (Larher, 1991) ainsi que l'étude des variables numériques et modales. Pour ce faire, il fallait créer un logiciel d'analyse de données intégrant les traitements suivants dont S. Ag Almouloud s'est chargé :

- la saisie et les éditions de données, ainsi que les différentes opérations sur ces données (100 sujets et 54 variables maximum),
- la classification hiérarchique de similarité entre variables (binaires, ou fréquentielle s) de I. C. Lerman,
- l'intensité d'implication entre variables (binaires, modales ou fréquentielles),
- la hiérarchie implicative de classes de variables et le calcul de leur cohésion,
- la construction de graphe implicatif (programmé par H. Rostam (Rostam, 1981) et repris en turbo Pascal par M. Mouradi, Université de FES, Maroc),
- le repérage des nœuds significatifs des la hiérarchie des similarités
- du repérage des nœuds significatifs de la hiérarchie implicative de classes (programme réalisé par H. Ratsimba-Rajohn (Ratsimba-Rajohn, 1992) dans le cadre de sa thèse),
- le calcul des paramètres (moyenne, écart-type et corrélation).
- la possibilité de concaténer deux fichiers ayant le même nombre de variables binaires.

2.3 Une évolution dans une seconde étape

L'utilisation de ce premier module d'édition de données conduisant à des restrictions préjudiciables, (impossibilité d'ajouter ou d'éliminer des variables ou des individus par exemple,) S. Ag Almouloud l'a progressivement amélioré :

- l'extension de la dimension du tableau de données (1000 individus et 100 variables) grâce à une meilleure gestion de la mémoire ;
- la possibilité de supprimer ou d'adjoindre des individus et des modalités ;

- la concaténation de deux fichiers ayant le même nombre d'individus ;
- la disjonction d'un sous-programme permettant de modifier la valeur d'une donnée ;
- la relecture de fichier de données ; ces données pouvant être des données binaires, fréquentielles ou des intensités ;
- une meilleure gestion de la mémoire et des ports graphiques
- le calcul des intensités d'implication.

Le changement principal que S. Ag Almouloud a alors apporté consiste dans la précision des résultats. En collaboration avec A. Totohasina (Totohasina, 1992), la loi normale est simulée. Aussi, remplace-t-on dans le programme initial de R. Gras (traduit en turbo pascal par P. Gentil) l'approximation grossière de cette loi par une simulation plus fine. Pour l'analyse implicative, sont intégrées au programme :

- la possibilité d'une ou plusieurs sorties des intensités d'implication à 1'écran. L'option offre à l'utilisateur la sélection des intensités d'implication les plus fortes afin de faciliter la construction manuelle du graphe implicatif,
- l'impression à l'écran ou à l'imprimante des couples (i, j), j fixe et i variable tels que ($i \Rightarrow j$). Cette option s'inscrit également dans le cadre d'une aide à la construction du graphe implicatif,
- la sauvegarde des intensités d'implication dans un fichier explicitement nommé. Leur stockage était indispensable au calcul des indices de cohésion et à la hiérarchie cohésitive de classes,
- la possibilité également de stocker la matrice d'incidence dans un fichier dénommé par l'utilisateur. Pour la construction automatique d'un graphe implicatif, les intensités d'implication sont multipliées par 100. Sa construction est essentielle pour le didacticien qui cherche à mettre en évidence de façon visuelle, les enchaînements dans les comportements de réponses des élèves.

Signalons qu'à chacune des options "calcul des intensités d'implication" et "calcul des indices de similarité", nous obtenons les occurrences des variables. L'analyste peut alors en éliminer certaines s'il estime que la faiblesse de leurs occurrences n'est pas suffisante ou pertinente et ceci sans changer le contenu du fichier initial. Cette élimination entraîne éventuellement une nouvelle numérotation des variables non éliminées et une sortie à l'écran du nombre de variables sur lesquelles portera l'analyse. Pour, d'une part, vérifier la fiabilité du nouvel outil statistique qu'est la hiérarchie implicative de classes, et d'autre part, informer les utilisateurs des potentialités de ce nouvel outil, S. Ag Almouloud a intégré au programme "cohésion", disponible dans la partie "hiérarchie implicative de classes", les indices d'implication entre les classes de variables.

En résumé, la dernière version en 1992 de CHIC écrite en turbo pascal 6, se compose d'un programme principal et d'un ensemble de sous-programmes, selon l'organigramme figure 1 correspondant à la structure de CHIC réalisée par S. Ag Almouloud.

Parallèlement aux travaux réalisés par S. Ag Almouloud, Harrisson Ratsimba-Rajohn (Ratsimba-Rajohn, 1992) a introduit dans sa thèse la notion de typicalité de variables supplémentaires et des individus et a réalisé le programme informatique qui en permet le traitement.

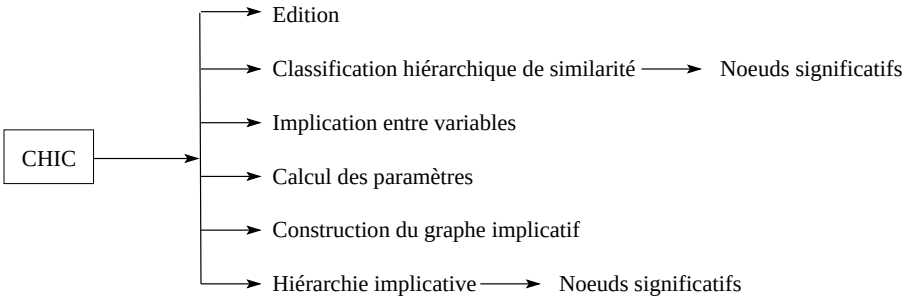

FIG. 1 – *Organigramme du logiciel CHIC*

2.4 Le CHIC contemporain

Depuis 1993, Raphaël Couturier a repris la réalisation informatique du logiciel CHIC en lui apportant des améliorations et une évolution prenant en compte les nouveaux développements théoriques de l'ASI et les vœux des utilisateurs. Cette évolution se traduit, entre autres, par l'introduction des couleurs, par l'enrichissement du menu et par son écriture en C++ qui a conduit à des dizaines de versions différentes jusqu'à l'actuelle version 4.2. Celle-ci permet comme nous allons le voir plus en détail :

- de traiter différents de variables, autres que binaires,
- de quantifier la significativité des valeurs attribuées à la qualité, la consistance de la règle associée, de classes ordonnées de règles, à la typicalité et la contribution des sujets ou de catégories de sujets à certaines règles,
- de représenter, par un graphe, pour un seuil de qualité choisi, des chemins de règles et, par une hiérarchie, des règles sur des règles que l'on appelle aussi règles généralisées,
- de supprimer, d'ajouter, de conjoindre des variables.

3 Variables

Initialement CHIC tout comme l'ASI ont été pensés pour traiter des variables binaires. Par la suite, l'ASI a été enrichie par l'ajout d'autres types de variables et CHIC en a bénéficié. Actuellement, CHIC offre la possibilité de traiter des variables binaires, des variables fréquentielles, des variables définies sur intervalles et des variables intervalles. Le cas des variables binaires est évidemment le cas le plus simple. Les variables fréquentielles, quant à elles, prennent leur valeur entre 0 et 1. Ce type de variable permet de modéliser les variables modales pour lesquelles il existe un nombre fixe de valeurs comprises entre 0 et 1 qui correspondent aux différentes modalités. La manière de définir les modalités est très importante, parce qu'elle intervient fortement dans les résultats de CHIC selon que les valeurs des modalités ordonnées sont proches de 0 ou de 1. Cette remarque est évidemment valide pour les variables fréquentielles.

L'utilisateur doit prêter une très grande attention au processus qu'il utilise pour transformer une variable réelle en une variable fréquentielle. En effet, plusieurs stratégies sont envisageables en fonction des valeurs. Si les valeurs sont positives, elles peuvent être divisées par la valeur maximale. Une autre possibilité consiste à considérer que la valeur minimale représente le 0 et que la valeur maximale représente le 1, toutes les autres variables sont, dans ce cas, proportionnellement distribuées entre la valeur minimale et la valeur maximale. Si une variable réelle possède des valeurs négatives et positives, il est possible de constituer deux variables, l'une contenant les valeurs positives et l'autre contenant les valeurs négatives. Dans ce cas, les précédentes remarques sont toujours opportunes pour les deux variables nouvellement constituées. Cependant, il est également possible de considérer que la valeur minimale (même si elle est négative) représente 0 et la valeur maximale représente 1 et, ainsi, les autres valeurs sont proportionnellement converties dans l'intervalle $[0, 1]$.

	poids p	taille p	femme s	homme s
i1	95	200	0	1
i2	43	140	1	0
i3	75	186	0	1
i4	60	174	1	0
i5	110	183	0	1
i6	140	180	0	1
i7	100	176	0	1
i8	79	172	1	0
i9	69	170	1	0
i10	45	165	1	0
i11	65	180	0	1
i12	60	175	0	1
i13	120	175	0	1
i14	80	160	1	0
i15	100	180	0	1
i16	121	175	0	1
i17	97	162	1	0
i18	72	176	0	1
i19	40	160	1	0
i20	57	175	1	0

FIG. 2 – *Un simple exemple de données avec des variables sur intervalles (avant partitionnement) et des variables supplémentaires*

Les variables sur intervalles et les variables intervalles sont utilisées pour modéliser des situations complexes. Nous les détaillons par la suite. Les variables intervalles permettent de faire face au problème rencontré par la conversion d'une variable réelle en une variable fréquentielle, comme nous l'avons expliqué précédemment (voir paragraphe suivant). En utilisant les mêmes valeurs réelles, une variable sur intervalle procède différemment. Elle découpe les valeurs de la variable en un nombre fixe d'intervalles. Le nombre d'intervalles est choisi par l'utilisateur et ensuite l'algorithme des

nuées dynamiques Diday (1971) constitue automatiquement les intervalles qui ont des limites distinctes. Cet algorithme a la particularité de construire des intervalles en minimisant l'inertie de chaque intervalle et en maximisant l'inertie interclasse de l'ensemble des intervalles. Ensuite, un intervalle est représenté par une variable binaire et un individu a la valeur 1 s'il appartient à cet intervalle et 0 sinon. En utilisant une telle décomposition, un individu appartient à un seul intervalle. Ainsi, le nombre de variables croit avec cette méthode.

Prenons un exemple. Supposons que nous disposions d'un ensemble d'individus et que pour chacun d'entre eux, nous connaissions sa taille et son poids. Supposons également que ces individus pèsent entre $40kg$ et $140kg$ et que leur taille varie entre $140cm$ et $200cm$. La figure 2 montre un exemple avec quelques individus, les valeurs ont été choisies arbitrairement et ne sont pas encore partitionnées. Supposons que nous souhaitions décomposer chaque variable en quatre intervalles, en fonction de la distribution des deux variables, nous pouvons obtenir les intervalles $[40, 60[$, $[60, 95[$, $[95, 110[$, $[110, 140]$ que nous appelons respectivement $poids1$, $poids2$, $poids3$ et $poids4$ et les intervalles $[140, 160[$, $[165, 174[$, $[174, 186[$, $[186, 200]$ pour les tailles que nous appelons respectivement $taille1$, $taille2$, $taille3$ et $taille4$. Dans la suite du calcul, toutes les unions des intervalles d'une variable sont considérées. Ainsi avec la variable $taille$, nous obtenons les intervalles $taille12$, $taille23$, $taille34$, $taille1-3$, $taille2-4$ et $taille1-4$. Les intervalles de la forme $nomAB$ correspondent à l'union de deux intervalles consécutifs, par exemple $taille23$ correspond à l'union des intervalles $taille2$ et $taille3$. Les intervalles de la forme $nomA-B$ correspondent à l'union de tous les intervalles entre $nomA$ et $nomB$, par exemple $taille1-3$ correspond à l'union des intervalles $taille1$, $taille2$ et $taille3$. Bien évidemment, l'utilisation des variables sur intervalles est d'autant plus intéressante quand il est possible de constituer des partitions les plus petites possibles, c'est-à-dire rassembler les intervalles pour lesquels on sait qu'ils sont naturellement proches les uns des autres. CHIC permet d'utiliser un tel algorithme qui est décrit mathématiquement, par exemple, dans (Gras et al., 1996; Gras, 2005). Avec l'exemple précédent, si d'autres variables renseignent sur les tendances vers telle ou telle caractéristique des individus, alors il est possible d'obtenir des informations entre ces variables et le poids et la taille de la population étudiée. Par exemple, il est possible de savoir que les personnes mesurant entre $140cm$ et $180cm$ ont plutôt une attirance pour telle ou telle chose ou que les personnes avec telles hauteurs pèsent principalement entre $90kg$ et $150kg$. Bien évidemment le nombre d'intervalles peut agir fortement sur les résultats.

Alors que pour une variable sur intervalles un individu prend la valeur 1 pour un seul intervalle, une variable intervalle offre la particularité qu'un individu ait différentes valeurs sur plusieurs intervalles. De plus, les intervalles peuvent être continus et peuvent représenter une décomposition discrète, comme c'est le cas en utilisant une méthode de décomposition automatique telle que celle des nuées dynamiques, mais ils peuvent également être définis par l'utilisateur selon les critères personnels de celui-ci. En prenant l'exemple précédent avec la taille et le poids, un utilisateur peut préférer choisir que les personnes soient élancées, normales, en surcharge pondérale, petite, moyenne ou grande. Néanmoins, une variable intervalle offre la possibilité qu'un individu puisse prendre différentes valeurs parmi les différents intervalles

mais impose que la somme de ces valeurs soit inférieure ou égale à 1. Dans la plupart des cas, la somme sera égale à 1 mais ce n'est pas une obligation. Dans la pratique, l'utilisation de variables intervalles permet de classifier plus facilement un objet ou un individu parce qu'il est fréquent que les opinions divergent sur le fait que quelque chose ou quelqu'un soit plutôt petit ou normal. Par conséquent, on peut exprimer que quelqu'un est mince en donnant à cet individu 0.75 pour élancé et 0.25 pour normal. Il faut également noter que ce mécanisme permet de prendre en compte les variables floues qui sont souvent utilisées dans certains types de problèmes Bojadziev et Bojadziev (1996). L'utilisation de variables floues vient soit d'une appréciation humaine, qui par définition est subjective, soit par une mesure imprécise qui pour une raison quelconque introduit une incertitude (cf Partie1, chap 7).

CHIC utilise le format CSV (avec un point virgule comme séparateur) comme format de données pour les fichiers, celui-ci est utilisé classiquement dans les tableurs. Les individus sont rangés dans la première colonne. Les variables sont disposées sur la première ligne. Les valeurs des individus sont représentées dans un tableau à deux dimensions tel que les valeurs pour chaque variable d'un individu sont rangées dans une ligne du tableau (le premier élément étant le nom de l'individu). Les valeurs d'une variable pour tous les individus sont disposées dans les colonnes du tableau (le premier élément étant le nom de la variable). Bien entendu, le type des valeurs diffère dans un tableau selon le type des variables (binaires, fréquentielles, ...).

Les variables supplémentaires peuvent être utilisées dans CHIC afin d'expliquer la formation de certaines règles. Ce type de variable n'intervient pas directement dans le calcul des règles mais il est utilisé dans le calcul des typicalités et des contributions. Prenons un exemple. Supposons que nous souhaitions étudier l'impact d'un nouveau tramway dans une ville et qu'un questionnaire ait été élaboré à cet effet. Ce dernier rassemble de nombreuses informations sur les besoins et les espoirs associés à ce projet. Dans ce genre de questionnaire, le sexe des personnes est renseigné. Par exemple, il est possible que CHIC, avec un tel questionnaire, génère des règles telles que les personnes travaillant et habitant loin de leur lieu de travail sont généralement très intéressées par le projet, ou les familles avec des enfants jeunes sont parties prenantes du projet. En utilisant le sexe des personnes comme variable supplémentaire, il est possible de savoir si les personnes responsables de la construction des précédentes règles sont plutôt des hommes, des femmes ou s'il n'y a pas de distinction.

Avant de lancer un calcul, l'utilisateur doit choisir quel type de calcul il désire utiliser. En effet, il est possible de choisir le calcul classique de l'intensité d'implication ou la version entropique de celle-ci, comme nous l'avons signalé en introduction. Ce choix de type de calcul influe très fortement sur les règles produites. Généralement, il faut utiliser la version entropique dès lors que l'effectif de la population devient grand, par exemple plus grand que 300 à 400. Elle est plus sévère que la version classique de l'intensité d'implication qui produit plus de règles mais qui n'est pas appropriée aux grands ensembles de données.

Pour plus d'informations sur l'intensité d'implication, le lecteur intéressé est invité à consulter la partie 1 de ce livre.

4 Calcul des conjonctions

Pour calculer efficacement les indices d'implication CHIC est basé sur l'algorithme de calcul des règles d'association défini par Agrawal Agrawal et al. (1993). Cet algorithme permet de calculer efficacement règles d'implications composées de conjonctions dans la partie prémisse. Prenons un exemple avec 5 variables A, B, C, D et E et cherchons les règles composées de 3 variables (donc de la forme $A \wedge B \Rightarrow C$). Pour cela, l'algorithme va déterminer les occurrences des triplets de variables, c'est-à-dire des 8 triplets : ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE and CDE. Pour chacun de ces triplets, il a fallu déterminer les occurrences des couples AB, AC, AD, AE, BC, BD, BE, CD, CE et DE. À partir de ces couples et triplets, il est possible de calculer de nombreuses règles d'implication. Par exemple avec les occurrences de ABC, de AB, BC, et AC il est possible de calculer l'intensité des règles $A \wedge B \Rightarrow C$, $B \wedge C \Rightarrow A$ et $A \wedge C \Rightarrow B$.

Même avec un seuil d'implication élevé, le nombre de règles produit par les conjonctions peut s'avérer très élevé si le nombre de variables initial est grand. De plus, le fait d'utiliser des conjonctions peut être source de ressemblance voire de superfluité (cf Partie 1, chap 9) entre les règles, c'est pourquoi nous avons introduit un critère d'originalité entre les règles. Celui-ci permet de sélectionner uniquement les conjonctions de règles présentant un critère d'originalité que nous définissons par le fait que les sous-règles qui la composent ne sont pas triviales. Prenons par exemple la règle $A \wedge B \Rightarrow C$, elle est originale si son intensité d'implication est forte et si les règles $A \Rightarrow C$ ont $B \Rightarrow C$ une faible intensité d'implication. Les détails des calculs sont dans Couturier (2008).

Dans la suite nous présentons les modes de représentations graphiques offerts par CHIC. L'utilisateur souhaitant plus d'informations sur la manière de calculer les règles pourra consulter par exemple Gras et al. (2004) et les références s'y trouvant.

5 Hiérarchie des similarités et hiérarchique cohésitive

Dès que CHIC a calculé l'ensemble de toutes les règles en fonction des paramètres choisis par l'utilisateur, il est possible de construire une hiérarchie à partir de ces règles. Cette hiérarchie peut s'apparenter à une méthode de classification orientée ou non en fonction du type de calcul choisi "similarité ou implication". Cependant les manières de construire chacune de ces hiérarchies comportent certaines similitudes. Dans la suite une règle est appelée classe, elle agrège deux variables dans sa forme la plus simple. À chaque niveau de la classification, CHIC choisit la classe qui possède la plus grande cohésion (de similarité ou d'implication). Ensuite, à chaque étape, CHIC calcule un ensemble de nouvelles classes à partir des classes présentes dans la hiérarchie. Pour créer une nouvelle classe, on agrège une classe existante avec soit une variable qui n'a pas été agrégée pour l'instant, soit avec une autre classe de la hiérarchie. Néanmoins, chaque couple de variables lors de l'agrégation de deux classes doit avoir une intensité valide. Par exemple, la formation de la classe $((a,b),c)$ nécessite que les classes (a,c) et (b,c) aient une bonne cohésion (avec l'implication) ou soient similaires (avec l'analyse des similarités). La classe $((a,b),c)$ représente la règle

$(a \Rightarrow b) \Rightarrow c$ avec l'analyse implicative et représente le fait que a implique b et que la classe ((a,b),c) admet une bonne cohésion et que la classe (a, b) est similaire à c avec l'analyse des similarités. Pour plus de détails sur la formation de classes, nous invitons le lecteur intéressé à se référer à (Lerman, 1981; Gras et al., 1996).

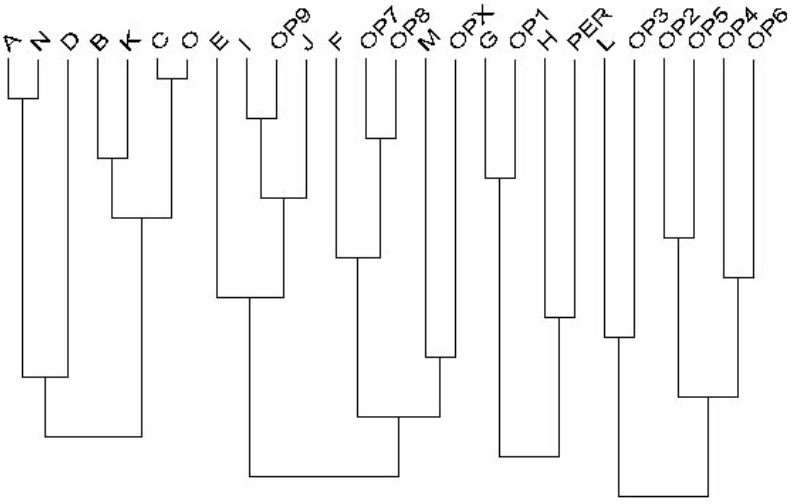

FIG. 3 – *Un exemple de hiérarchie des similarités*

Si l'utilisateur se demande quelle allure aurait la hiérarchie sans une ou plusieurs variables, il peut simplement les déselectionner grâce à l'interface prévue à cet effet. Cette possibilité est offerte pour toutes les représentations de CHIC (hiérarchie ou graphe). Malheureusement, une modification (même petite) portant sur la présence des variables implique une reconstruction totale de la hiérarchie. Cette étape dépend fortement du nombre de variables concernées dans le calcul (l'algorithme a une complexité qui dépend de la factorielle du nombre de variables dans le pire des cas). Avant de lancer une analyse, l'utilisateur peut choisir dans les options de calcul de détecter les niveaux significatifs de la hiérarchie.

La figure 3 montre une hiérarchie des similarités et la figure 4 illustre une hiérarchie cohésitive. Pour cette dernière, les niveaux significatifs sont affichés. Ils sont représentés par un trait rouge (dans CHIC) et ils signifient que le niveau signalé est plus significatif que le précédent et le suivant. Pour plus de détails sur leur construction sur les articles s'y rapportant, consulter par exemple Gras et al. (1996) et dans cet ouvrage Partie 1, chap 4.

L'indice de similarité est défini en utilisant soit la théorie classique soit la théorie entropique. Il est clair que le dernier choix est préférable avec un grand nombre d'individus. De plus, la construction de la hiérarchie des similarités avec la théorie classique conduit à la fin à une seule classe qui rassemble toutes les autres. Au contraire,

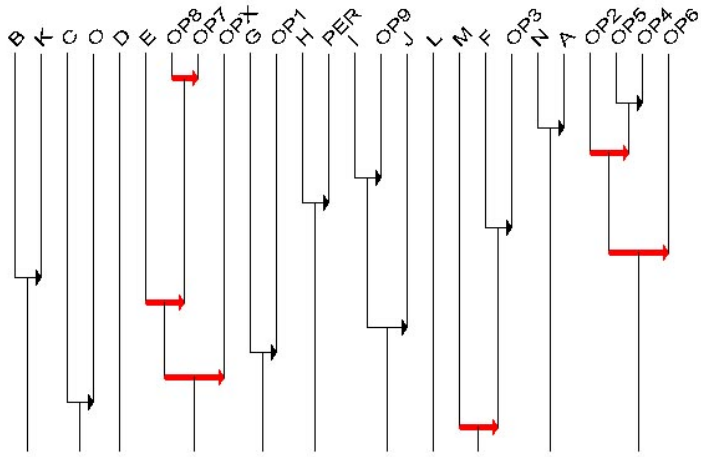

FIG. 4 – *Un exemple de hiérarchique cohésitive*

avec la version entropique de l'index de similarité, il est fréquent que l'algorithme de construction des classes conduise à plusieurs classes distinctes à la fin du processus. Le nombre de classes dépend en fait de la similarité des données.

Pour plus d'informations sur la hiérarchie cohésitive, le lecteur intéressé est inviter à consulter le chapitre 4 de la partie 1 de ce livre.

6 Graphe implicatif

Comme nous l'avons expliqué précédemment, les deux classifications de CHIC mettent en valeur certaines règles significatives en mettant de côté certaines autres règles. Si aucune variable ne doit être privilégiée, l'utilisation du graphe implicatif peut sembler judicieuse. Dans ce cas, l'utilisateur peut visualiser les règles dont l'intensité est plus grande qu'un seuil choisi. Dans la pratique, quatre seuils sont disponibles et CHIC propose des couleurs différentes les identifier plus rapidement. La figure 5 illustre un exemple de représentation d'un graphe implicatif. Une flèche est utilisée pour représenter l'implication entre deux variables (la règle $A \Rightarrow B$ est représentée par une flèche entre A et B). Comme le nombre de règles peut être important, l'utilisateur a la possibilité de sélectionner uniquement certaines variables. Ainsi seules les règles impliquant les variables présentes sont représentées. Par conséquent, ceci réduit le nombre de règles. De plus, afin de rendre le graphe plus lisible, CHIC utilise un algorithme de dessin automatique de graphes qui essaie de minimiser le nombre de croisements parmi les règles. Par défaut les fermetures transitives ne sont pas affichées sur le graphe implicatif. Un simple clique sur la souris dans la boite à outils les affiche. CHIC les calcule une fois pour toutes au début de chaque nouveau

graphe. Ensuite, même si l'utilisateur sélectionne ou désélectionne certaines variables, change le seuil d'affichage des règles, choisit d'afficher ou non les fermetures transitives, CHIC affiche le graphe sans aucun calcul supplémentaire. Cela permet à l'utilisateur de mettre en évidence les caractéristiques importantes de ses données. Néanmoins, l'utilisation de la procédure de dessin automatique de graphe est coûteuse en ressource de calculs, c'est pourquoi il n'est pas souhaitable de l'utiliser systématiquement.

Pour plus d'information sur la hiérarchie cohésitive, le lecteur intéressé pourra trouver de plus amples information dans la partie 1 de ce livre.

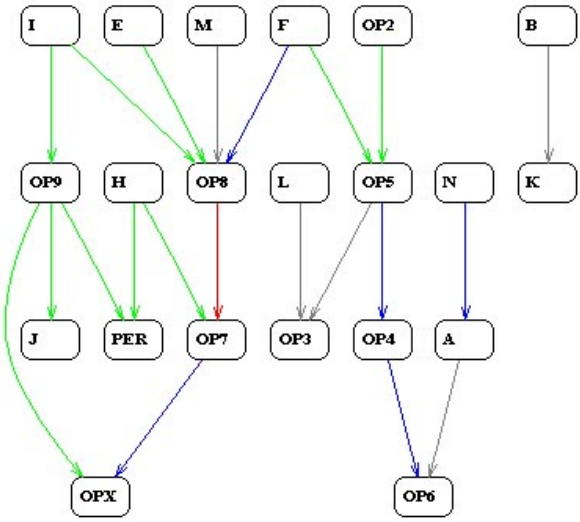

FIG. 5 – *Une exemple de graphe implicatif*

7 Autres possibilités

En plus des modes de représentation précédemment décrits, CHIC fournit quelques outils intéressants. Pour chaque mode de représentation graphique, il est possible de calculer la contribution et la typicalité d'un individu à une règle donnée. De la même manière, CHIC permet de calculer la contribution et la typicalité d'un ensemble d'individus à une règle donnée (cf Partie 1, chap 5).

La notion de contribution est définie pour déterminer les individus qui contribuent bien à la création de la règle. Ces individus sont plus responsables de la création de la règle que les autres. Par exemple si une règle $A \Rightarrow B$ possède une intensité implicative égale à $0,7$, alors les individus les plus contributifs sont ceux qui ont la valeur 1 pour les variables A et B. Par opposition, la notion de typicalité est définie par le fait que

certains individus soient "typiques" du comportement de la population, c'est-à-dire avec une intensité d'implication similaire à celle de la règle. Avec l'exemple précédent, les individus les plus "typiques" de la règle sont ceux qui possèdent respectivement des valeurs proches de $0,5$ et 1 pour A et B (ces valeurs dépendent du mode de calcul choisi pour l'analyse, classique ou entropique, et des cardinalités des ensembles A et B).

Il est facile de se rendre compte que les notions de typicalité et de contribution sont différentes. De la même manière, la notion de typicalité (respectivement de contribution) d'un ensemble d'individus (ou d'une catégorie d'individus) est définie pour savoir si un ensemble d'individus particulier est typique (respectivement contributif) d'une règle, d'une règle généralisée ou d'un chemin.

8 Illustration avec les variables intervalles et le calcul des typicalités et des contributions

Cette partie décrit un exemple simple et concret avec les deux variables sur intervalles de la section 3. La figure 6 montre un graphe implicatif issu des données de la figure 2. Les deux variables *poids* et *taille* sont automatiquement découpées en 4 intervalles par CHIC en suivant la méthode décrite dans la section 3. Dans ce graphe on peut remarquer quelques propriétés intéressantes. Par exemple, on peut remarquer les règles : $poids1 \Rightarrow taille12$ et $taille34 \Rightarrow poids2\text{-}4$. En raison du petit nombre d'individus pour cet exemple, par conséquent non significatif, et parce que les valeurs ont été générées de manière arbitraire, on ne peut rien dire de plus que : "les individus légers sont généralement petits et les personnes les plus grandes ne sont pas les plus légères". Néanmoins ces règles montrent une implication entre les partitions de deux variables. En considérant que les données puissent avoir du sens pour un expert, alors nous pourrions calculer la typicalité et la contribution d'un groupe d'individus. Par exemple, à propos de la règle $taille34 \Rightarrow poids2\text{-}4$, CHIC détermine que la variable *homme* contribue le plus à cette variable. Au contraire, la variable la plus typique à la règle $poids1 \Rightarrow taille12$ est la variable $femme$. Ces deux résultats ne sont pas surprenants compte tenu des données.

9 Conclusion

CHIC permet de mettre en pratique la plupart des méthodes et techniques liées à l'ASI. Dans ce chapitre, nous décrivons les caractéristiques principales de CHIC. Tout d'abord nous détaillons les types de variables que CHIC permet de traiter. Quelques options utiles pour comprendre CHIC sont détaillées. Ensuite, nous présentons les trois principaux modes de représentation. La hiérarchie des similarités et la hiérarchie cohésitive fournissent respectivement une classification orientée et une classification non orientée. Le graphe implicatif, qui est de loin le plus interactif, permet à l'utilisateur de "fouiller" parmi ses données et ainsi mettre en évidence les règles qui peuvent intéresser l'expert.

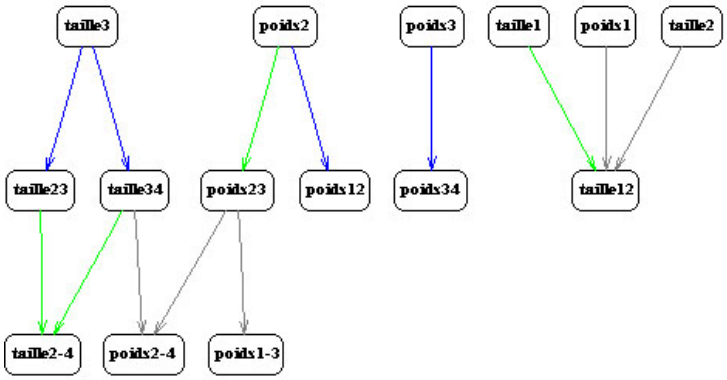

FIG. 6 – *Un exemple de graphe implicatif avec des variables intervalles*

Références

Ag Almouloud, S. (1992). *L'ordinateur, outil d'aide à l'apprentissage de la démonstration et de traitement de données didactiques*. Thèse de doctorat, Université de Rennes 1.

Agrawal, R., T. Imielinski, et A. Swami (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pp. 207–216.

Bojadziev, G. et M. Bojadziev (1996). *Fuzzy sets, fuzzy logic, applications*. World scientific.

Couturier, R. (2005). Un système de recommandation basé sur l'a.s.i. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 157–162.

Couturier, R. (2008). Statistical implicative analysis. In *CHIC : Cohesive Hierarchical Implicative Classification*, Volume 127 of *Studies in Computational Intelligence*, pp. 41–52. Springer Verlag.

Couturier, R., R. Gras, et F. Guillet (2004). Reducing the number of variables using implicative analysis. In *International Federation of Classification Societies, IFCS 2004*, pp. 277–285. Springer Verlag : Classification, Clustering, and Data Mining Applications.

Diday, E. (1971). La méthode des nuées dynamiques. *Revue de statistique appliquée 19*(2), 19–34.

Froissard, G. (2005). Chic et les études docimologiques. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 187–197.

Gras, R. (1979). *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques,*. Thèse d'état, Université de Rennes I.

Gras, R. (2005). Panorama du développement de l'A.S.I. à travers des situations fon-

datrices. In *Actes de la 3ème Rencontre Internationale A.S.I.*, pp. 9–33. Université de Palerme.

Gras, R., S. Ag Almouloud, M. Bailleul, A. Lahrer, M. Polo, H. Ratsimba-Rajohn, et A. Totohasina (1996). *L'implication Statistique*. La Pensée Sauvage.

Gras, R., R. Couturier, J. Blanchard, H. Briand, P. Kuntz, et P. Peter (2004). *Mesures de qualité pour la fouille de données*, Chapter Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, pp. 3–32. RNTI-E-1, Cepaduès Editions.

Larher, A. (1991). *Implication statistique et applications à l'analyse de démarches de preuve mathématique*. Thèse de doctorat, Université de Rennes I.

Lerman, I. C. (1981). *Classification et analyse ordinale des données*. Dunod.

Orus, P. et P. Gregori (2005). Des variables supplémentaires et des élèves "fictifs", dans la fouille didactique de données avec chic. In *Troisième rencontre internationale de l'Analyse Statistique Implicative (ASI3)*, pp. 279–291.

Ramstein, G. (2008). Statistical implicative analysis. In *Statistical Implicative Analysis of DNA microarrays*, Volume 127 of *Studies in Computational Intelligence*, pp. 205–225. Springer Verlag.

Ratsimba-Rajohn, H. (1992). *Contribution à l'étude de la hiérarchie implicative, application à l'analyse de la gestion didactique des phénomènes d'ostension et de contradictions*. Thèse de doctorat, Université de Rennes 1.

Rostam, H. (1981). Construction automatique et évaluation d'un graphe d'implication. Technical Report 150, IRISA, Rennes.

Totohasina, A. (1992). *Méthode implicative en analyse de données et application à l'analyse de conceptions d'étudiants sur la notion de probabilité conditionnelle*. Thèse de doctorat, Université de Rennes 1.

Summary

CHIC is a data analysis tool based on SIA. Its aim is to discover the more relevant implications between states of different variables. It proposes two different ways to organize these implications into systems: i) In the form of an oriented hierarchical tree and ii) as an implication graph. Besides, it also produces a (non oriented) similarity tree based on the likelihood of the links between states. The paper describes its history, its main features and its usage.