TRAITEMENT DE L'ANALYSE STATISTIQUE DANS CHIC

Raphaël COUTURIER[1]

RESUME - *Cet article a pour but de montrer les possibilités offertes par le logiciel CHIC pour effectuer de l'analyse de données. CHIC signifie Classification Hiérarchique Implicative et Cohésitive. Il repose sur la théorie implicative développée par Régis Gras et ses collaborateurs. Les fondements théoriques sont introduits dans sa conférence. L'article présent explique la démarche à suivre pour utiliser le logiciel ainsi que les possibilités offertes par celui-ci en présentant différents exemples.*

MOTS-CLES - *Traitements avec CHIC, analyse implicative*

PROCESSING STATISTICAL ANALYSIS WITH CHIC

SUMMARY - *The goal of this article is to show how the software CHIC can be used to process data analysis. CHIC stands for Cohesitive and Hierarchical Implicative Classification. It is founded on the implicative theory developped by Régis Gras and its collaborators. The theory details are introduced into his conference. This article explains the use of the software and the features on different examples.*

KEYWORDS - *Processing with CHIC, implicative analysis*

[1] Maître de Conférences à l'I.U.T. de Belfort, BP 527, rue E.Gros, 90016 Belfort cedex,

couturier@iut-bm.univ-fcomte.fr

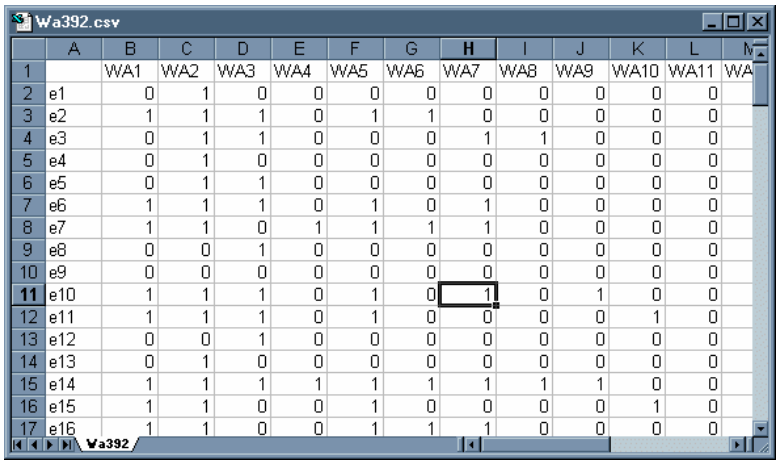
1 INTRODUCTION

CHIC est le résultat informatique des travaux sur l'analyse implicative. La version actuelle de ce logiciel a été portée en C++ sous Windows il y a 5 ans environ à partir d'une version antérieure en Pascal, mais avec des développements importants et avec une plus grande convivialité [2]. Depuis elle a subi régulièrement de nombreuses modifications tant au niveau pratique qu'au plan théorique en intégrant de nombreux nouveaux modes de calculs. 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 [4] et [5] et reprise dans ces Actes. La documentation de CHIC [2] peut apporter d'autres éclaircissements sur les concepts introduits dans cette présentation de CHIC.

Dans la partie 2, nous commençons par présenter 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 un arbre cohésitif. La partie 6 introduit le graphe inclusif. La partie 7 détaille toutes les possibilités communes aux différents types d'analyse pour affiner les résultats et les interprétations d'une analyse. Finalement la partie 8 donne une conclusion et des perspectives. En annexe, nous présentons les travaux dirigés qui ont été effectués lors des journées sur la fouille de données en trois séances.

2 MISE EN FORME DES DONNEES

Les données sont disposées sous forme d'un tableau de contingence, c'est-à-dire qu'à chaque variable que nous souhaitons évaluer, nous faisons correspondre le résultat de l'évaluation de chaque objet ou individu à cette variable. La figure ci-dessous montre comment sont disposées les données avec le logiciel Excel. Les variables sont disposées en colonne (ici de WA1 à WA12) et les individus sont arrangés en colonne (ici de e1 à e16). Toutes les cases du tableau doivent pour le moment être remplies. 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 secondaires 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 ramenées 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 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 des tableurs, chaque champ étant


Wa392.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		WA1	WA2	WA3	WA4	WA5	WA6	WA7	WA8	WA9	WA10	WA11	WA
2	e1	0	1	0	0	0	0	0	0	0	0	0	
3	e2	1	1	1	0	1	1	0	0	0	0	0	
4	e3	0	1	1	0	0	0	1	1	0	0	0	
5	e4	0	1	0	0	0	0	0	0	0	0	0	
6	e5	0	1	1	0	0	0	0	0	0	0	0	
7	e6	1	1	1	0	1	0	1	0	0	0	0	
8	e7	1	1	0	1	1	1	1	0	0	0	0	
9	e8	0	0	1	0	0	0	0	0	0	0	0	
10	e9	0	0	0	0	0	0	0	0	0	0	0	
11	e10	1	1	1	0	1	0	1	0	1	0	0	
12	e11	1	1	1	0	1	0	0	0	0	1	0	
13	e12	0	0	1	0	0	0	0	0	0	0	0	
14	e13	0	1	0	0	0	0	0	0	0	0	0	
15	e14	1	1	1	1	1	1	1	1	1	0	0	
16	e15	1	1	0	0	1	0	0	0	0	1	0	
17	e16	1	1	0	0	1	1	1	0	0	0	0	

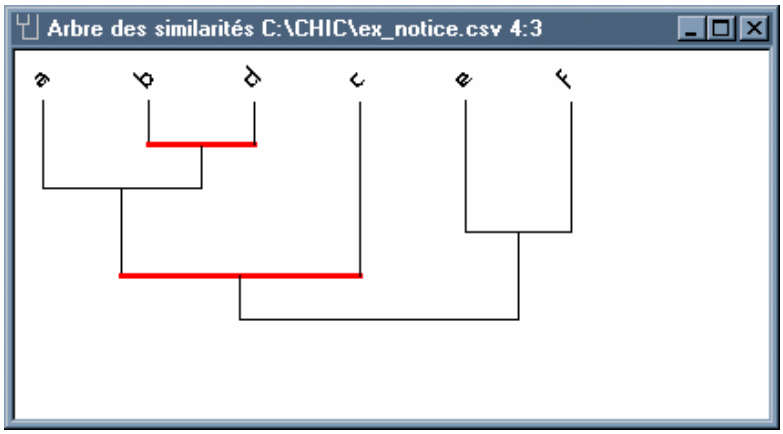
Wa392

séparé par un point virgule. Notons d'ailleurs qu'il est possible de générer automatiquement des données en respectant ce format avec un outil quelconque.

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 I.C. LERMAN [6] mais avec notre propre programmation, du graphe implicatif, d'un arbre cohésitif et d'un graphe inclusif qui sont spécifiques de l'analyse implicative. Nous allons maintenant les présenter dans cet ordre.

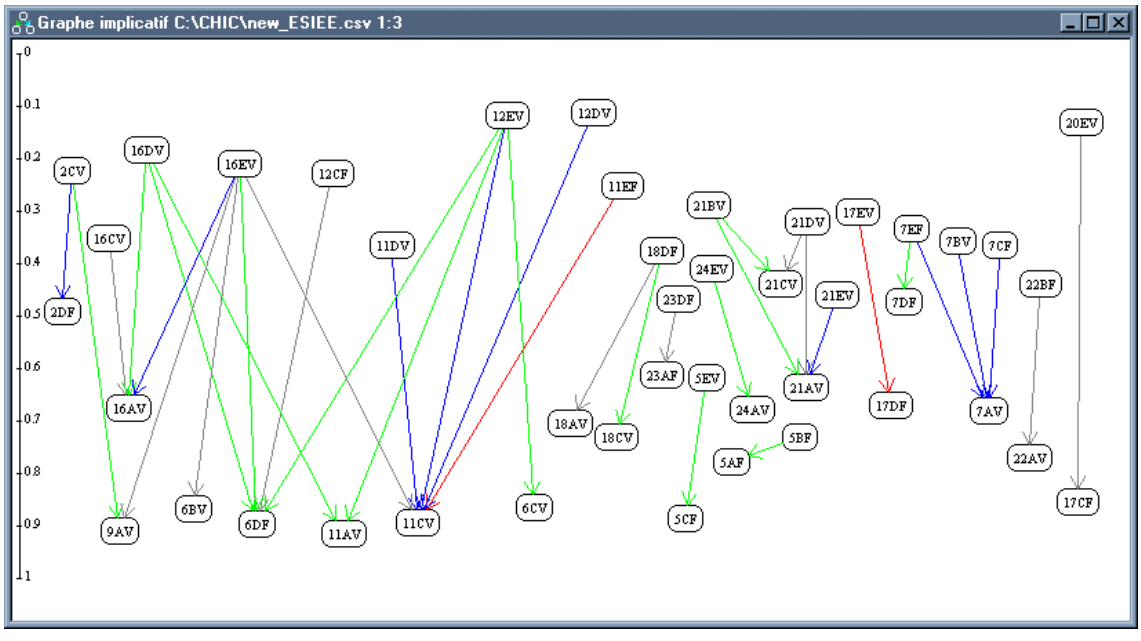
3 CALCUL DE L'ARBRE DES SIMILARITES

L'arbre des similarités calcule pour chaque couple de variables la similarité entre celles-ci. Ensuite, il agrège des classes constituées elles-mêmes d'autres classes. Sur l'arbre ci-dessous, les variables *b* et *d* sont dans un premier temps les variables les plus similaires. Ensuite l'algorithme choisit d'associer la variables *a* à la classe (*b*,*d*), ainsi la nouvelle classe est (*a*,(*b*,*d*)). A l'itération (ou au niveau) 3, la classe (*e*,*f*) est formée. Finalement au dernier niveau, nous obtenons une seule classe, mais les classes (a,b,d,c) et (e,f) sont dissemblables. Les 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.



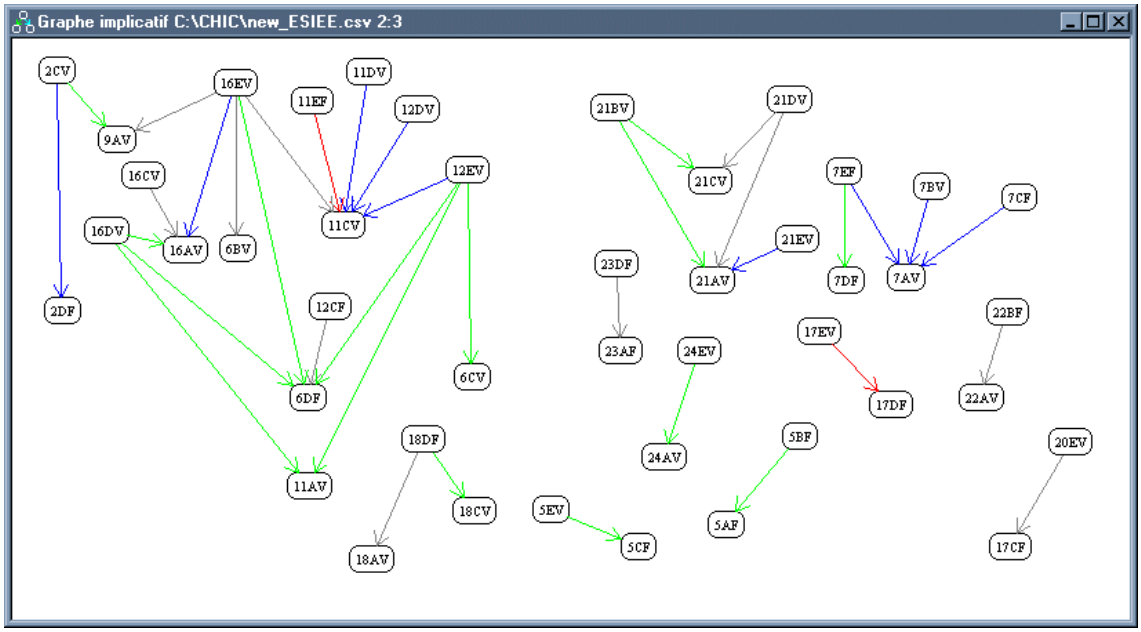
4 CALCUL DU GRAPHE IMPLICATIF

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

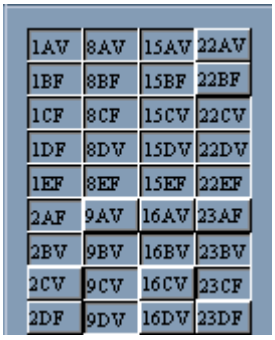


une flèche représentant l'implication. CHIC permet de sélectionner 4 seuils différents d'implication. Sur le graphe précédent, nous avons choisi 4 seuils identifiées par des couleurs différentes (à l'écran). L'utilisateur peut disposer les valeurs comme il le souhaite. Dans ce cas, CHIC ne nécessite pas de refaire les calculs d'implications puisqu'il les mémorise. Ainsi l'utilisateur peut explorer très rapidement de nombreuses pistes. Il peut choisir les couleurs qu'il désire et peut modifier la taille des flèches ainsi que celle des variables. Dans un premier temps, CHIC place les variables sans essayer de minimiser, de façon optimale, le nombre de croisements. Les fermetures transitives ne sont pas affichées sur le graphe mais une option permet de le faire à tout moment et également de les supprimer de manière transparente pour l'utilisateur.

Une autre option permet de conserver les occurrences des variables. Elle est très utile pour l'interprétation des résultats. Par exemple, sur le graphe que nous avons représenté, les variables 9AV, 6BV, 6DF, 11AV, 11CV, 6CV, 5CF et 17CF sont des variables dont le support est proche de 90%. Ce sont donc des variables qui seront, probablement, dans la conclusion d'une règle implicative, alors que les variables avec un faible support seront plutôt dans la prémisse d'une règle. Cependant étant donné la taille de certains graphes, il est préférable de ne pas respecter les occurrences des variables afin de simplifier la lecture du graphe. CHIC dispose d'une option à cet effet. La figure ci-dessous représente le même graphe sur lequel nous avons déplacé, avec la souris, les variables afin de supprimer les croisements, mais, cette fois, nous ne respectons pas les occurrences.



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. Nous avons représenté une partie de la boite de dialogue permettant de faire le choix des variables intervenant dans le graphe. Ensuite, CHIC met à jour à nouveau le graphe des implications. A tout moment il est possible d'ajouter ou de

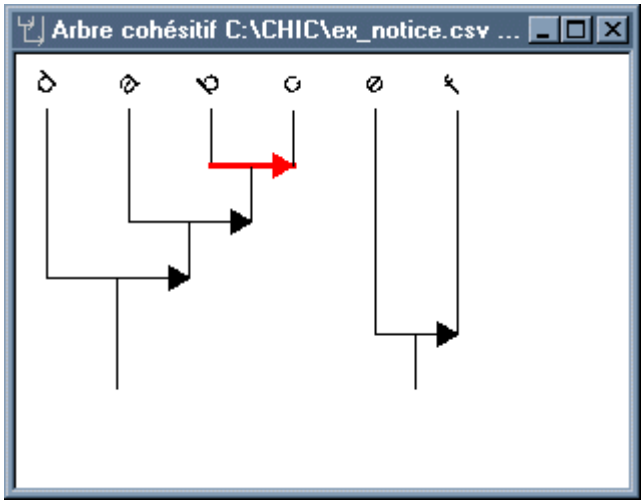


supprimer des variables dans l'analyse que l'on effectue. Par l'intermédiaire du menu contextuel de la souris, il est possible de déselectionner toutes les variables qui ne sont pas reliées par une implication afin de rendre la lecture plus simple.

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.

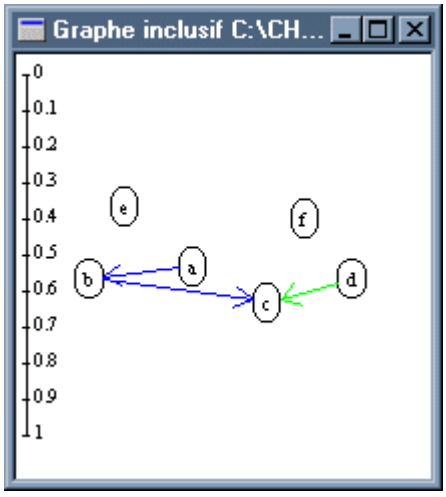
5 CALCUL DE L'ARBRE COHESITIF

L'arbre cohésitif est, en première approche, à l'implication ce que l'arbre des similarités est à la similarité. Dans cet arbre, 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 ci-contre représente l'arbre cohésitif obtenu avec les mêmes données que l'exemple de l'arbre de similarités. Au premier niveau de la hiérarchie, on remarque que la classe (*b*, *c*) est créée. Elle représente le fait que la variable *b* implique la variable *c* 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). Au second niveau, la classe (*a*, (*b*, *c*)) est formée. Cette classe à trois composantes admet la plus forte cohésion parmi celles de toutes les classes possibles à trois composantes et celle de tout autre classe à deux composantes. Puis finalement la classe (*e*, *f*) est créée à la dernière étape. Contrairement à l'algorithme de l'arbre de similarité, l'algorithme construisant l'arbre cohésitif 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.



6 CALCUL DU GRAPHE INCLUSIF

Le graphe inclusif hérite de la plupart des caractéristiques du graphe implicatif. La différence fondamentale se situe au niveau du calcul de l'inclusion définie par A. BODIN [1]. Plusieurs seuils représentées par plusieurs couleurs permettent de visualiser les différents niveau d'inclusion. Le changement des valeurs des seuils nécessite de recalculer la totalité des inclusions car les calculs avec d'autres seuils de confiance sont différents. De ce point de vue, nous perdons la rapidité offerte par le graphe implicatif lorsque les graphes sont complexes. L'exemple ci-

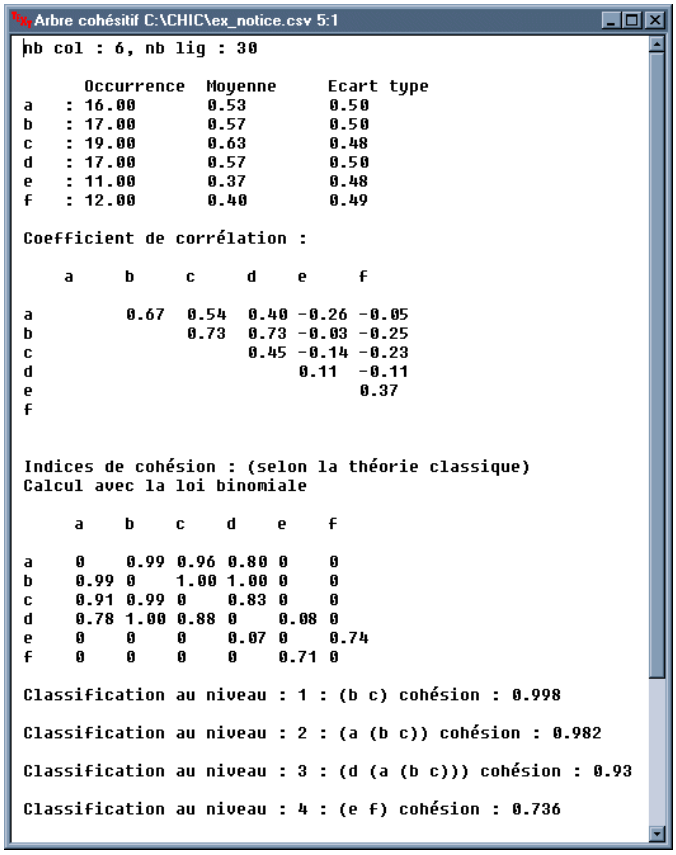


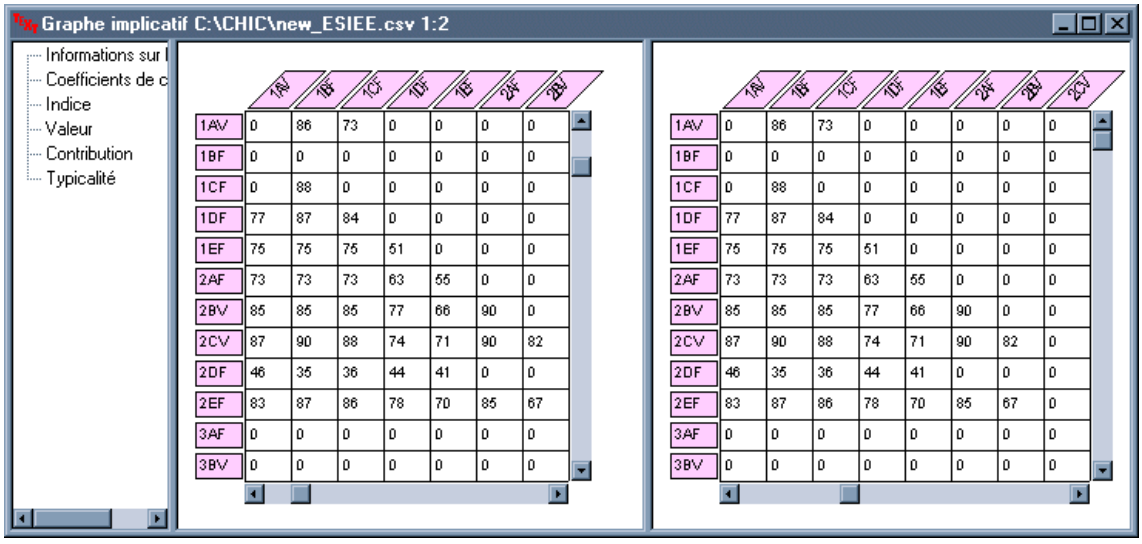
contre montre le graphe inclusif obtenu à partir des données utilisées initialement pour l'arbre des similarités. Dans les options, nous pouvons définir la valeur du niveau de confiance pour le graphe inclusif.

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

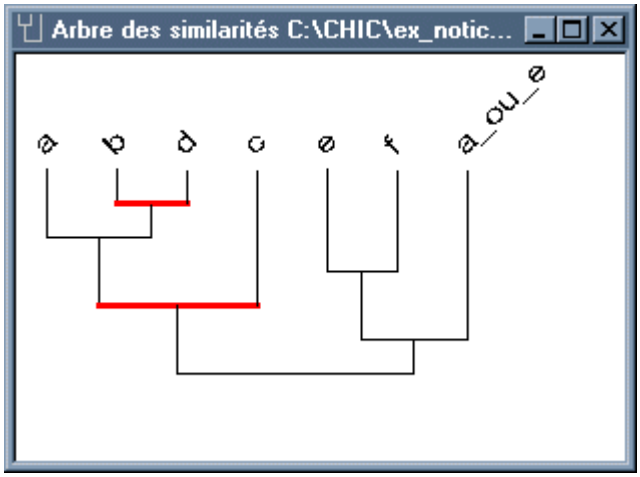
7 AUTRES POSSIBILITES DE CHIC

A l'initialisation de chaque calcul, CHIC commence par établir le nombre de colonnes et de lignes du tableau de contingence sur lequel nous travaillons. Ensuite pour chaque variable, il affiche son nombre d'occurrences, sa moyenne et son écart-type. Il affiche également les coefficients de corrélation linéaire entre les variables. Ses données sont souvent très utiles pour interpréter les résultats d'une analyse. L'image ci-contre représente une partie des renseignements que donne CHIC sur le fichier en cours d'analyse (ici les données sont celles de l'arbre cohésitif de la partie 5). L'image ci-dessous montre également une partie de l'affichage que fournit CHIC au niveau graphique. Lorsqu'il s'agit de tableaux de contingences, l'affiche sous forme graphique est plus facile à lire (les données se rapportent au graphe implicatif de la partie 4).

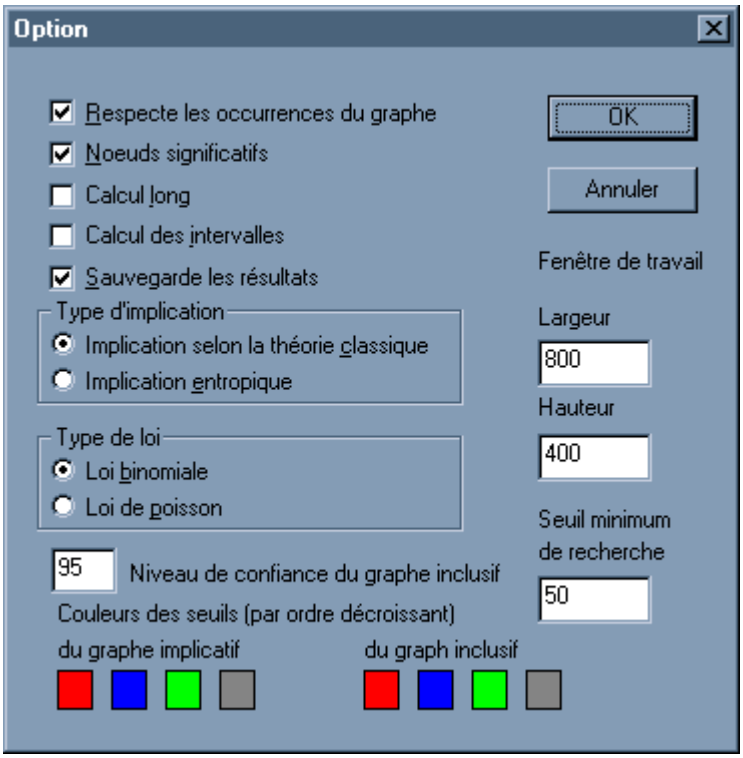




Lors d'une analyse, il est parfois intéressant de pouvoir créer de nouvelles variables par conjonction ou disjonction de variables existantes. CHIC permet de créer de telles variables par l'intermédiaire d'une boîte de dialogue qui renseigne l'utilisateur sur la nouvelle variable qu'il souhaite créer. Ensuite la nouvelle variable possède le même statut que les variables initiales de CHIC. La figure suivante montre l'arbre des similarité de la partie 3 dans lequel nous avons ajouté la variable *a_ou_e*.



L'interprétation des résultats fournis par CHIC peut être facilitée grâce à l'utilisation de variables secondaires (ou supplémentaires) [3]. De telles variables n'interviennent pas directement dans les calculs Par contre, elles interviennent dans l'interprétation qu'un utilisateur peut faire. Prenons un exemple : supposons que l'on souhaite analyser un questionnaire auquel ont répondu des individus appartenant à différentes catégories sociales. Dans ce cas, l'origine sociale des individus n'intervient pas comme variable principale. A partir des variables principales, CHIC construit une hiérarchie ou un graphe à l'intérieur desquels certaines variables sont liées par similarité, implication ou inclusion. Pour interpréter finement les résultats, nous pouvons étudier la responsabilité des catégories sociales sur la formation des graphes ou des classes. Ainsi, il est possible de savoir si une catégorie contribue plus fortement qu'une autre à la création d'une structure (une classe dans une hiérarchie ou un chemin dans un graphe). Cette information est d'une grande richesse car elle nous renseigne sur les catégories d'individus qui sont à la base des structures obtenues.



Afin d'établir ces calculs, CHIC définit pour chaque relation entre les variables, le groupe optimal à cette relation. Un groupe optimal est un ensemble d'individus qui contribue plus à la formation d'une relation que son ensemble complémentaire, c'est-à-dire… les autres.

Lorsque la population étudiée possède un grand nombre d'individus, A. Bodin à mis en évidence un

biais dans le calcul classique de l'implication. Ce biais est dû à la confusion que nous avons tendance à faire entre étonnement et implication. L'analyse implicative est fondée sur l'étonnement statistique. Or celui-ci varie très fortement avec la dilation des ensembles. Dans la pratique, avec de grands ensembles, il est étonnant au niveau statistique que certains ensembles soient disposés d'une certaine manière (une inclusion que l'on pourra qualifier de faible) sans pour autant que l'on souhaite parler d'implication au niveau mathématique. Cette raison nous a conduit à définir une nouvelle mesure implicative qui rende davantage compte de l'inclusion. Cette mesure est basée sur la notion d'entropie d'où le nom d'implication entropique. Elle ne possède pas le biais que nous venons de présenter. Dans les options de CHIC, nous pouvons choisir d'utiliser l'implication classique évoquée précédemment ou l'implication entropique. L'image précédente nous montre les différentes options de CHIC.

8 CONCLUSION ET PERSPECTIVES

Le logiciel CHIC permet d'effectuer différents traitements statistiques basé 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.

Pour le moment, le logiciel est capable de traiter rapidement des tableaux de contingences d'une taille 200x100 000 (ces limites dépendent de la capacité de calcul et mémoire de l'ordinateur)[2]. 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. Les autres articles de ce colloque témoignent des utilisations très diverses.

Actuellement nous essayons d'intégrer une méthode permettant d'améliorer la disposition des variables sur les graphes. Il existe de nombreux algorithmes pour placer des sommets en minimisant les croisements d'arcs mais ces algorithmes sont complexes à implanter et peuvent être longs à exécuter. De plus nous devons les adapter afin d'intégrer certaines contraintes inhérentes à CHIC (par exemple : les variables n'ont pas la même taille sur l'écran). De plus, un nouvel algorithme va prochainement, permettre d'effectuer une réduction du nombre de variables, sans grande perte d'information.

Il est également prévu que l'interface graphique subissent quelques améliorations, par exemple, nous comptons permettre à l'utilisateur de choisir quelles sont les variables retenues comme principales ou comme supplémentaires directement dans CHIC, sans être l'obliger à retourner dans le tableur.

Enfin, nous comptons intégrer le calcul automatique des variables conjonctives. Cette technique est actuellement en cours d'élaboration. Notre objectif est de donner à l'utilisateur un ensemble de règles dans lequel les variables sont obtenues par conjonction des variables initiales. Les algorithmes traitant ce genre de problème sont confrontés à deux difficultés essentielles. La première est due aux grandes capacités

---

[2] En utilisant une base de données nous sommes capables de traiter des fichiers de plus d'un million d'individus avec plusieurs milliers de variables, mais pour cela nous utilisons le système d'exploitation Linux sur lequel nous avons porté une partie des calculs de CHIC.

requises tant au niveau mémoire qu'au niveau temps de calcul. La seconde réside dans le nombre de règles générées qui peut dépasser couramment le million ou le milliard.

BIBLIOGRAPHIE

[1] BODIN A., « Modèles sous-jacents à l'analyse implicative et outils complémentaires ». Prépublication 97-32, IRMAR, Décembre 1997.

[2] BODIN A., COUTURIER R., GRAS R., « Analyse d'une épreuve de concours par la méthode implicative, présentation interactive », SFC'96, Vannes, France, 1996.

[3] COUTURIER R, GRAS R., « Introduction de variables supplémentaires dans une hiérarchie de classes et application à CHIC », SFC'99, Nancy, France, 1999.

[4] GRAS R. et coll., L'implication Statistique, Grenoble, La Pensée Sauvage, 1996.

[5] GRAS R., KUNTZ P., « Les fondements de l'analyse statistique implicative et leur prolongement pour la fouille de données », Mathématique et Sciences Humaines, à paraître.

[6] I.C. LERMAN, Classification et analyse ordinale des données, Dunod, 1981.