CHIC : traitement de données avec l'analyse implicative

Raphaël Couturier*, Régis Gras**

*LIFC - IUT Belfort
raphael.couturier@iut-bm.univ-fcomte.fr
**École polytechnique de l'université de Nantes
regisgra@club-internet.fr

Résumé. Cet article a pour but de montrer les possibilités offertes par le logiciel CHIC (Classification Hiérarchique Implicative et Cohésitive) pour effectuer certaines analyses de données. Il est basé sur la théorie de l'Analyse Statistique Implicative ou A.S.I. développée par Régis Gras et ses collaborateurs. Le principe premier de l'A.S.I. repose sur la problématique d'une mesure des règles d'association du type : «si a alors b» dans une population instanciant les variables a et b. CHIC enrichit sa réponse, établie sur des bases statistiques, en évaluant la responsabilité des sujets dans l'élection de la règle. L'article présent explique la démarche à suivre pour utiliser le logiciel ainsi que les possibilités offertes par celui-ci.

1 Introduction

CHIC est le fruit informatique des travaux sur l'analyse statistique implicative. La version actuelle de ce logiciel a été portée en C++ sous Windows il y a 10 ans environ à partir d'une version antérieure en Pascal, mais avec des développements importants et avec une plus grande convivialité Couturier (2000). Depuis elle a subi régulièrement de nombreuses modifications tant au niveau pratique que sur le plan théorique en intégrant de nombreux nouveaux modes de calculs et de nouveaux concepts. Dans cet article nous faisons le point sur les possibilités actuelles de CHIC pour l'usager des méthodes d'analyse de données. La théorie de l'analyse implicative est présentée dans Gras et al. (1996, 2001b, 2004). La documentation de CHIC peut apporter d'autres éclaircissements sur les concepts introduits dans cette présentation de CHIC. Rappelons-en brièvement les grandes lignes.

Partie d'une question de didactique (organiser les comportements de réponse d'élèves à des exercices de mathématiques et en valider une taxonomie de complexité), la problématique de l'A.S.I. s'est développée au fil des problèmes rencontrés dans d'autres situations où la recherche de règles –relations non symétriques, i.e. directionnelles– se pose et est résumée par l'énoncé : «si l'on observe a dans la population alors on observe généralement b». La fonction première de la règle est donc d'être un prédicteur potentiel. L'extension de la problématique initiale s'est développée dans plusieurs directions :

- la nature des variables binaire, puis numérique, ordinale, intervalle Gras et al. (2001a), floue ;
- la mesure affectée à la règle : classique sur une base adaptée de l'algorithme de la vraisemblance du lien de Lerman (1981), puis entropique afin d'atténuer l'effet

de taille des échantillons ;
- la mise en évidence de la contribution des sujets ou des descripteurs et de leur typicalité par rapport aux règles et aux familles de règles (chemin de graphe, classe de hiérarchie orientée).

Ces extensions s'inscrivent généralement comme réponses à des applications dont les champs se sont élargis en psychologie, sociologie, biologie, médecine, économie, autant de disciplines ouvertes à l'intelligence artificielle. Nous avons également défini une méthode pour réduire le nombre de variables basée sur l'A.S.I. Couturier et al. (2004).

Dans la partie 2, nous présentons tout d'abord comment formater les données afin de préparer un traitement. La partie 3 montre un arbre des similarités. Dans la partie 4, nous détaillons un exemple d'utilisation du graphe implicatif qui constitue sans doute la richesse et l'une des originalités de CHIC. Dans la partie 5, nous présentons une hiérarchie cohésitive. La partie 6 détaille toutes les possibilités communes aux différents types d'analyse pour affiner les résultats et leurs interprétations. Finalement la partie 7 donne une conclusion et des perspectives.

2 Mise en forme des données

Les données sont disposées sous forme d'un tableau numérique, dans lequel à chaque variable que nous souhaitons évaluer, nous faisons correspondre le résultat de l'évaluation (≥ 0 dans notre cas) de chaque objet ou individu à cette variable. Les variables à étudier peuvent avoir différents types, à savoir : binaire, modale et fréquentielle, quantitative ou intervalle. De plus elles peuvent être principales, c'est-à-dire qu'elles interviennent directement dans tous les calculs ou elles peuvent être supplémentaires comme il est fait en analyse factorielle. Les variables modales et fréquentielles doivent avoir une valeur réelle comprise entre 0 et 1. Les valeurs des variables quantitatives sont normalisées dans l'intervalle [0-1] en divisant toutes les valeurs par la valeur maximum obtenue par la variable. Il faut effectuer cette manipulation dans un tableur pour le moment. Les variables-intervalles sont automatiquement découpées en différents intervalles par un algorithme approprié, de type « nuées dynamiques »[1], qui, à partir d'un nombre d'intervalles choisi par l'utilisateur, constitue des intervalles tout en maximisant la variance inter-classe. Ayant formaté les données, il faut sauvegarder le fichier avec le type CSV qui est un format standard, chaque champ étant séparé par un point virgule.

Après le formatage des données, nous pouvons lancer les différents traitements proposés par CHIC. Il s'agit d'un arbre des similarité selon la théorie de Lerman (1981) mais avec notre propre programmation, du graphe implicatif, d'une hiérarchie cohésitive qui sont spécifiques de l'analyse implicative. Nous allons maintenant les présenter dans cet ordre.

3 Calcul de l'arbre des similarités

L'arbre des similarités, figure 1, calcule pour chaque couple de variables la similarité entre celles-ci. Ensuite, il agrège des classes constituées elles-mêmes d'autres classes. Les

[1] d'autres algorithmes sont également utilisables comme celui de Fisher

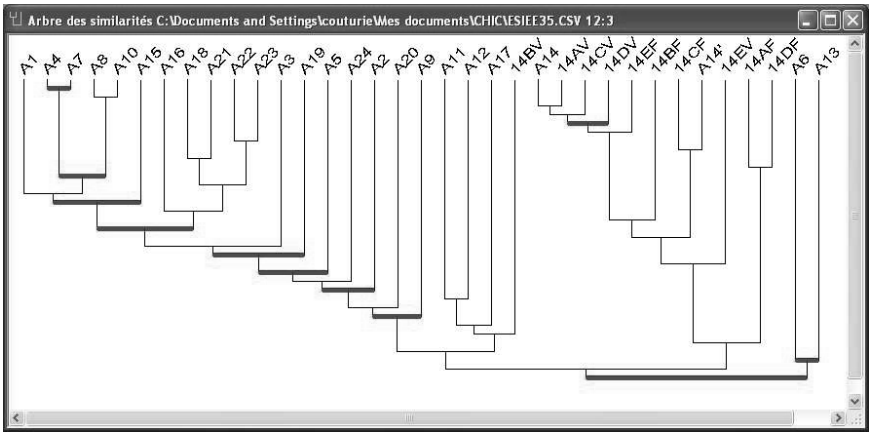

FIG. 1 – *Arbre des similarités.*

niveaux identifiés par un trait rouge (en gras sur la figure) sont des niveaux significatifs dans la mesure où ceux-ci ont plus de signification classifiante que les autres niveaux. L'algorithme utilisé est l'algorithme de la vraisemblance du lien (AVL) de Lerman (1981).

4 Calcul du graphe implicatif

Le calcul du graphe implicatif permet d'obtenir un graphe sur lequel les variables qui possèdent une intensité d'implication supérieure à un certain seuil sont reliées par une flèche représentant l'implication. La figure 2 représente un exemple de graphe implicatif. CHIC permet de sélectionner 4 seuils différents et modulables d'implication. L'utilisateur peut disposer les valeurs comme il le souhaite. Les différentes options permettent par exemple de faire apparaître les fermetures transitives, de minimiser le nombre de croisements (par l'utilisation d'un algorithme de dessin de graphe).

Il est possible de choisir une zone de travail par défaut et la faire évoluer au fil de l'utilisation. Au début d'un traitement de grande taille, il est préférable de faire intervenir toutes les données et donc de disposer d'une grande surface de travail qui peut être largement supérieure à la taille de l'écran. Puis au cours de l'interprétation, l'utilisateur peut se rendre compte que seules certaines variables lui semblent utiles pour son interprétation. Dans ce cas, il supprime temporairement les variables désirées grâce à une boite de dialogue prévu à cet effet. Ensuite, CHIC met à jour à nouveau le graphe des implications. A tout moment il est possible d'ajouter ou de supprimer des variables dans l'analyse que l'on effectue. De plus il est possible de calculer des règles avec des conjonctions de variables dans la partie prémisse. On peut avoir des règles de la forme $a \wedge b \wedge c \Rightarrow d$. Dans ce cas le nombre de règles est souvent très important, c'est

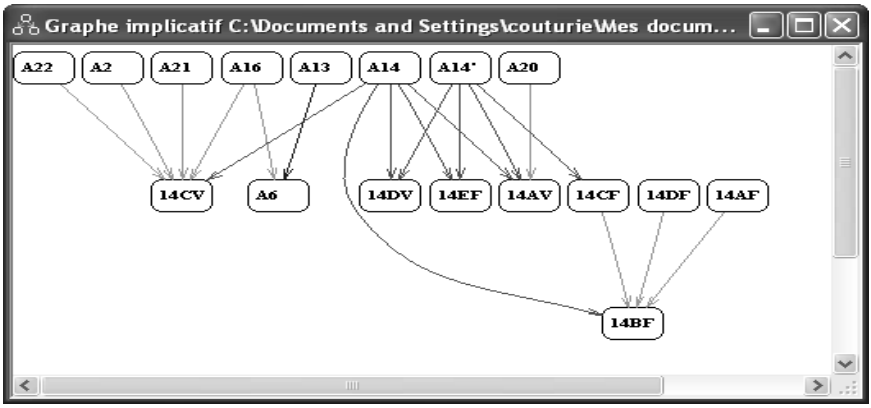

FIG. 2 – *Graphe implicicatif.*

pourquoi un indice dit d'originalité ne retient que les conjonctions qui ne paraissent pas triviales.

Il est possible de sauvegarder l'état d'un graphe, c'est-à-dire la disposition des variables, les seuils d'implication, la sélection ou non de chaque variable. Ainsi l'utilisateur peut reprendre un graphe qu'il avait organisé soigneusement lors d'une précédente session. De plus, il est possible de sauvegarder plusieurs états sur le même graphe et ainsi mettre en évidence différentes parties du graphe. On peut également exporter un graphe sous word ou excel.

5 Calcul de la hiérarchie cohésitive

La hiérarchie cohésitive est, en première approche, à l'implication ce que l'arbre des similarités est à la similarité. Mais cette hiérarchie est orientée Gras and Kuntz (2005). Des classes de variables ou de règles entre variables sont constituées à partir des implications entre celles-ci. L'algorithme agrège à chaque étape les variables conduisant à la cohésion la plus forte à cette étape. L'exemple de la figure 3 représente la hiérarchie cohésitive obtenue avec les mêmes données que l'exemple de l'arbre de similarités.

Au premier niveau de la hiérarchie, on remarque que la classe $(A7,\ A1)$ est créée. Elle représente le fait que la variable $A7$ implique la variable $A1$ avec une intensité plus forte que tous les autres couples de variables. Ce premier niveau de la hiérarchie est d'ailleurs significatif comme l'indique la flèche rouge (en gras sur la figure). Plus loin dans la hiéarchie, la classe $(A8, (A7, A1))$ est formée, elle représente $A8 \Rightarrow (A7 \Rightarrow A1)$ c'est-à-dire $A8 \wedge A7 \Rightarrow A1$. Contrairement à l'algorithme de l'arbre de similarité, l'algorithme construisant la hiérarchie cohésitive constitue de manière quasi systématique plusieurs classes et arrête son processus de construction dès que la cohésion entre variables ou entre règles devient trop faible.

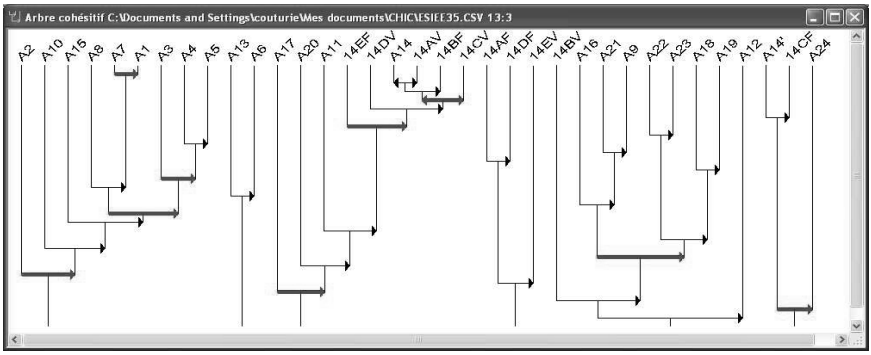

FIG. 3 – *Arbre cohésitif.*

Après avoir montré les possibilités de traitements de CHIC, nous allons regarder quelles possibilités complémentaires permettent d'affiner les interprétations des résultats précédents.

6 Autres possibilités de CHIC

A l'initialisation de chaque calcul, CHIC indique à l'utilisateur des informations telles que le nombre d'occurrences, la moyenne et l'écart-type de chaque variable.

Lorsqu'on analyse des données contenant des grands effectifs, l'option calcul entropique est préférée à l'option calcul classique car le calcul d'une règle prend également en compte la contraposée de celle-ci. Ainsi la qualité de la règle s'en trouve renforcée.

De plus, à l'instar des méthodes factorielles, il est possible de connaître quels sont les sujets et les catégories de sujets (variables supplémentaires, notées v.s., comme l'âge, la catégorie socio-professionnelle,...) les plus en accord ou les plus « responsables » des implications calculées. Un indice, dit de typicalité, permet de désigner les v.s. qui présentent un comportement, vis-à-vis des attributs constitutifs d'une classe, comparable à celle de la tendance générale de la population.

Un second indice, dit de contribution, permet de désigner les responsabilités externes des variables supplémentaires sur les implications obtenues dès lors que celles-ci présentent une certaine significativité. Il se calcule de façon comparable au précédent mais en faisant cette fois référence à une implication qui serait stricte au sein des règles génériques. L'expert peut ainsi interpréter plus aisément la correspondance entre les sujets, leurs catégories et les attributs qu'ils instancient.

7 Conclusion et perspectives

Le logiciel CHIC permet d'effectuer différents traitements statistiques basés sur l'étonnement statistique (analyse des similarités ou analyse implicative). Nous avons

essayé d'uniformiser les différentes techniques ou options (variables supplémentaires, contribution des individus, utilisation de l'entropie) pour chaque traitement afin de faciliter l'utilisation du logiciel. Il est possible de sauver les calculs intermédiaires et ainsi accélérer les prochaines utilisations d'un même fichier de données. Les données peuvent être de différents types (binaires, fréquentielles, modales, intervalles) ce qui permet d'utiliser CHIC pour de nombreuses analyses dans lesquelles les variables ne sont pas du même type.

Références

Couturier, R. (2000). Traitements de l'analyse implicative avec chic. In *Journées sur l'implication statistique*, pages 33–50, Caen.

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

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

Gras, R., Couturier, R., Blanchard, J., Briand, H., Kuntz, P., and Peter, P. (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, pages 3–32. RNTI-E-1, Cepaduès Editions. I.S.B.N. 2.85428.646.4.

Gras, R., Diday, E., Kuntz, P., and Couturier, R. (2001a). Variables sur intervalles et variables-intervalles en analyse implicative. In *8ème Congrès de la SFC*, pages 166–173.

Gras, R. and Kuntz, P. (2005). Discovering r-rules with a directed hierarchy. *Soft Computing*. Sous presse.

Gras, R., Kuntz, P., and Briand, H. (2001b). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines*, 154-155 :9–29. ISSN 0987 6936.

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

Summary

This paper aims at showing the features of the software CHIC (Cohesive Hierarchical and Implicative Classification) to carry out some data analyses. It is based on the Implicative Statistical Analysis (A.S.I. in french) developed by Régis Gras and its collaborators. The main objective of A.S.I. focuses on the measure of association rules of kind : "if a then b" in a population with variables a and b. CHIC improves its response, based on statistical theories, by evaluating the contribution of the subjects to build the rule. This paper presents the principle to follow to use the software and its features.