Recoder les variables pour obtenir un modèle implicatif optimal

Martine Cadot

PRE-PUBLICATION

Recoder les variables pour obtenir un modèle implicatif optimal

Martine Cadot*

* LORIA/UdL, 54600 Villers-lès-Nancy, France
Martine.Cadot@loria.fr

Résumé. Il existe un certain nombre de méthodes permettant d'obtenir à partir de données individuelles un modèle de catégorisation du type X→Y, X représentant un ensemble de caractéristiques numériques des individus et Y leur catégorie. Nous faisons un tour rapide de ces méthodes en appliquant les plus utilisées aujourd'hui au jeu de données des « Iris de Fisher ». La confrontation des divers modèles obtenus nous incite à privilégier l'A.S.I. (Analyse Statistique Implicative) pour traiter ce type de données, après un recodage particulier des variables quantitatives. Ce chapitre reprend et élargit une étude qui a fait l'objet d'une communication lors du colloque A.S.I.8 (Cadot et al. 2015) dans laquelle nous montrions l'intérêt de la méthodologie choisie (A.S.I. après recodage particulier) pour le traitement de données acoustiques.

1 Introduction

Notre but est de choisir une méthode permettant d'extraire à partir de données individuelles un modèle reliant les caractéristiques numériques X des individus et leur catégorie Y. Le modèle que nous visons est de la forme X→Y, se lisant selon les cas : X implique Y, X cause Y, X explique Y, X produit Y, etc. (Gras *et al.,* 2013). Il est « asymétrique » dans la mesure où les variables n'ont pas toutes le même rôle[1] : la variable Y est la « variable à expliquer » (appelée également variable dépendante), et les variables X sont les « variables explicatives » (appelées également variables indépendantes).

L'A.S.I. n'est pas le seul cadre théorique permettant d'obtenir de tels modèles : depuis plus d'un siècle « le modèle linéaire » des statisticiens fournit des solutions, ainsi que plus récemment certaines branches de l'informatique comme « l'apprentissage automatique » dédiée au traitement des données, ou « les modèles de aide au raisonnement » dédiés à la prise de décision.

Dans un premier temps nous faisons un tour rapide de ces diverses approches méthodologiques aboutissant à un modèle de liaison entre une variable catégorielle Y et un ensemble de variables quantitatives X. Nous voyons comment cette liaison peut s'exprimer quantitativement de diverses façons, notamment après recodages des variables quantitatives, et être validée statistiquement. Pour rendre notre exposé plus lisible, nous l'avons illustré à l'aide d'un

[1] Pour un exposé détaillé sur le statut des variables dans la planification des expériences en psychologie expérimentale voir Hoc (1983), en agronomie voir Dagnelie.(2003), en médecine et biologie voir Schwartz (1991).

jeu de données bien connu, « les Iris de Fisher » (Fisher 1936), formé des valeurs de 5 variables (X : 4 mesures de fleurs, et Y : la catégorie de la fleur parmi 3 catégories) recueillies sur 150 fleurs. À partir de cette réflexion nous détaillons dans un deuxième temps les méthodes de l'A.S.I. (Gras *et al.* 1996, 2009, 2013) et l'utilisation du logiciel C.H.I.C. (Version 6.0, Copyright© 2012) sur les données Iris, d'abord sans recodage des variables quantitatives puis en les recodant. Dans un troisième temps nous donnons les raisons qui nous ont fait choisir l'A.S.I. pour traiter nos données avant de conclure.

A part le logiciel C.H.I.C. pour l'A.S.I., HUGIN (version Lite disponible gratuitement http://www.hugin.com/) pour les réseaux bayésiens et le tableur Excel de Microsoft pour la régression linéaire ainsi que le tableau de contingence et une partie des graphiques, tous les modèles testés l'ont été à l'aide de scripts écrits dans le langage/logiciel R (disponible gratuitement sous licence GNU http://www.r-project.org/), dont le code est mis en annexe de la communication présentée à A.S.I.8 (Cadot et al. 2015).

Les Iris de Fisher

Ce jeu de données est téléchargeable depuis UCI repository (https://archive.ics.uci.edu/ml/datasets/Iris). Il est formé d'une variable catégorielle Y (l'espèce) à 3 modalités (iris-setosa, iris-versicolor, iris-virginica), avec 50 fleurs de chaque espèce, et d'un ensemble de 4 variables numériques X, qui sont la longueur et la largeur moyennes des sépales et des pétales (Sepal_Length, Sepal_Width, Petal_L, Petal_W). Il a été utilisé par de nombreux chercheurs en analyse de données désirant éprouver une nouvelle méthode statistique, soit en classification pour retrouver les 3 groupes de fleurs, soit en discrimination pour trouver les règles d'attribution d'une fleur à un groupe. C'est dans ce deuxième cadre que nous nous plaçons. Dans la figure 1, nous avons représenté les 150 fleurs dans l'espace formé de 3 variables de X (Petal_L, Petal_W et Sepal_W), avec une couleur par espèce.

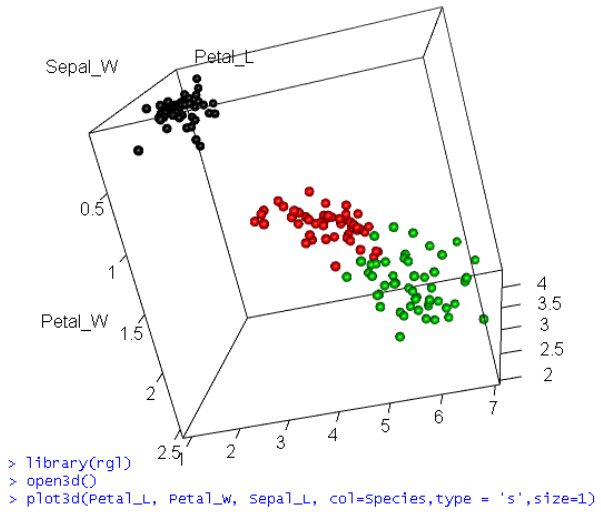

FIG. 1 – Les 150 Iris de Fisher dans l'espace de 3 variables, avec une couleur par espèce.

On peut remarquer que la longueur et la largeur des pétales permettent d'affecter plus ou moins facilement une espèce à chaque iris : les plus petits pétales correspondent à l'espèce

Setosa (en noir), les moyens à l'espèce Versicolor (en rouge) et les plus grands à l'espèce Virginica (en vert). Plus précisément, parmi les 3 nuages de points de couleurs différentes de la figure 1, seuls le nuage de points rouges et celui de points verts ont une petite partie commune d'une dizaine d'individus. Dans la mesure où un simple découpage en quelques sous-espaces de la figure 1 permet d'établir l'espèce des 150 iris avec une dizaine d'erreurs, on attend de chacune des méthodes de discrimination étudiées qu'elles atteignent des taux de reconnaissance de plus de 90%.

2 Approches méthodologiques courantes

Pour obtenir un modèle à partir de données, on peut procéder de trois façons différentes, que nous décrivons chacune dans une sous-section. D'abord dans le cadre de la statistique classique qui permet de valider un modèle à partir d'échantillons tirés des données. Utilisée depuis plus d'un siècle, la statistique a l'avantage de fournir un cadre rigoureux pour des modèles simples de relations entre quelques variables, mais cela suppose que les données vérifient un certain nombre d'hypothèses contraignantes. Ensuite les méthodes d'apprentissage automatique, utilisées depuis quelques dizaines d'années, qui produisent un modèle à partir de calculs intensifs faits en plusieurs étapes avec des découpages variés des données. Le modèle est parfois complexe, pas toujours explicite, mais il a un fort pouvoir de prédiction. On exposera pour finir des méthodes produisant un modèle de aide au raisonnement, qui s'appuient sur de la combinatoire. La qualité de ce dernier type de modèle dépend cette fois essentiellement des interventions de l'utilisateur.

Dans les trois parties suivantes, nous faisons un tour rapide des modèles courants en donnant pour chacun une référence renvoyant à un manuel pédagogique dans lequel cette méthode est détaillée parmi d'autres, une partie restreinte de ces références pouvant donc suffire pour couvrir l'ensemble des modèles décrits.

2.1 Les modèles statistiques classiques

Un des modèles les plus utilisés est le ***modèle linéaire*** (Prum 1996). Sa version de base est le ***modèle de régression linéaire***, pour lequel X et Y sont deux variables quantitatives. Il se résume à la connaissance des 2 coefficients a et b de la ***droite de régression*** $Y_{pred}=aX+b$ sur laquelle sont situés les points de coordonnées (X, Y_{pred}) dans le plan de X et Y (par exemple la première équation à droite de la figure 2 exprime la liaison entre la longueur prédite du pétale et la longueur du sépale pour les iris setosa, et elle est représentée par la droite bleue du graphique). Quand X est formé de p variables quantitatives (p>1), le modèle de régression linéaire s'écrit $Y_{pred}=XA+B$ avec A et B des vecteurs de p valeurs fixées, et c'est l'équation d'un hyper-plan dans l'espace de dimension p+1 de X et Y.

La régression linéaire peut s'étendre à des variables X non toutes quantitatives et on parlera plutôt de ***modèle linéaire généralisé*** (Baillargeon 2000). Par exemple en Figure 2, on a exprimé la dépendance de Y (longueur des pétales) en fonction de X (petal_L, la longueur des pétales, I_{vers} et I_{virg} les variables indicatrices de 2 types d'iris (respectivement iris-versicolor et iris-virginica, le troisième type étant la référence par défaut, obtenu quand les deux variables sont nulles simultanément). Si au lieu d'être quantitative, Y est binaire (1 : réussite, 0 : échec), on utilise le modèle ***logistique*** (Besse 2003) pour lequel l'équation devient $logit(P_{est})=XA+B$, où P_{est} est la probabilité que Y soit égal à 1, et $logit(P)= log(P/(1-P))$.

Le modèle logistique peut encore s'étendre à une variable Y de comptage (Y : nombre de réussites) ou ordinale, mais plus difficilement à une variable catégorielle. Toutefois, si on « éclate » en p variables binaires la variable catégorielle Y à p catégories, on peut alors chercher p modèles logistiques, un par variable binaire.

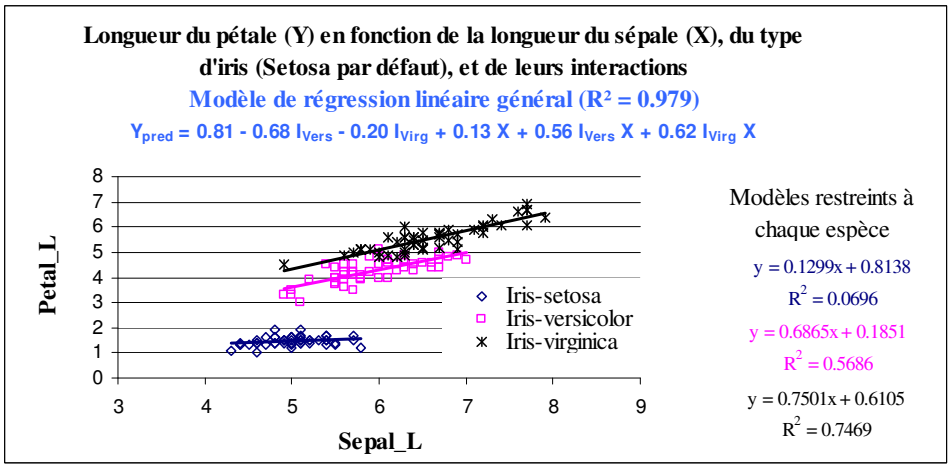

FIG. 2 – Un modèle de régression de la longueur du pétale sur la longueur du sépale et l'espèce d'iris. En haut, en bleu le modèle global prenant en compte les 150 mesures, et à droite les 3 modèles spécifiques à chaque type d'iris, ne prenant en compte que 50 mesures chacun.

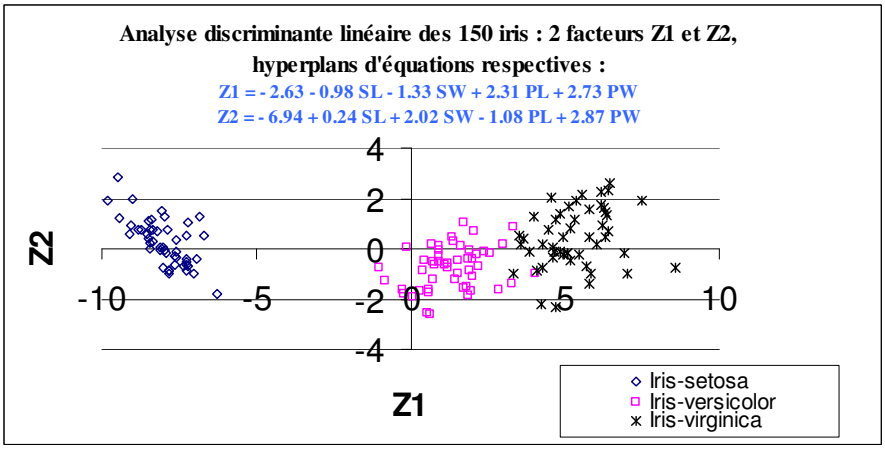

FIG. 3 – Les facteurs discriminants des 150 iris : plus de 99% de variance selon Z1.

Une extension dans une autre direction est le modèle ***log-linéaire*** (Morineau, 1996), pour lequel toutes les variables sont catégorielles, X comme Y, et c'est alors chaque variable de X qu'il convient d'éclater en intervalles, ces intervalles devenant les catégories de la variable,

et dans ce modèle, ce n'est plus Y qui dépend linéairement de X mais log(Effectif$_{pred}$) qui est une fonction linéaire de X et Y. La ***discrimination linéaire*** (Nakache 2003) est un modèle un peu différent, elle consiste à trouver un changement de repère dans X pour lequel les catégories de Y sont séparées le mieux possible. En Figure 3, on a représenté la projection des 150 iris dans l'espace des deux hyperplans obtenus pour la discrimination linéaire des 3 espèces d'Iris (un seul aurait suffit, car il produit plus de 99% de la variance).

Le « modèle linéaire » et ses extensions, font partie du cadre de la statistique classique. Ils font partie des ***modèles paramétriques*** et sont assortis de conditions d'application qui permettent de garantir la qualité des estimations produites. Quand les conditions d'application ne sont pas vérifiées (par exemple la normalité de la distribution des erreurs), des modèles non paramétriques (Siegel 1988) peuvent être utilisés, par exemple, si le nombre d'ex-æquo est réduit, en remplaçant les variables par leurs rangs (Droesbeke, 1996).

2.2 Les modèles obtenus par apprentissage automatique

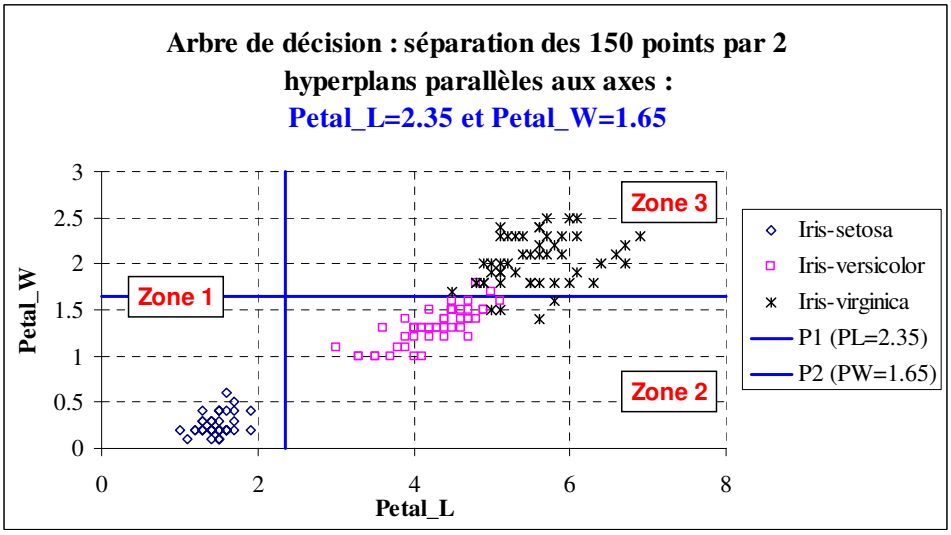

FIG. 4 – Résultat sur les 150 iris d'un arbre de décision entraîné une fois sur 75 iris tirés au hasard (25 par classe) : 2 hyperplans séparateurs et 6 erreurs (2 pour la partie apprentissage et 4 pour l'autre).

A côté de ces méthodes qui, depuis plus d'un siècle, ont permis de produire des prédictions « garanties » à partir de mesures faites sur un seul échantillon choisi avec soin (Morin 1999), depuis quelques dizaines d'années de nouvelles méthodes de prédiction sont apparues avec l'avènement de l'informatique, méthodes qui se sont généralisées dernièrement avec le libre accès à des ressources sur Internet. Issues de « l'apprentissage automatique » (Mitchell 1997) ces méthodes se proposent d'évaluer la qualité des prédictions en partitionnant l'ensemble des données en plusieurs parties de diverses façons (par exemple k parties pour la ***validation croisée***, le modèle étant construit sur la réunion de k-1 parties, testé sur la partie restante, on itère le processus k fois en changeant la partie restante). Le nombre de ces mé-

thodes est en constante augmentation (une grande partie des 6500 packages R disponibles à ce jour <http://cran.r-project.org/>). En effet, leur validité peut s'établir sans avoir besoin de théories statistiques ***asymptotiques*** (par exemple s'appuyant sur l'hypothèse de normalité), mais par une mise à l'épreuve sur des jeux de données caractéristiques d'un problème spécifique à résoudre.

Nous ne nous intéresserons ici qu'aux méthodes qui peuvent prédire une variable Y catégorielle de plus de 2 catégories à partir d'un ensemble de variables X quantitatives, et donc s'appliquer au jeu de données des Iris de Fisher. Les plus pratiquées sont des extensions des réseaux neuronaux (Ripley 1996), les MSVM (Multiple Support Vector Machine, Weston 1998) et les arbres de décision (Breiman 1984). Dans la Figure 4, nous montrons le partage en régions proposé par un arbre de décision sur les données Iris. Deux hyperplans parallèles aux axes ont suffi pour déterminer les 3 zones de points formées quA.S.I.-exclusivement d'une seule espèce d'Iris. La zone 1 correspond à Petal_L<2.35 et ne contient que les iris Setosa. La zone 2 correspond à Petal_L≥2.35 et Petal_W<1.65, et contient tous les iris Versicolor, sauf 1, ainsi que 5 iris Virginica. Et dernière zone correspondant à Petal_L≥2.35 et Petal_W≥1.65 contient la quA.S.I.-totalité des iris Virginica, ainsi qu'un iris Versicolor. Il y a 6 erreurs sur 150 iris, soit 4% d'erreurs. Du seul point de vue de l'interprétation graphique, on peut dire que les MSVM sont une extension des arbres de décision, avec des hyperplans frontières qui ne sont plus nécessairement parallèles aux axes, des frontières qui ne sont plus strictes, mais floues (ce sont des bandes), et dans un espace qui est une extension de X.

2.3 Les modèles d'aide au raisonnement

Les modèles les plus utilisés actuellement pour la prise de décision sont les ***arbres de décision*** (que nous avons exposés dans la section précédente), ***les règles d'association***, et ***les réseaux bayésiens***.

Les règles d'association sont obtenues en lançant des algorithmes de recherche d'associations dans des bases de données, ils font partie de la « fouille de données » et mettent au jour les liens entre items (Han et al. 2000). A titre d'exemple bien connu, on peut citer le lien entre les deux items « achat de bières » et « achat de couches-culottes » extrait de la base des tickets de caisse d'une chaîne de supermarchés des Etats-Unis. Plus généralement, partant d'un jeu de données formé des valeurs de q variables catégorielles X_i sur un certain nombre de sujets, l'algorithme extrait automatiquement des règles de la forme (A et B et C)→D, où le membre de gauche contient un certain nombre d'affirmations (ici il y en a 3, A, B et C) du type $X_i=C_{ik}$ (la variable X_i a pour valeur sa catégorie C_{ik}), celui de droite n'en contenant qu'une en général. Si on ne garde que les règles ayant Y en partie droite, et si on a transformé les variables quantitatives en catégorielles pour les utiliser en partie gauche, cette méthode peut répondre à notre recherche en nous fournissant des règles de discrimination. Cependant, toute combinaison de variables et de catégories fournissant une règle, leur nombre augmente de façon exponentielle avec le nombre de variables. Pour en obtenir un nombre raisonnable, on calcule des indices basés sur les comptages d'individus appartenant aux diverses catégories C_{ik} et le défi est de trouver le (ou les) indices de qualité en adéquation au problème posé. Contrairement aux ***arbres de décision*** qui sont validés par des méthodes éprouvées d'apprentissage automatique (ils ont été inclus dans la section précédente), les méthodes de validation des règles d'association sont encore à l'état de recherches (Cadot 2006). Dans notre exemple, plus de 50 règles sont générées, mais à notre connaissance les

indices de qualité actuels ne permettent pas d'en extraire les plus significatives (au sens statistique) ou les plus généralisables (au sens de l'apprentissage automatique).

La première apparition des ***réseaux bayésiens*** date de Pearl (1988) qui a représenté par un graphe les liens probabilistes entre des faits. Les liens au sein d'un groupe de variables peuvent être complexes, faisant appel à des notions de probabilités jointes, d'indépendance conditionnelle (Naïm et al. 2007). On y retrouve la notion de règles qui sont représentées par les flèches entre les variables (voir figures 5 et 6). La construction de chaque flèche du réseau peut se faire en dehors de toute donnée, d'après la connaissance de l'expert qui doit indiquer les dépendances, et donner les valeurs de probabilité associées. Le raisonnement se fait alors de façon assez intuitive en suivant les flèches du réseau. On peut aussi construire le réseau automatiquement, en utilisant un algorithme spécifique, ce qui s'appelle « apprentissage de la structure ». C'est cette voie que nous avons choisie, afin d'extraire par cette méthode un modèle des données Iris. Nous avons utilisé la version Lite du Logiciel HUGIN (www.hugin.com) sur les variables X quantitatives et la variable Y d'espèce transformée en 3 variables dichotomique, une par espèce, et parmi les 6 algorithmes proposés dans le logiciel nous avons choisi d'abord l'algorithme Rebane-Pearl Polytree. Le réseau obtenu est en figure 5, avec en noir les liens significatifs[2] à $p<0.05$. On y voit que les 3 espèces sont liées, et que les 4 caractéristiques des fleurs sont liées à la seule espèce Setosa. Bien que les liens entre les espèces soient justifiés (l'opposition entre 2 espèces est bien une dépendance), nous avons décidé d'essayer de les faire disparaître en abaissant le seuil de probabilité, comme le logiciel nous le permettait. Et avec $p<3.10^{-13}$ nous avons obtenu que seules les flèches de la partie bleue sont restées.

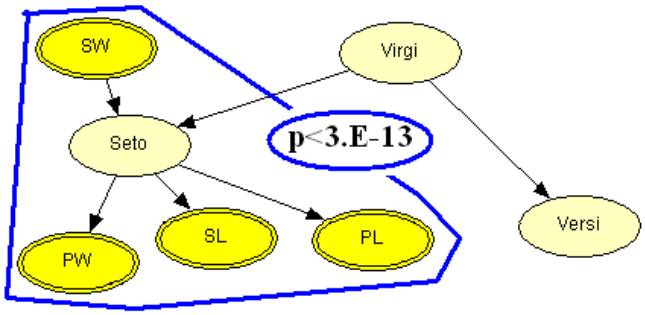

FIG. 5 – Réseau bayésien des données Iris (HuginLiteR8, algorithme Rebane-Pearl Polytree). En noir les flèches correspondant à p<0.05, Dans la zone bleue celles qui restent en imposant $p<3.10^{-13}$.

Même en ne gardant que la partie bleue, ce réseau ne nous satisfaisait pas, dans la mesure où il ne permettait pas de prédire chaque espèce Y d'après les caractéristiques X des fleurs. Nous avons alors transformé chaque variable quantitative en 3 variables binaires comme proposé dans C.H.I.C. (voir le détail de la décomposition section suivante) et avec l'algorithme « Chow-Liu tree », en prenant comme paramètre « root=steto », nous avons

[2] p est la probabilité qu'on aurait d'observer un lien alors qu'il n'y en a pas, à cause des fluctuations d'échantillonnage. Par défaut le logiciel propose de ne dessiner que les liens pour lesquels $p<0.05$, correspondant à un risque (appelé risque α) inférieur à 5% de se tromper en décidant à tort que le lien n'est pas dû au hasard.

obtenu le réseau bayésien de la figure 6. Les relations que nous obtenons avec cet algorithme entre chaque espèce d'iris et les variables quantitatives nous paraissent appropriées. Mais les relations obtenues avec les 5 autres algorithmes proposés par le logiciel ne nous ont pas satisfaite.

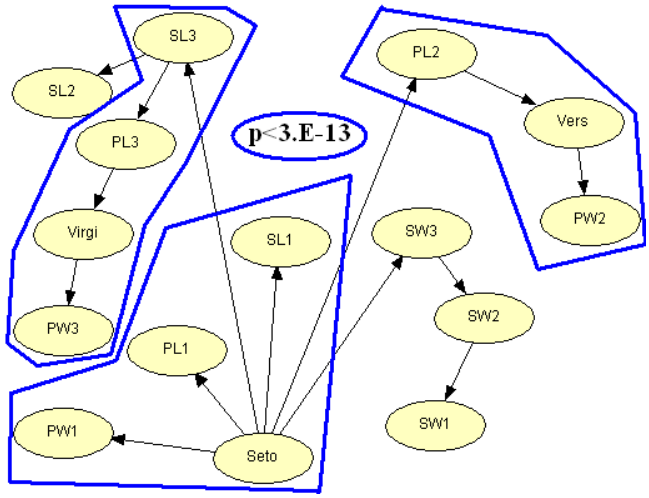

FIG. 6 – Réseau bayésien des données Iris avec les variables quantitatives de X recodées chacune en 3 variables (HuginLiteR8 avec « algorithme=Chow-Liu tree », et « root=steto »). En noir les flèches correspondant à p<0.05, Dans les zones bleues celles qui restent en imposant $p<3.10^{-13}$.

3 Où se situe l'approche méthodologique de l'A.S.I. ?

Dans sa version originale, l'A.S.I. s'applique à des variables binaires, qu'on obtient souvent par éclatement de variables catégorielles (la variable Genre avec 2 catégories F et M, donnera par éclatement 2 variables binaires F et M). Les règles fournies ressemblent donc aux règles d'association (RA) mais la méthode présente en outre 2 avantages qu'on retrouve dans le logiciel Hugin : 1) l'interface graphique de C.H.I.C. permet de « voir » les règles et de les manipuler 2) seules les règles valides s'affichent. Dans C.H.I.C., la validité des règles est établie selon la théorie statistique asymptotique (comptages vérifiant la loi de poisson ou binomiale, selon un choix à préciser dans les options de C.H.I.C.).

Dans la version actuelle, on peut utiliser des variables quantitatives telles quelles, ou les faire recoder en variables catégorielles. Nous avons utilisé les données des Iris pour confronter les réponses que l'A.S.I. offre à notre problème de discrimination X→Y. Nous avons d'abord utilisé les variables sans les découper mais en les remplaçant par une valeur entre 0 et 1 selon la formule indiquée dans l'aide :

valeur_nouvelle=(valeur_ancienne-min_valeurs)/(max_valeurs-min_valeurs),

Puis nous les avons fait découper en 2, 3, ou 4 parties par le logiciel C.H.I.C.. Le graphe implicatif des variables quantitatives non recodées, seulement réajustées pour être dans l'intervalle [0 ; 1] est en figure 7, à gauche, et à droite on a celui obtenu avec les variables

découpées en 3 par C.H.I.C., les 2 autres découpages testés, qui se sont avérés moins bons, ont été mis en annexe de Cadot et al. (2015).

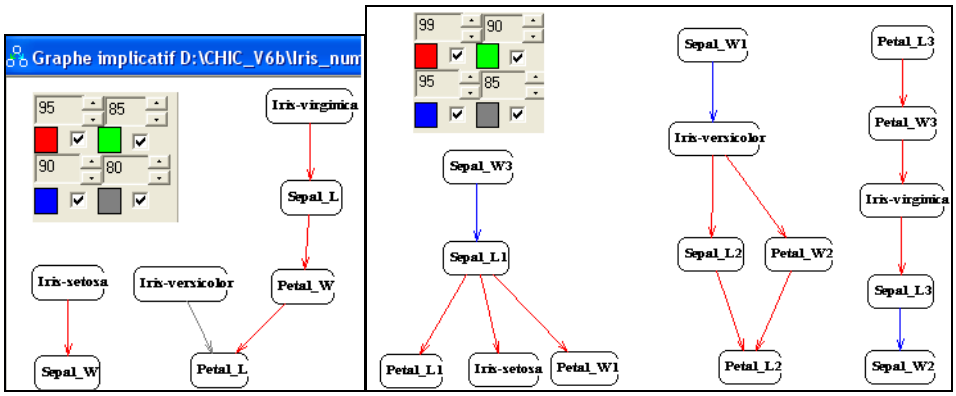

FIG. 7 – Graphes implicatifs produits par C.H.I.C. sur les 150 iris. À gauche les 4 variables quantitatives à valeurs réajustées dans [0 ; 1], à droite recodées en 3 catégories chacune par C.H.I.C..

La première différence entre les deux graphiques de la figure 7 porte sur les seuils. Comme nous voulions faire apparaître la totalité des 3 catégories, dans le graphique de gauche, nous avons dû descendre jusqu'à un seuil d'implication statistique de 0.80 pour faire apparaître Iris-versicolor, ce qui fournit donc une règle de piètre qualité pour cette espèce. La recherche de règles de bonne qualité pour chaque espèce d'iris est la première raison qui nous fait préférer le graphique de droite.

La deuxième raison qui nous fait préférer le graphique de droite est que chaque espèce apparaît dans un graphe séparé. Cette séparation exprime graphiquement ce que les autres modèles nous ont indiqué dans des tables de confusion croisant les espèces estimées par le modèle et les espèces observées : connaissant les mesures des pétales et des sépales, on peut en déduire avec peu d'erreurs l'espèce de l'iris. Comparons maintenant la figure 7 à la figure 2. La régression de Petal_L sur Sepal_L n'était significative que pour iris-virginica et iris-versicolor, pas pour iris-setosa, comme on peut le voir dans les équations à droite du graphique de la figure 2 et nous retrouvons un peu cela dans le graphique de gauche de la figure 7, dans la mesure où les 2 variables Petal_L et Sepal_L sont reliées entre elles fortement ($p>0.95$) et également avec les 2 espèces d'iris convenables. Le modèle de régression linéaire nous en dit plus, notamment, dans l'équation globale, les coefficients qui diffèrent significativement de 0 sont les seules interactions entre Sepal_L et chacune des 2 catégories. Mais ce qu'il dit est d'interprétation difficile, et si l'on faisait la régression pour chaque sous-ensemble de variables cela poserait des problèmes d'interprétation encore plus complexes. Si on prend maintenant le graphique de droite de la figure 7 nous pouvons dire que la liaison des variables Sepal_L et Petal_L est forte car chaque fois qu'elles sont reliées par une flèche, c'est avec la même catégorie (1, 2 ou 3), et à chaque catégorie correspond une espèce d'iris. Sauf pour l'espèce setosa, un chemin rouge (seuil=0.99) fait de 2 flèches qui se suivent joint l'espèce et les 2 catégories de Sepal_L et Petal_L, on retrouve ainsi les éléments déjà cités de la figure 2, mais en plus on a une information fine sur la relation entre les espèces et les va-

riables Sepal_L et Petal_L : les valeurs les plus faibles des 2 variables sont pour iris-setosa, les valeurs intermédiaires pour iris-versicolor et les plus fortes pour iris-virginica. Et on voit dans ces 3 graphes que les niveaux de Petal_W sont associés aux espèces de la même façon. Par contre la variable Sepal_W se comporte différemment : le niveau le plus faible est associé à versicolor, le moyen à virgina et le plus fort à setosa, et ces associations sont plus faibles (entre 0.95 et 0.99).

Jusqu'ici, la balance penche fortement du côté du graphique de droite de la figure 7 : Il nous donne des informations qui rejoignent celles que nous avons trouvées avec les autres modèles donnés en exemples, ainsi que des informations supplémentaires que nous n'avons pas su extraire des autres modèles, pour autant qu'elles y soient.

Maintenant se pose la question de l'interprétation du sens des flèches, et notamment : pourquoi la position en hauteur de l'espèce diffère-t-elle d'un graphe à l'autre dans le graphique de droite de la figure 7. Faut-il interpréter différemment ces 3 graphes ? Ce qui nous renvoie à la sémantique de ces flèches. Dans l'ouvrage collectif sur l'A.S.I. (Gras *et al*, 2013) nous lisons dans l'introduction (Gras et Regnier, section 4, p. 16-17) que cette flèche peut indiquer une inclusion plutôt qu'une causalité, ce que nous appellerons un « effet d'effectif ».

Niveau	Sepal_L	Sepal_W	Petal_L	Petal_W
1	46	47	50	50
2	53	67	54	52
3	51	36	46	48

TAB. 1 – *Effectifs selon les catégories obtenues en découpant chaque variable quantitative en 3 par C.H.I.C. (ces « informations sur le fichier » figurent sous cette dénominations dans les sorties de C.H.I.C.).*

Avant donc d'essayer d'interpréter le sens des flèches pour les iris, examinons la relation entre l'ordre des effectifs et la hauteur dans le graphe. Dans le tableau 1 on a reporté les effectifs de chacune des 3 catégories des variables selon le découpage fait par C.H.I.C.. On voit que la variable Sepal_W a des catégories déséquilibrées, ce qui n'est pas le cas des 3 autres variables, dont les effectifs restent entre 46 et 54. Dans le niveau 3, si on range ces 3 variables ainsi que l'espèce par effectif croissant, on obtient (le symbole « < » se lira « précède ») :

Petal_L (46) < Petal_W (48) < Iris-virginica (50) < Sepal_L (51).

C'est exactement dans cet ordre que les flèches successives joignent les items du 3ème graphe, figure 7, à droite. Pour le niveau 2, on a :

Iris-versicolor (50) < Petal_W (52) < Sepal_L (53) < Petal_L (54),

Et c'est quA.S.I.ment dans cet ordre qu'on trouve le graphe 2, sauf que Petal_W et Sepal_L (53) ne sont pas l'un sous l'autre mais l'un à côté de l'autre. Pour le premier graphe, on retrouve exactement l'ordre des effectifs.

Dans notre exemple, le sens des flèches pourrait ainsi être un effet d'effectif. D'ailleurs, il paraît difficile de justifier sémantiquement un chemin de causalité différent pour chaque espèce d'iris, par exemple le fait d'avoir l'implication « sépales courtes → iris-setosa » et l'implication inverse « iris-virginica→sépales longues ». Nous décidons donc de ne pas essayer d'interpréter en termes de causalité le sens des flèches dans les graphes implicatifs.

Nous essayons maintenant d'en déduire des règles de discrimination. Pour cela, nous utilisons le fichier de données recodées après découpage en 3 des variables de X par C.H.I.C., et nous en extrayons sous Excel un tableau de contingence (voir tableau 2) croisant la variable Y des catégories avec les 3 variables de X (Petal_L, Petal_W et Sepal_L) liées aux catégories de Y par des flèches rouges dans le graphe implicatif (Figure 7, à droite).

Nombre de Id			Species			
Petal_L	Petal_W	Sepal_L	Iris-setosa	Iris-versicolor	Iris-virginica	Total
1	1	1	40			40
		2	10			10
2	2	1		5		5
		2		29		29
		3		13		13
	3	1			1	1
		2		1	4	5
		3			1	1
3	2	2		1	2	3
		3			2	2
	3	2			6	6
		3		1	34	35
Total			50	50	50	150

TAB. 2 – Répartition des 150 iris selon leur espèce et 3 de leurs dimensions, en rouge les iris « mal classés ».

La lecture de ce tableau nous permet d'écrire le jeu de trois règles de discrimination suivant :

- Si Petal_L1 alors iris_setosa (N=50, nbVrai=50, nbFaux=0) ;
- Si Petal_L2 et Petal_W2 alors iris_versicolor (N=47, nbVrai=47, nbFaux=0) ;
- Si (Petal_L2 et Petal_W3) ou (Petal_L3) alors iris_virginica (N=53, nbVrai=50, nbFaux=3).

Et en récupérant les intervalles de valeurs des variables figurant dans le journal de C.H.I.C. (voir tableau 3), on peut remplacer les modalités des variables quantitatives dans les règles par leur appartenance à un intervalle (par exemple Petal_L1 devient Petal_L ∈ [1 ; 1.9]).

Niveau	Sepal_L	Sepal_W	Petal_L	Petal_W
1	de 4.3 à 5.3	de 2.0 à 2.8	de 1 à 1.9	de 0.1 à 0.6
2	de 5.4 à 6.2	de 2.9 à 3.3	de 3 à 4.9	de 1 à 1.6
3	de 6.3 à 7.9	de 3.4 à 4.4	de 5 à 6.9	de 1.7 à 2.5

TAB. 3 – *Intervalle de valeurs de chaque niveau des variables quantitatives donné par C.H.I.C..*

Nous constatons que ce jeu de trois règles permet de discriminer les 150 iris avec seulement 3 erreurs. Ce qui nous permet d'affirmer que C.H.I.C. nous a fourni pour les iris un modèle tout à fait pertinent de la relation X→Y, où X est la matrice de 4 variables quantitatives et Y une variable catégorielle.

4 Pourquoi et comment choisit-on le cadre théorique de l'A.S.I.

Comme on peut le constater plus haut, les diverses méthodes visant à modéliser X→Y arrivent presque toutes à discriminer les 3 catégories d'iris avec plus de 95% de réussite, ce qui n'est pas surprenant sachant que dans le sous-espace de X formé de la longueur et de la largeur des pétales (voir figure 1), une espèce se détache nettement des deux autres, ces dernières formant 2 blocs distincts juxtaposés avec une zone commune de moins de dix individus plus difficiles à catégoriser. C'est donc sur d'autres critères, qualitatifs, que nous choisissons la méthode destinée à nous fournir un modèle.

4.1 Comparaison entre les méthodes de l'A.S.I. et d'autres méthodes de modélisation

Nous avons donné une liste de méthodes visant à produire un modèle simple et fiable de la relation X→Y. D'abord les méthodes statistiques linéaires pour lesquelles la fiabilité est assurée par la théorie statistique (tests d'hypothèses), et la simplicité par la linéarité, ce qui signifie que la valeur prédite de Y est combinaison linéaire des valeurs de X. C'est la formule la plus simple quand Y est quantitative, comme dans la figure 2 où Y est la longueur du pétale. Mais quand Y n'est pas quantitative, ce qui est le cas de l'espèce, il faut faire des transformations préalables des données pour conserver la linéarité, et le modèle se complexifie, même si les résultats peuvent s'exprimer par des équations et des graphiques comme en figure 3. Quand X contient peu de variables, on peut s'y retrouver, mais avec 29 variables comme dans les données d'intonation que nous avons traitées dans Cadot et al (2015), le choix et l'utilisation d'un tel modèle s'avère délicat. De plus les conditions d'application exigées par ces méthodes statistiques linéaires ou dérivées sont vérifiées par les variables X des Iris de Fisher, mais pas par celles des données d'intonation, qu'il faut transformer pour les adapter. Une fois ces étapes franchies, on dispose d'un modèle explicite utilisable pour la discrimination, et de la mesure de sa fiabilité, qui peut s'avérer plus ou moins forte. Quand elle est suffisamment forte, elle permet d'établir la théorie visée, d'où son intérêt.

Les méthodes de discrimination suivantes que nous avons vues (MSVM, arbres de décision) fournissent des résultats fiables sans exiger que les données vérifient autant de conditions que les méthodes statistiques précédentes, mais soit le modèle détaillé n'est pas fourni (fonctionnement des MSVM en « boîte noire »), soit il est simpliste (arbres de décision), la qualité se situant au niveau du résultat (taux élevé de prédictions réussies), pas du modèle[3]. Notons toutefois que ce n'est pas parce que ces méthodes n'exigent pas de conditions sur les distributions des données qu'elles pourront traiter correctement des données mal distribuées. Quant aux règles d'association, elles fournissent un modèle riche, mais complexe et qui reste attaché aux données traitées, par manque de méthode de validation statistique universellement reconnue. Leur avantage est de faire découvrir des relations locales inattendues, qui peuvent entrer dans la composition d'un modèle global qu'il conviendra de valider ensuite.

L'A.S.I. est à la frontière entre les deux premiers types de méthodes : elle fournit un modèle des données dont on peut évaluer la fiabilité grâce aux seuils indiqués dans l'interface

[3] A nos yeux, la qualité d'un modèle est fonction de son intelligibilité, de sa concision, opérationnalité, …

graphique, mais sans avoir d'exigences sur la distribution des données. Son modèle sous forme de graphe implicatif permet d'exprimer des relations complexes entre les variables au sein d'un modèle global, mais comme elles ont toutes le même statut (pas de variable à expliquer ou explicative), la prédiction ne peut se faire directement avec le logiciel C.H.I.C.. L'interface graphique permet de modifier les éléments du modèle (ajouter/retirer des variables, changer les seuils, déplacer des groupes de flèches, …) par simples clics, permettant le cas échéant de se servir des résultats dans une optique de prédiction.

Quant aux réseaux bayésiens, ils nous ont fourni un modèle convenable (voir figure 6) pour seulement un algorithme parmi les six proposés : son utilisation nous semble bien risquée. De plus, nous avions déjà comparé dans Cadot (2009) l'A.S.I., les réseaux bayésiens et les treillis de Galois (graphe associé aux règles d'association) sur d'autres données et nous avions déjà privilégié l'utilisation de l'A.S.I. avec C.H.I.C..

Notre but étant d'extraire de nos données un modèle fiable, simple mais riche, nous avons choisi C.H.I.C. et ses graphes implicatifs, sachant qu'avec la facilité de manipulation procurée par l'interface graphique de C.H.I.C., nous pouvons ajuster par clics jusqu'à obtenir le modèle qui nous convient le mieux sur les données d'intonation, sans être obligées de diagnostiquer leurs « défauts » pour les corriger au préalable (liaisons fortes entre certaines variables de X, distributions déséquilibrées, nombreux ex-æquo, valeurs extrêmes, etc.).

4.2 Découper ou non des variables quantitatives ?

Transformer une variable quantitative en 3 catégories fait perdre de la précision sur les valeurs (on ne dispose plus que de 3 valeurs non ordonnées), et on pourrait penser qu'il vaut mieux l'éviter pour avoir un modèle plus fiable. Nous avons vu que ce n'est pas le cas avec le logiciel C.H.I.C. pour les Iris, le modèle avec les variables découpées en 3 catégories ayant été préféré à celui avec les variables simplement recodées sur l'intervalle [0 ; 1]. Il semblerait donc que le recodage de variables quantitatives en variables catégorielles ne diminue par la précision du graphe implicatif. Et dans le cas des Iris, il lui en a fait gagner. Cela s'explique par le fait que 3 dimensions sur 4 (Petal_L, Petal_W, Sepal_L) de la fleur d'iris sont très liées entre elles et ont tendance à varier dans le même sens, et si on ne les découpe pas, elles se retrouvent dans le même graphe et liées aux mêmes espèces d'iris (voir figure 7, à gauche), la dernière variable (Sepal_W) se trouvant dans un autre graphe avec les espèces restantes d'iris, ce qui ne fait que deux graphes pour 3 espèces, donc une mauvaise discrimination. Ce n'est qu'en découpant les variables qu'on a pu obtenir une discrimination fine de chaque catégorie, avec des graphes séparés (voir Figure 7, à droite).

Ayant établi que découper les variables quantitatives peut augmenter la précision du modèle global, il s'agit maintenant de définir la façon de procéder. La découpe d'une variable quantitative se fait automatiquement dans le logiciel C.H.I.C., une fois choisi le nombre de parties. C'est un algorithme qui détermine les seuils de découpe en optimisant le rapport entre variance inter et variance intra, indépendamment des autres variables, contrairement à l'algorithme des arbres de décision, qui optimise en fonction de la variable à discriminer. Ce choix permet au logiciel C.H.I.C. d'obtenir des niveaux de significativité non biaisés en conservant sa méthode d'estimation asymptotique, alors que les arbres de décision utilisent des méthodes de validation croisée pour éviter ce biais. L'algorithme de C.H.I.C. fonctionnant avec des variances, il faut que la distribution des valeurs soit suffisamment équilibrée, ce qui est le cas des Iris. Dans le cas contraire, comme celui de certaines variables acoustiques, il est préférable de découper soi-même les variables selon d'autres critères, ne faisant

pas intervenir la variable Y, pour éviter d'introduire de biais dans la discrimination, en utilisant par exemple des quantiles pour avoir des catégories d'effectifs proches.

Le découpage peut se faire selon une théorie (par exemple pour la tension artérielle, découpage selon les seuils de l'hypotension et de l'hypertension). Comme nous ne disposions pas de théorie sur les dimensions des fleurs d'iris, nous avons profité de l'algorithme de découpage automatique proposé par le logiciel C.H.I.C. pour faire essayer divers découpages, en commençant par un découpage identique de toutes les variables, en deux, trois, puis quatre parties. Le découpage en trois parties, procuré par le logiciel C.H.I.C., a donné de meilleurs modèles : en demandant des seuils entre 0.80 et 0.99, nous avons obtenu 3 graphes séparés, un par espèce, et dans chaque graphe, plus d'une flèche rouge (indices≥0.99), une seule flèche bleue (0.95≤indice<0.99) et aucune flèche d'une autre couleur, donc des indices d'implication statistique tous supérieurs à 0.95. En effet, on peut voir en annexe qu'en fixant le nombre de parties à 2 ou 4 pour le découpage de toutes les variables, les espèces iris-virginica et iris_versicolor apparaissent dans le même graphe. Ensuite nous avons tenté deux autres découpages en un nombre différent de parties selon les variables des pétales et celles des sépales, qui sont joints en annexe de la partie 2.2.3. On peut y voir qu'un de ces deux graphes implicatifs l'emporte sur tous les autres découpages, y compris ceux avec toutes les variables découpées en 3 parties, c'est celui obtenu en découpant les pétales en trois et les sépales en deux, car il est formé de 3 graphes séparés (un par espèce), il ne contient que des flèches rouges (indices≥0.99) et il est plus concis (moins de variables après découpage). Toutefois, en créant à partir de ce graphe le tableau et le jeu de règles de discrimination comme nous l'avons fait dans la section précédente pour le découpage de chaque variable en 3 parties, nous constatons que sa qualité de discrimination n'est pas meilleure (également 3 erreurs sur les 150 iris), nous privilégions donc le modèle issu du découpage de toutes les variables en trois parties, plus facile à justifier théoriquement ici (nombre de parties des variables de X égal au nombre de catégories de Y).

Ayant perdu de la qualité en passant le nombre de parties du découpage de trois à quatre, nous n'avons pas essayé au-delà de quatre parties.

5 Conclusion et perspectives

Comme nous l'avons indiqué dans Cadot et al. (2015), en utilisant sur les données acoustiques le logiciel C.H.I.C. comme nous l'avions utilisé sur les données Iris, nous avons pu 1) confirmer les descriptions prosodiques théoriques reconnues, à l'aide d'indices spécialement conçus pour vérifier ces schémas ; 2) disposer d'éléments précis sur la manière dont ces indices codent les principales intonations du français.

Pour conclure, la méthodologie que nous avons choisie se base sur l'A.S.I., à travers le logiciel C.H.I.C., à la condition expresse de transformer les variables quantitatives en qualitatives. Et à la question « Pourquoi et comment transformer des variables quantitatives en catégorielles ? », nous pouvons répondre ainsi :

- « Pourquoi ? » : pour prédire une variable catégorielle, quand on dispose de variables explicatives quantitatives très liées entre elles, il est mieux de les transformer en variables catégorielles.
- « Comment ? » : si on ne dispose pas de théorie donnant des seuils de coupe, et donc le nombre de parties pour chaque variable, il est préférable d'essayer d'abord un découpage de chaque variable quantitative en k intervalles pour prédire k catégo-

ries. Le choix de ces intervalles peut se faire avec des quantiles pour obtenir des effectifs voisins.

Mais c'est une heuristique que nous proposons ici, bien sûr, à éprouver sur d'autres données et d'autres problématiques.

Références

Baillargeon, (2000), *La régression linéaire*, Edition des trois Sources, Quebec.

Besse, P. (2003), *Pratique de la modélisation Statistique*, Document interne du Laboratoire de Statistique et Probabilités, Université Paul Sabatier de Toulouse, http://www.math.univ-toulouse.fr/~besse/pub/modlin.pdf .

Breiman L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984), *Classification and Regression Trees*. Wadsworth.

Cadot, M. et A. Bonneau (2015), Pourquoi et comment transformer des variables quantitatives en catégorielles ? Application à l'intonation de la langue française, Actes de *A.S.I.8*, Radès, Tunisie.

Cadot, M. (2009), Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayésiens et les treillis de Galois, *Revue des Nouvelles Technologies de l'Information*, Hermann, E (16) : 223-250.

Cadot, M. (2006), *Extraire et valider les relations complexes en sciences humaines : statistiques, motifs et règles d'association,* Thèse de l'université de Franche-Comté.

Dagnelie P. (2003), *Principes d'expérimentation : planification des expériences et analyse de leurs résultats*. Grenoble, Presses Agronomiques de Gembloux, Edition électronique : http://www.dagnelie.be

Droesbeke J.-J. et J. Fine, éditeurs (1996), *Inférence non paramétrique, les statistiques de rangs*. Journées d'Etude en Statistiques de l'Association pour la Statistique et ses Utilisations, Edition de l'Université de Bruxelles, Ellipses

Fisher, R.A. (1936), The use of multiple measurements in taxonomic problems *Annual Eugenics*, **7**: II:179-188.

Gras R. et collaborateurs (1996), *L'implication statistique, une nouvelle méthode exploratoire de données*, La pensée sauvage, Grenoble.

Gras, R., J.-C. Régnier, C. Marinica, et F. Guillet (2013), *L'analyse statistique implicative : méthode exploratoire et confirmatoire à la recherche de causalités*, 2ème éd., Cépaduès, Toulouse.

Han, J. and M. Kamber (2001), *Data Mining: Concepts and Techniques, Morgan Kaufmann*, Publishers, San Francisco.

Hoc J.-M. (1983), *L'analyse planifiée des données en psychologie*, PUF, Paris.

Logiciel C.H.I.C. : Classification Hiérarchique Implicative et Cohésitive, Version 6.0, Copyright (c) 2012.

Mitchell, T. (1997). *Machine Learning*, McGraw Hill.

Morin H., (1999), *Théorie de l'échantillonnage*, Les presses de l'université, Laval.

Morineau, A., J.-P. Nakache, et C. Krzyzanowski (1996), *Le modèle log-linéaire et ses applications*, Cisia-Ceresta, Paris.

Naïm, P., P.-H. Willemin, P. Leray, O. Pourret, et A. Becker (2007), *Réseaux bayésiens*, 3ème éd., Eyrolles, Paris.

Nakache, J.-P. et J. Confais (2003), *Statistique explicative appliquée : analyse discriminante, modele logistique, segmentation*. Editions Technip, Paris, France.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, Cambridge University Press, New York.

Prum, B., (1996), *Modèle linéaire. Comparaison de groupes et régression*, Eyrolles, Collection INSERM, Paris.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*. Cambridge.

Schwartz, D. (1991), *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion, Paris.

Siegel S., et Jr. Castellan N.J. (1988), *Nonparametric statistics for the behavorial sciences*, McGraw-Hill, London.

Weston J. and C. Watkins (1998), *Multi-class support vector machines*. Technica l Report CSD-TR-98-04,Royal Holloway, University of London, Department of Computer Science.

Summary

A number of methods are available for deriving a categorization model of type X→Y out of a set of individual data, where X is a set of individual numerical features and Y their categories. We develop a brief overview of these methods by making use of the most popular ones for processing the well-known "Fisher's Iris" dataset. The comparison of the resulting models encourages us to give preference to ISA (Implicative Statistical Analysis) for this specific type of data, on condition of a thorough recoding of the quantitative variables. This paper incorporates and expands a communication made during A.S.I.8 conference (Cadot et al. 2015) in which we show the interest of the chosen methodology (ISA after a specific recoding step) for the processing of acoustic data.